ESAIM: PS 24 (2020) 69–99
https://doi.org/10.1051/ps/2019021

ESAIM: Probability and Statistics
www.esaim-ps.org

AN ADAPTIVE MULTICLASS NEAREST NEIGHBOR CLASSIFIER*

Nikita Puchkin[1,2,**] and Vladimir Spokoiny[1,2,3]

Abstract. We consider a problem of multiclass classification, where the training sample $S_n = \{(X_i, Y_i)\}_{i=1}^n$ is generated from the model $\mathbb{P}(Y = m|X = x) = \eta_m(x)$, $1 \leqslant m \leqslant M$, and $\eta_1(x), \dots, \eta_M(x)$ are unknown α-Holder continuous functions. Given a test point X, our goal is to predict its label. A widely used k-nearest-neighbors classifier constructs estimates of $\eta_1(X), \dots, \eta_M(X)$ and uses a plug-in rule for the prediction. However, it requires a proper choice of the smoothing parameter k, which may become tricky in some situations. We fix several integers $n_1, \dots, n_K$, compute corresponding n_k-nearest-neighbor estimates for each m and each n_k and apply an aggregation procedure. We study an algorithm, which constructs a convex combination of these estimates such that the aggregated estimate behaves approximately as well as an oracle choice. We also provide a non-asymptotic analysis of the procedure, prove its adaptation to the unknown smoothness parameter α and to the margin and establish rates of convergence under mild assumptions.

Mathematics Subject Classification. 62H30, 62G08.

Received December 23, 2018. Accepted October 24, 2019.

1. Introduction

Multiclass classification is a natural generalization of the well-studied problem of binary classification. It is a problem of supervised learning when one observes a sample $S_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $Y_i \in \mathcal{Y} = \{1, \dots, M\}$, $1 \leqslant i \leqslant n$, $M > 2$. The pairs (X_i, Y_i) are generated independently according to an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Given a test pair (X, Y), which is generated from $\mathcal{D}$ independently of S_n, the learner's task is to propose a rule $f : \mathcal{X} \to \{1, \dots, M\}$ in order to make a probability of misclassification

$$R(f) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}\left(Y \neq f(X)\right)$$

as small as possible. In practice, it is a common situation when one has to discern between more than two classes, so multiclass classification has a wide range of applications and arises in such areas as bioinformatics, when one tries to predict a protein's fold [14] or when one wants to classify DNA microarrays [32], finance when

*Financial support by the Russian Academic Excellence Project 5-100 and by the German Research Foundation (DFG) through the Collaborative Research Center 1294 is gratefully acknowledged.

Keywords and phrases: Multiclass classification, k nearest neighbors, adaptive procedures.

[1] National Research University Higher School of Economics, 20 Myasnitskaya ulitsa, 101000 Moscow, RF.
[2] Institute for Information Transmission Problems RAS, Bolshoy Karetny per. 19, 127051 Moscow, RF.
[3] Weierstrass Institute and Humboldt University, Mohrenstr. 39, 10117 Berlin, Germany.

** Corresponding author: npuchkin@hse.ru

one predicts a corporate credit rating [2], image analysis [20] when one tries to classify an object on an image, speech recognition [18] and others.

Concerning the multiclass learning problem, one can distinguish between two main approaches. The first one is by reducing to the binary classification. The most popular and straightforward examples of these techniques are One-vs-All (OvA) and One-vs-One (OvO). Another example of reduction to the binary case is given by error correcting output codes (ECOC) [13]. In [3], this approach was generalized for margin classifiers. A similar approach uses tree-based classifiers. Methods of the second type solve a single problem such as it is done in multiclass SVM [9] and multiclass one-inclusion graph strategy [29]. Daniely, Sabato, and Shalev-Shwartz in [11] provided a theoretical comparison of OvA, OvO, ECOC, tree-based classifiers and multiclass SVM for linear discrimination rules in a finite-dimensional space. From their study, multiclass SVM outperforms the OvA method. In [9], Crammer and Singer also showed a superiority of multiclass SVM on several datasets. Nevertheless, in our work, we will use One-vs-All for two reasons. First, we will consider a broad nonparametric class of functions and results in [11] do not cover this case. Second, in [26], Rifkin and Klautau showed that OvA behaves comparably to multiclass SVM if binary classifier in OvA is strong enough.

For each class m, we construct binary labels $\mathbb{1}(Y_i = m)$. We denote a marginal distribution of X by $\mathbb{P}_X$ and suppose that $\mathbb{P}_X$ has a density $p(X)$ with respect to the Lebesgue measure μ. Given X, we denote the conditional probability $\mathbb{P}(Y = m|X)$, $1 \leqslant m \leqslant M$ by $\eta_m(X)$. For this model, the optimal classifier f^* can be found analytically

$$f^*(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \eta_m(X).$$

Unfortunately, true values $\eta_1(X), \dots, \eta_M(X)$ are unknown but can be estimated. Since for any classifier f it holds $R(f) \geqslant R(f^*)$, then it is reasonable to consider the excess risk

$$\mathcal{E}(f) = R(f) - R(f^*),$$

which shows the quantitative difference between the classifier f and the best possible one. One of the most popular approaches to tackle the classification problem is a (weighted) k-nearest neighbors rule. Given a test point $X \in \mathcal{X}$, this rule constructs nearest neighbor estimates $\widehat{\eta}_1^{(NN)}(X), \dots, \widehat{\eta}_M^{(NN)}(X)$ of $\eta_1(X), \dots, \eta_M(X)$ and predicts the label Y at the point X by a plug-in rule:

$$\widehat{f}^{(NN)}(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \widehat{\eta}_m^{(NN)}(X).$$

Although the method is simple and known for a long time, several new finite sample results in the binary setting were obtained quite recently. In [30], the author considers weighted and bagged nearest neighbor estimates with smooth function $\eta_1(x)$ and finds optimal vector of non-negative weights. Moreover, the author goes further and derives faster rates under additional smoothness assumptions if the weights are allowed to be negative. However, the analysis in [30] requires that the marginal distribution of features must have a compact support and its density must be bounded away from zero (strong density assumption). In [8] and [16], authors address this issue. In [8], the authors introduce a novel Holder-like smoothness condition on $\eta_1(x)$ tailored to nearest neighbor. This trick allows to avoid the strong density assumption and boundedness of features. Although in both papers [8] and [16] the authors work under similar assumptions, their approach is different. In [8], the authors obtain convergence rates in terms of size of the effective boundary, while in [16] the authors use a more classical approximation-estimation tradeoff technique of statistical analysis. The disadvantage of the modified smoothness condition in [8] is that it is implicit. Instead of this condition, in [17], the authors introduce the minimal mass assumption and the tail assumption, which are proved to be necessary for quantitative analysis of nearest neighbor estimates and cover the case when marginal distribution of features has an unbounded support and has a density, which may be arbitrarily close to zero. Note that the nearest neighbor estimate $\widehat{\eta}_m^{(NN)}(X)$

strongly depends on the parameter k and its choice determines the performance of the classifier $\widehat{f}^{(NN)}$. Moreover, as pointed out in [7] and [17], the global nearest neighbor classifier (*i.e.* the number of neighbors is the same for all test points) may be suboptimal, while the nearest neighbor classifier with point-dependent choice of k shows a better performance. In multiclass setting, the situation is even more difficult, because for each class the optimal number of neighbors may be different and this fact becomes crucial when for two classes (say, 1 and 2) and a test point X the values $\eta_1(X)$ and $\eta_2(X)$ are very close, and one has to estimate these values as precisely as possible to avoid misclassification. Since for each test point X and each class m, the optimal value of k may be different, the tuning procedure becomes tricky. To solve this problem, we consider a sequence of integers $n_1, \dots, n_K$, compute weighted nearest neighbor estimates for each of them and use a plug-in classifier based on a convex combination of these estimates.

An aggregation of the nearest neighbor estimates is a key feature of our procedure. We use a multiclass spatial stagewise aggregation (SSA), which originates from [5], where an aggregation of binary classifiers was studied. Unlike many other aggregation procedures, such as exponential weighting [21, 28, 33], mirror averaging [19, 34], empirical risk minimization [22], and Q-aggregation [10, 23], which perform *global* aggregation, SSA makes *local* aggregation yielding a point dependent aggregation scheme. This means that the aggregating coefficients depend on the point X where the classification rule is applied. The drawback of the original SSA procedure [5] is that it is tightly related to the Kullback-Leibler aggregation and, therefore, puts some restrictions, which are usual for such setup and appear in other works on this topic (for instance [6, 27]) but are completely unnecessary for the classification task. We show that, in a special case of the multiclass classification, one can omit those restrictions and obtain the same results under weaker assumptions.

Finally, it is worth mentioning that nonparametric estimates have slow rates of convergence especially in the case of high dimension d. It was shown in [4] and then in [15] that plug-in classifiers can achieve fast learning rates under certain assumptions in both binary and multiclass classification problems. We will use a similar technique to derive fast learning rates for the plug-in classifier based on the aggregated estimate.

Main contributions of this paper are the following:

- we propose a computationally efficient algorithm of multiclass classification, which is based on aggregation of nearest neighbor estimates;
- the procedure automatically chooses an almost optimal number of neighbors for each test point and each class;
- the procedure adapts to an unknown smoothness of the functions $\eta_1(\cdot), \dots, \eta_M(\cdot)$;
- we provide theoretical guarantees on large deviations of the excess risk and on its mean value as well under mild assumptions; theoretical guarantees claim optimal accuracy of classification with only a logarithmic payment for adaptation.

The rest of the paper is organized as follows. In Section 2, we give auxiliary definitions and introduce some notations. In Section 3, we formulate the multiclass classification procedure and then provide its theoretical properties in Section 4. Section A is devoted to the proof of the main result, which is given in Theorem 4.1. Some auxiliary results and proofs are moved to the appendix. Finally, in Section 5, we demonstrate a performance of the procedure on both artificial and real datasets.

2. Preliminaries and notations

We start with a simple observation. Introduce

$$\varphi(t) = \left(\frac{1}{2M} \vee t\right) \wedge \left(1 - \frac{1}{2M}\right). \tag{2.1}$$

It is easy to show that for the truncated function

$$\theta_m(X) = \varphi(\eta_m(X)) \equiv \left(\frac{1}{2M} \vee \eta_m(X)\right) \wedge \left(1 - \frac{1}{2M}\right),$$

it holds

$$\operatorname*{argmax}_{1 \leqslant m \leqslant M} \eta_m(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \theta_m(X),$$

and, instead of the value $\eta_m(x)$, one can estimate $\theta_m(x)$ at a point x. In our approach, we consider a plug-in classifier

$$\widehat{f}(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \widehat{\theta}_m(X),$$

where $\widehat{\theta}_m(x)$ stands for an estimate of $\theta_m(x)$, $1 \leqslant m \leqslant M$, at the point x.

Now, the problem is to estimate $\theta_m(x)$, $1 \leqslant m \leqslant M$. Fix some m and transform the labels into binarized ones: $\mathbb{1}(Y_i = m)$. It is clear that

$$\left(\mathbb{1}(Y_i = m) \,|X_i\right) \sim \text{Bernoulli}(\eta_m(X_i)).$$

This approach is nothing but the One-vs-All procedure for multiclass classification. Then a weighted k-nearest-neighbor estimate of $\theta_m(x)$ at the point x can be expressed as $\widetilde{\theta}_m^w(x) = \varphi(\widetilde{\eta}_m^w(x))$ and

$$\widetilde{\eta}_m^w(x) = \frac{\sum\limits_{i=1}^n w_i(X_i, x)\mathbb{1}(Y_i = m)}{\sum\limits_{i=1}^n w_i(X_i, x)} \equiv \frac{S_m^w(x)}{N_w(x)}, \tag{2.2}$$

where $S_m^w(x) = \sum\limits_{i=1}^n w_i(X_i, x)\mathbb{1}(Y_i = m)$, $N_w(x) = \sum\limits_{i=1}^n w_i(X_i, x)$, is a weighted nearest neighbor estimate of $\eta_m(x)$. The non-negative weights $w_i(X_i, x)$ depend on the distance between X_i and x and $w_i(X_i, x) > 0$ if and only if X_i is among k nearest neighbors of x; otherwise, $w_i(X_i, x) = 0$. In this paper, we consider the weights of the following form:

$$w_i = w_i(X_i, x) = \mathscr{K}\left(\frac{\|X_i - x\|}{h}\right), \tag{2.3}$$

where a bandwidth $h = h(\mathsf{k})$ is a distance to the k-th nearest neighbor and the kernel $\mathscr{K}(\cdot)$ fulfills the following conditions:

- $\mathscr{K}(t)$ is a non-increasing funciton,
- $\mathscr{K}(0) = 1$,
- $\mathscr{K}(1) \geqslant \dfrac{1}{2}$,
- $\mathscr{K}(t) = 0, \quad \forall\, t > 1.$

(A1)

This assumption can be easily satisfied. First, note that the rectangular kernel $\mathscr{K}(t) = \mathbb{1}(0 \leqslant t \leqslant 1)$ meets these requirements and, therefore, (A1) holds for the case of ordinary nearest neighbor estimates. There are

other examples of such kernels $\mathscr{K}$. For instance, one can easily check that Epanechnikov-like and Gaussian-like kernels, $\mathscr{K}(t) = (1 - t^2/2)\mathbb{1}(0 \leqslant t \leqslant 1)$ and $\mathscr{K}(t) = e^{-t^2/2}\mathbb{1}(0 \leqslant t \leqslant 1)$ respectively, fulfill (A1). It is also important to mention that here and further in this paper, without loss of generality, we suppose that a tie (*i.e.* a situation, when there are several candidates for the k-th nearest neighbor) does not happen almost surely. Otherwise, one can use the tie-breaking procedure described in [8].

The nearest neighbor estimate (2.2) requires a proper choice of the parameter k. Moreover, an optimal value of k may be different for each test point x and each class m, and the problem of a fine parameter tuning may become tricky. Instead of using one universal value of the number of neighbors, we fix an increasing sequence of integers $\{n_k : 1 \leqslant k \leqslant K\}$. We only require that there exist constants $0 < u_0 < u < 1$ such that

$$2u_0 \leqslant \frac{n_{k-1}}{n_k} \leqslant \frac{u}{2}, \quad 1 \leqslant k \leqslant K, \tag{A2}$$

and there are positive constants a and b such that $n_1 \leqslant a$ and $n_K \geqslant bn^{2/(d+2)}$. Each n_k induces a set of weights $w_1^{(k)}, \dots, w_n^{(k)}$ with

$$w_i^{(k)} = w_i^{(k)}(X_i, x) = \mathscr{K}\left(\frac{\|X_i - x\|}{h_k}\right), \tag{2.4}$$

where h_k stands for the distance to the n_k-th nearest neighbor, and a weighted n_k-NN estimator:

$$\widetilde{\theta}_m^{(k)}(x) = \varphi\left(\widetilde{\eta}_m^{(k)}(x)\right) \equiv \left(\frac{1}{2M} \vee \widetilde{\eta}_m^{(k)}(x)\right) \wedge \left(1 - \frac{1}{2M}\right), \tag{2.5}$$

$$\widetilde{\eta}_m^{(k)}(x) = \frac{S_m^{(k)}(x)}{N_k(x)}, \tag{2.6}$$

where $S_m^{(k)}(x) = \sum\limits_{i=1}^{n} w_i^{(k)}(X_i, x)\mathbb{1}(Y_i = m)$, $N_k(x) = \sum\limits_{i=1}^{n} w_i^{(k)}(X_i, x)$. Then one can use the SSA procedure [5] to construct aggregated estimates $\widehat{\theta}_1(x), \dots, \widehat{\theta}_M(x)$. The final prediction of the label at the point x is given by the plug-in rule (2). The detailed description of the procedure for multiclass classification is given in Section 5. We will refer to it as MSSA (short for Multiclass Spatial Stagewise Aggregation).

To show a consistency of the MSSA procedure, we will derive upper bounds for the generalization error $\mathbb{P}_{(X,Y)\sim\mathcal{D}}\left(Y \neq \widehat{f}(X) \middle| S_n\right)$ of the classifier $\widehat{f}$, which hold in mean and with high probability over training samples S_n. As a byproduct, we will provide convergence rates for the pointwise error $\max\limits_{1\leqslant m\leqslant M} |\widehat{\theta}_m(x) - \theta_m^*(x)|$ and obtain a user-friendly bound on the performance of the nearest neighbor estimates under mild assumptions. Namely, along with (A1) and (A2), we assume the following. First, the functions $\eta_m(\cdot)$ are (L, α)-Holder continuous, *i.e.* there exist $L > 0$ and $\alpha > 0$ such that for all $x, x' \in \mathcal{X}$ and $1 \leqslant m \leqslant M$ it holds

$$|\eta_m(x) - \eta_m(x')| \leqslant L\|x - x'\|^\alpha. \tag{A3}$$

Second, since we deal with the problem of nonparametric classification, even the optimal rule can show poor performance in the case of a large dimension d. Low noise assumptions are usually used to speed up rates of convergence and allow plug-in classifiers to achieve fast rates. We can rewrite

$$\begin{aligned} R(f) &= 1 - \mathbb{E}_{(X,Y)\sim\mathcal{D}}\mathbb{1}(Y = f(X)) \\ &= 1 - \mathbb{E}_X\mathbb{P}(Y = f(X)|X) = 1 - \mathbb{E}_X\eta_{f(X)}(X). \end{aligned} \tag{2.7}$$

In the case of binary classification, a misclassification often occurs, when $\eta_1(X) \equiv \mathbb{P}(Y = 1|X)$ is close to $1/2$ with a high probability. The well-known Mammen-Tsybakov noise condition [24] ensures that such a situation appears rarely. More precisely, it assumes that there exist non-negative constants B and β such that for all $t > 0$ it holds

$$\mathbb{P}_X\left(|2\eta_1(X) - 1| < t\right) \leqslant Bt^{\beta}.$$

This assumption can be extended to the multiclass case. Let $\eta_{(1)}(x) \geqslant \eta_{(2)}(x) \geqslant \ldots \geqslant \eta_{(M)}(x)$ be the ordered values of $\eta_1(x), \ldots, \eta_M(x)$. Then the condition (2) for the multiclass classification can be formulated as follows (see [1, 25]): there exist $B > 0$ and $\beta \geqslant 0$ such that for all $t > 0$ it holds

$$\mathbb{P}_X\left(\eta_{(1)}(X) - \eta_{(2)}(X) < t\right) \leqslant Bt^{\beta}. \tag{A4}$$

We will use this assumption to establish fast rates for the plug-in classifier $\widehat{f}(X)$ in Section 4.

There are two more requirements we need: the minimal mass assumption and the tail assumption introduced in [17]. The first one assumes that there exist $\varkappa > 0$ and $r_0 > 0$, such that for all $r \in (0, r_0]$ and $x \in \mathrm{supp}(\mathbb{P}_X)$ it holds

$$\mathbb{P}_X(X \in B(x, r)) \geqslant \varkappa p(x) r^d, \tag{A5}$$

where $B(x, r)$ stands for the Euclidean ball of radius r centered at x and $p(x)$ is a density of the marginal distribution $\mathbb{P}_X$ with respect to the measure μ. The tail assumption admits that there are positive constants C, ε_0 and p such that for every $\varepsilon \in (0, \varepsilon_0]$ it holds

$$\mathbb{P}_X\left(p(X) < \varepsilon\right) \leqslant C\varepsilon^p. \tag{A6}$$

It was discussed in [17] (Thm. 4.1) that the conditions (A5) and (A6) are necessary for quantitative analysis of classifiers and cannot be removed.

One can pick out a simple case of a bounded away from zero density when for any $x \in \mathrm{supp}(\mathbb{P}_X)$ it holds $p(x) \geqslant p_0 > 0$ with a positive constant p_0. The most difficult points x for classification with the nearest neighbor rule are those points, which are close to the decision boundary or where the density $p(x)$ approaches zero, because in this case a vicinity of x may not contain the sample points at all. One of the ways to control the misclassification error in the low-density region is to impose a modified smoothness condition on the regression function $\eta(\cdot)$, as it is done in [8, 16]: they assume that there are constants $L > 0$ and $\alpha \in (0, 1]$, such that for all $x, x' \in \mathcal{X}$ it holds

$$|\eta(x) - \eta(x')| \leqslant L\left(\mathbb{P}_X\left\{B(x, \|x - x'\|)\right\}\right)^{\alpha/d}.$$

This assumption ensures that in the regions with a small density $p(x)$ the function $\eta(x)$ is (L', α)-Holder continuous with a small constant L'. An approach, considered in [17], uses assumptions (A5) and (A6) instead of the modified smoothness condition. The assumption (A5) helps to control the minimal probability mass of the ball $B(x, r)$ in regions where the density $p(x)$ is close to zero. A curious reader can ensure that all the results we formulate will also hold if $p(x)$ and $\varkappa$ in (A5) are replaced with p_0 and $\mu(B(0, 1))$ respectively in the case of a bounded away from zero density $p(x)$. Also, note that in this case, the assumption (A6) is satisfied with $\varepsilon_0 < \min\{1, p_0\}$ and the power $p = \infty$.

We proceed with several examples of distributions when the tail assumption (A6) holds. For instance, univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$, exponential distribution $\mathrm{Exp}(\lambda)$, gamma-distribution $\mathrm{Gamma}(k, \lambda)$, Cauchy and Pareto $\mathrm{P}(k, 1)$ distributions meet (A6) with the powers 1, 1, $1 + \varepsilon$ (with arbitrary $\varepsilon > 0$), $1/2$ and $k/(k + 1)$ respectively

(see [17], Ex. 4.1 for the details). A special case, in which one may be interested in, is the case when $\mathrm{supp}(\mathbb{P}_X)$ is compact. In this case,

$$\mathbb{P}_X\left(p(X)<\varepsilon\right)=\int\limits_{\mathrm{supp}(\mathbb{P}_X)}\mathbb{1}\left(p(X)<\varepsilon\right)p(x)\mathrm{d}x\leqslant\varepsilon\int\limits_{\mathrm{supp}(\mathbb{P}_X)}\mathrm{d}x=\varepsilon\mu\left(\mathrm{supp}(\mathbb{P}_X)\right),$$

so (A6) is satisfied with $p=1$ and $C=\mu\left(\mathrm{supp}(\mathbb{P}_X)\right)$, where μ stands for the Lebesgue measure. In general, the assumption (A6) admits that $\mathbb{P}_X$ has an unbounded support. For this case, we give a simple sufficient condition to check (A6).

Proposition 2.1. *Let $X\in\mathbb{R}^d$ be such that $\mathbb{E}\|X\|^r<\infty$. Then X satisfies (A6) with $p=r/(r+d)$ and*

$$C=\left(\left(\frac{r}{d}\right)^{\frac{d}{r+d}}+\left(\frac{d}{r}\right)^{\frac{r}{r+d}}\right)\omega_d^{\frac{r}{r+d}}\left(\mathbb{E}\|X\|^r\right)^{\frac{d}{r+d}},$$

where ω_d stand for the Lebesgue measure of the unit ball in $\mathbb{R}^d$.

Proof. The proof of the proposition is straightforward:

$$\begin{aligned}\mathbb{P}_X\left(p(X)<\varepsilon\right)&=\int\limits_{\mathbb{R}^d}\mathbb{1}\left(p(X)<\varepsilon\right)p(x)\mathrm{d}x\\&=\int\limits_{x\in B(0,R)}\mathbb{1}\left(p(X)<\varepsilon\right)p(x)\mathrm{d}x+\int\limits_{x\notin B(0,R)}\mathbb{1}\left(p(X)<\varepsilon\right)p(x)\mathrm{d}x\\&\leqslant\varepsilon R^d\omega_d+\int\limits_{x\notin B(0,R)}\frac{\|x\|^r}{R^r}p(x)\mathrm{d}x\leqslant\varepsilon R^d\omega_d+\frac{\mathbb{E}\|X\|^r}{R^r}.\end{aligned}$$

Taking $R^{r+d}=r\mathbb{E}\|X\|^r/(d\varepsilon\omega_d)$ to minimize the expression in the right-hand side, we obtain

$$\mathbb{P}_X\left(p(X)<\varepsilon\right)\leqslant\left(\left(\frac{r}{d}\right)^{\frac{d}{r+d}}+\left(\frac{d}{r}\right)^{\frac{r}{r+d}}\right)\left(\omega_d\varepsilon\right)^{\frac{r}{r+d}}\left(\mathbb{E}\|X\|^r\right)^{\frac{d}{r+d}}.$$

□

In what is going further, we require p in (A6) to be larger than $\alpha/(2\alpha+d)$. By Proposition 2.1, any $\mathbb{P}_X$, such that $\mathbb{E}\|X\|^r<\infty$ for some $r>\alpha d/(\alpha+d)$, satisfies (A6) with $p>\alpha/(2\alpha+d)$.

3. The algorithm

In this section, we present the multiclass spatial stagewise aggregation (MSSA) procedure, which is precisely formulated in Algorithm 1. The procedure takes a sequence of integers $\{n_k:1\leqslant k\leqslant K\}$, which fulfills (A2), a training sample $S_n=\{(X_i,Y_i):1\leqslant i\leqslant n\}$, a test point $x\in\mathcal{X}$ and a set of positive numbers $\{z_k:1\leqslant k\leqslant K\}$. The numbers $z_1,\ldots,z_K$ will be referred to as critical values. This name is not occasional since the original spatial stagewise aggregation procedure is tightly related to hypothesis testing. More details can be found in [5]. It is important to mention that performance of the MSSA procedure crucially depends on a choice of the critical values z_k, $1\leqslant k\leqslant K$. At first glance, one can think that the problem of tuning of so large number of parameters is very time consuming and impracticable. However, in Section 5 with numerical experiments we provide a simple tuning procedure leading to a proper choice of the critical values.

Algorithm 1 Multiclass Spatial Stagewise Aggregation (MSSA)

Given a sequence of integers $\{n_k : 1 \leqslant k \leqslant K\}$ fulfilling (A2), a set of critical values$\{z_k : 1 \leqslant k \leqslant K\}$, a training sample $S_n = \{(X_i, Y_i) : 1 \leqslant i \leqslant n\}$ and a test point $x \in \mathcal{X}$, do the following:
for m **from** 1 **to** M **do**
  For each k from 1 to K compute the weights $w_i^{(k)} = w_i^{(k)}(X_i, x)$, $1 \leqslant i \leqslant n$, according to the formula (2.4)
  with a kernel $\mathscr{K}$ satisfying (A1) and calculate $\widetilde{\theta}_m^{(k)}(x)$ according to (2.5) and (2.6).
  Put $\widehat{\theta}_m^{(1)}(x) = \widetilde{\theta}_m^{(1)}(x)$.
  for k **from** 2 **to** K **do**
    Compute $N_k(x) = \sum_{i=1}^{n} w_i^{(k)}(X_i, x)$ and
    $\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right) = \widetilde{\theta}_m^{(k)}(x) \log \frac{\widetilde{\theta}_m^{(k)}(x)}{\widehat{\theta}_m^{(k-1)}(x)} + \left(1 - \widetilde{\theta}_m^{(k)}(x)\right) \log \frac{1-\widetilde{\theta}_m^{(k)}(x)}{1-\widehat{\theta}_m^{(k-1)}(x)}$.
    Find $\gamma_k = \mathbb{1}\left(N_k(x)\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right) \leqslant z_k\right)$.
    Update the estimate $\widehat{\theta}_m^{(k)}(x) = \gamma_k \widetilde{\theta}_m^{(k)}(x) + (1 - \gamma_k)\widehat{\theta}_m^{(k-1)}(x)$.
  end for Put the final estimate $\widehat{\theta}_m(x) = \widehat{\theta}_m^{(K)}(x)$.
end for
Return the predicted label $\widehat{f}(x) = \underset{1 \leqslant m \leqslant m}{\operatorname{argmax}} \left\{\widehat{\theta}_m(x)\right\}$.

We also emphasize that, by construction, $\widetilde{\theta}_m^{(k)}(x) \in [1/(2M), 1 - 1/(2M)]$ and, therefore, $\widehat{\theta}_m^{(k)}(x)$ also belongs to $[1/(2M), 1 - 1/(2M)]$ and $\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right)$ is defined correctly. In fact, $\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right)$ is nothing but the Kullback-Leibler divergence between two Bernoulli distributions with parameters $\widetilde{\theta}_m^{(k)}(x)$ and $\widehat{\theta}_m^{(k-1)}(x)$ respectively.

Concerning the computational time of the MSSA procedure, the assumption (A2) ensures that $K = O(\log n)$ and then it requires $O(Mn \log n)$ operations to compute nearest neighbor estimates for all classes and $O(\log n)$ operations to aggregate them. As a result, the computational time of the procedure, consumed for a prediction of the label of one test point, is $O(Mn \log n)$. If there are several test points, then the computations can be done in parallel.

4. Theoretical properties of the MSSA procedure

4.1. Main result

Theorem 4.1. *Let the conditions (A1)–(A5) hold and let (A6) hold with $p > \alpha/(2\alpha + d)$. Choose the parameters $z_1, \ldots, z_K$ according to the formula*

$$z_k = \frac{8M^2}{u_0} \log \frac{12KM}{\delta_*}, \quad 1 \leqslant k \leqslant K, \tag{4.1}$$

where

$$\delta_* = \begin{cases} \left(\frac{M^3 \log n}{np_0}\right)^{\frac{\alpha(2+\beta)}{2\alpha+d}}, & \text{if } \exists\, p_0 : p(x) \geqslant p_0 \,\forall\, x \in \mathit{supp}(\mathbb{P}_X), \\ \psi_*^{r_*}, & \text{otherwise}, \end{cases} \tag{4.2}$$

with $r_* = \log \psi_*^{-1}$ *and*

$$\psi_* = \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha}{\alpha\beta/p+2\alpha+d}}.$$

Then, if the sample size n *is sufficiently large, for the MSSA estimates* $\widehat{\theta}_1(\cdot), \dots, \widehat{\theta}_M(\cdot)$*, the excess risk of the plug-in classifier* $\widehat{f}(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \widehat{\theta}_m(X)$ *is bounded by*

$$\mathbb{E}_{S_n} \mathcal{E}(\widehat{f}) \lesssim \begin{cases} \left(\frac{M^3 \log n}{n p_0}\right)^{\frac{\alpha(1+\beta)}{2\alpha+d}}, & \text{if } \exists p_0 : p(x) \geqslant p_0 \, \forall x \in \mathit{supp}(\mathbb{P}_X), \\ \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha(1+\beta)}{\alpha\beta/p+2\alpha+d}}, & \text{otherwise.} \end{cases} \tag{4.3}$$

Moreover, for any $\delta \in (0,1)$*, if*

$$z_k = \frac{8M^2}{u_0} \log \frac{12KM}{\delta}, \quad 1 \leqslant k \leqslant K,$$

on an event with probability at least $(1-\delta)$ *over training samples, it holds*

$$\mathcal{E}(\widehat{f}) \leqslant \mathbb{P}_X(\widehat{f}(X) \neq f^*(X)) \lesssim \delta + \left(\frac{M^3 \log(12KM/\delta)}{n}\right)^{\frac{\alpha\beta}{\alpha\beta/p+(2\alpha+d)}}. \tag{4.4}$$

Here and further in the paper the relation $g(n) \lesssim h(n)$ means that there exists a universal constant $c > 0$ such that $g(n) \leqslant c h(n)$ for all $n \in \mathbb{N}$.

There are some comments we have. First, the rates (4.3) are optimal up to a logarithmic factor (see [4], Thm. 3.2 for the case of bounded away from zero density, [4], Thm. 4.1 for the case of bounded support (*i.e.* $p = 1$ in (A6)), and [17], Thm. 4.5 for the general case). Second, in the case of a bounded away from zero density one can take $p = \infty$. Then the inequality (4.4) transforms into

$$\mathbb{P}_X(\widehat{f}(X) \neq f^*(X)) \lesssim \delta + \left(\frac{M^3 \log(12KM/\delta)}{n}\right)^{\frac{\alpha\beta}{2\alpha+d}},$$

which revisits the result of Theorem 7 in [8].

4.2. Comparison with the nearest neighbor rule

Theorem 4.2. *Assume* (A1), (A3) *and* (A5). *Fix any* m, $1 \leqslant m \leqslant M$, *and a test point* $x \in \mathcal{X}$. *Then, for the weighted nearest neighbor estimate* $\widetilde{\eta}_m^w(x)$ *defined by* (2.2) *and* (2.3), *with probability at least* $(1-\delta)$ *over all training samples, it holds*

$$|\eta_m(x) - \widetilde{\eta}_m^w(x)| \leqslant \frac{L}{(n\varkappa p(x))^{\alpha/d}} \left(2\mathsf{k} + 4\log(2/\delta)\right)^{\alpha/d} + \sqrt{\frac{\log(4/\delta)}{\mathsf{k}}},$$

for any k *and* $\delta \in (0,1)$, *satisfying*

$$\left(\frac{2\mathsf{k} + 4\log(1/\delta)}{n\varkappa p(x)}\right)^{\alpha/d} \leqslant r_0.$$

The proof of this result is moved to Appendix C.1. The bound in Theorem 4.2 improves the result for the nearest neighbor regression obtained in [16] since it controls large deviations of $|\eta_m(x) - \widetilde{\eta}_m^w(x)|$ rather than its mean value. For the case of a bounded away from zero density, Theorem 4.2 and the union bound immediately yield

$$\mathbb{E}_{S_n}\mathbb{E}_X \max_{1\leqslant m\leqslant M} |\eta_m(X) - \widetilde{\eta}_m^w(X)|^r \lesssim \left(\frac{\mathsf{k}\log M}{n}\right)^{\alpha r/d} + \left(\frac{\log M}{\mathsf{k}}\right)^{r/2}$$

for any $r > 0$. This, together with Lemma A.4 below, implies a bound for the k-nearest neighbors classifier $\widehat{f}^{(\mathsf{k}-NN)}(x) = \operatorname*{argmax}_{1\leqslant m\leqslant M} \widetilde{\eta}_m^w(x)$:

$$\mathbb{E}_{S_n}\mathcal{E}\left(\widehat{f}^{(\mathsf{k}-NN)}(x)\right) \lesssim \left(\frac{\log M}{n}\right)^{\frac{\alpha(1+\beta)}{2\alpha+d}},$$

provided that $\mathsf{k} \asymp n^{2\alpha/(2\alpha+d)}$.

In the case of the bounded away from zero density, the nearest neighbor rule attains the minimax rate $n^{-(1+\beta)/(2\alpha+d)}$, while the MSSA classifier has an additional logarithmic factor. It can be easily explained by the fact that in the case $p(x) \geqslant p_0$, it is enough to take only one number of neighbors $\mathsf{k} \asymp n^{d/(2\alpha+d)}$ for all points $x \in \mathcal{X}$. At the same time, the MSSA procedure aggregates several nearest neighbor estimates and the factor $\log n$ can be considered as a payment for adaptation. Nevertheless, MSSA is capable to adapt to an unknown smoothness parameter $\alpha \in (0, 1]$ from the condition (A3), while the optimal choice of the smoothing parameter k of the classifier $\widehat{f}^{(\mathsf{k}-NN)}$ is based on the knowledge of α.

The situation is completely different in the case of a general density, fulfilling (A5) and (A6). In [17] (Thms. 4.3 and 4.5), it was shown that a universal choice of k for all points $x \in \mathcal{X}$ leads to a suboptimal rate $n^{-\frac{\alpha(1+\beta)}{\alpha(1+\beta)/p+2\alpha+d}}$, while Theorem 4.1 guarantees that the MSSA classifier has a minimax rate of convergence up to a logarithmic factor. It was also shown in [17] (Thms. 4.4 and 4.5) that a point-dependent choice $\mathsf{k}(x) \asymp (np(x))^{2\alpha/(2\alpha+d)}$ leads to the same rate $\big((\log n)/n\big)^{\frac{\alpha(1+\beta)}{\alpha\beta/p+2\alpha+d}}$, as for the MSSA classifier up to a logarithmic factor. However, it is not clear how to implement such a choice of k in practice, since a prior knowledge of the density $p(x)$ is required. Of course, one can try to estimate $p(x)$ but the density estimates are susceptible to the curse of dimensionality. In our turn, in Section 5, we describe a simple procedure of tuning parameters of MSSA. Moreover, by Theorem 3.1, the choice of critical values is the same for all test points, while the estimate of $p(x)$ must be recomputed at each test point x.

5. Numerical experiments

This section serves to illustrate the numerical performance of the proposed MSSA procedure on the artificial and real datasets. First we specify the choice of tuning parameters, then present the results.

5.1. Parameter tuning by propagation

Performance of the procedure critically depends on the choice of parameters z_k. We apply here the propagation approach originating from [31]. The basic idea is to ensure the desired properties of the method in a specific homogeneous situation. Let $x \in \mathcal{X}$ be a fixed test point. We generate artificial labels $\widetilde{Y}_1, \ldots, \widetilde{Y}_n$, which are sampled independently according to the distribution Bernoulli $\left(\frac{1}{2}\right)$. In this case, $\eta_1(x) = \mathbb{P}(Y = 1|X = x) \equiv 1/2$. Now, the proof of Lemma A.2 gives an insight, how to choose the critical values z_k: in the homogeneous situation $\eta_1(x) \equiv 1/2$, an event $\{\exists\, k : \widehat{\theta}_1^{(k)}(x) \neq \widetilde{\theta}_1^{(k)}(x)\}$ has to occur with a small probability. Such property of the MSSA procedure is called propagation. The preliminary critical values $\widetilde{z}_2, \ldots, \widetilde{z}_K$ are computed sequentially. Suppose that $\widetilde{z}_2, \ldots, \widetilde{z}_{k-1}$ have been already fixed for $k \geqslant 1$. This allows to compute $\widehat{\theta}_1^{(k-1)}(x)$ and the test statistic

TABLE 1. Information about artificial datasets. $\phi(\cdot, \mu, \Sigma)$ stands for the density of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.

	Experiment 1	Experiment 2	Experiment 3
Sample size, n	500	500	500
Number of classes, M	3	4	3
Prior class probabilities, π_m	1/3, 1/3, 1/3	1/4, 1/4, 1/4, 1/4	1/3, 1/3, 1/3
Class densities, $p_m(x)$	$p_1(x) = \phi(x, [0, -1], 0.5I_2)$, $p_2(x) = \phi\left(x, [\sqrt{3}/2, 0], 0.5I_2\right)$, $p_3(x) = \phi\left(x, [-\sqrt{3}/2, 0], 0.5I_2\right)$	$p_1(x) = \phi(x, [-1, -1], 0.7I_2)$, $p_2(x) = \phi(x, [1, -1], 0.7I_2)$, $p_3(x) = \phi(x, [-1, 1], 0.7I_2)$, $p_4(x) = \phi(x, [1, 1], 0.7I_2)$	$p_1(x) = 0.5\phi(x, [-1, 0], 0.5I_2) + 0.5\phi(x, [1, 0], 0.5I_2)$, $p_2(x) = 0.5\phi\left(x, [0.5, \sqrt{3}/2], 0.5I_2\right) + 0.5\phi\left(x, [-0.5, -\sqrt{3}/2], 0.5I_2\right)$, $p_3(x) = 0.5\phi\left(x, [-0.5, \sqrt{3}/2], 0.5I_2\right) + 0.5\phi\left(x, [0.5, -\sqrt{3}/2], 0.5I_2\right)$
Number of neighbors, n_k	$n_k = \lfloor 3 \cdot 1.25^k \rfloor$, $0 \leqslant k \leqslant 11$	$n_k = \lfloor 3 \cdot 1.25^k \rfloor$, $0 \leqslant k \leqslant 15$	$n_k = \lfloor 3 \cdot 1.25^k \rfloor$, $0 \leqslant k \leqslant 14$
Localization kernel, $\mathscr{K}(t)$	Rectangular, $\mathscr{K}(t) = \mathbb{1}(0 \leqslant t \leqslant 1)$	Rectangular, $\mathscr{K}(t) = \mathbb{1}(0 \leqslant t \leqslant 1)$	Rectangular, $\mathscr{K}(t) = \mathbb{1}(0 \leqslant t \leqslant 1)$

$T_k = N_k(x)\mathcal{K}\left(\widehat{\theta}_1^{(k)}(x), \widehat{\theta}_1^{(k-1)}(x)\right)$. Now $\widetilde{z}_k$ is defined from the condition

$$\mathbb{P}\left(T_k > \widetilde{z}_k \,\middle|\, T_2 \leqslant \widetilde{z}_2, \dots T_{k-1} \leqslant \widetilde{z}_{k-1}\right) \leqslant \delta/K.$$

for some small $\delta \in (0, 1)$, *e.g.* $\delta = 0.1$. This condition is checked by the Monte-Carlo simulations for the artificial dataset $\widetilde{S}_n = \{(X_i, \widetilde{Y}_i) : 1 \leqslant i \leqslant n\}$. After that we choose the critical values $z_1, \dots, z_K$ in the form $z_k = c\widetilde{z}_k$ for all k from 1 to K. The constant c is chosen by cross validation. The Monte-Carlo simulations are performed only for one test point, because, due to Theorem 4.1, the choice of z_k's is universal for all test points.

5.2. Experiments on artificial datasets

We start with presenting the performance of MSSA on artificial datasets. We generate points from a mixture model: $p(x|Y = m) = p_m(x)$, $\mathbb{P}(Y = m) = \pi_m$. Then the density of X is given by $p(x) = \sum_{m=1}^{M} \pi_m p_m(x)$, and the Bayes rule is defined as $f^*(X) = \operatorname{argmax}_{1 \leqslant m \leqslant M} \pi_m p_m(x)$. We provide results for three different experiments. The information about them is summarized in Table 1 and sample realizations are displayed in Figure 1. For example, in the first experiment, the sample consists of $n = 500$ points, each of them belongs to one of $M = 3$ classes, and the prior class probabilities π_m, $1 \leqslant m \leqslant 3$, are equal to 1/3. Class densities $p_1(x)$, $p_2(x)$ and $p_3(x)$ were taken $\phi(x, [0, -1], 0.5I_2)$, $\phi\left(x, [\sqrt{3}/2, 0], 0.5I_2\right)$ and $\phi\left(x, [-\sqrt{3}/2, 0], 0.5I_2\right)$ respectively, where $\phi(x, \mu, \Sigma)$ stands for the density of a Gaussian random vector with the mean μ and the variance Σ. Next, we took the sequence of integers $n_k = \lfloor 3 \times 1.25^k \rfloor$, $0 \leqslant k \leqslant 11$, and considered n_k-nearest-neighbors estimates with the rectangular kernel $\mathscr{K}(t) = \mathbb{1}(0 \leqslant t \leqslant 1)$. We computed leave-one-out cross-validation errors for the MSSA classifier and all n_k-nearest neighbors classifiers. The second and the third experiments on artificial datasets were carried out in the same way. The results, which are shown on Figure 2, indicate that even the best n_k-nearest neighbors classifier is outperformed by the properly tuned MSSA classifier.

5.3. Experiments on the real datasets

We proceed with experiments on datasets from the UCI repository [12]: Ecoli, Iris, Glass, Pendigits, Satimage, Seeds, Wine and Yeast. Short information about these datasets is given in Table 2.

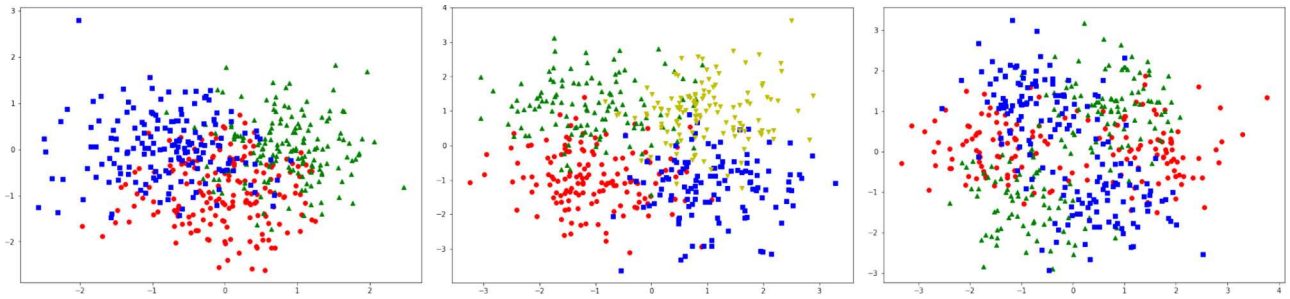

FIGURE 1. Sample realizations in the first (left, $M = 3$ classes, $n = 500$ points), the second (center, $M = 4$ classes, $n = 500$ points) and the third (right, $M = 3$ classes, $n = 500$ points) experiments with artificial datasets.

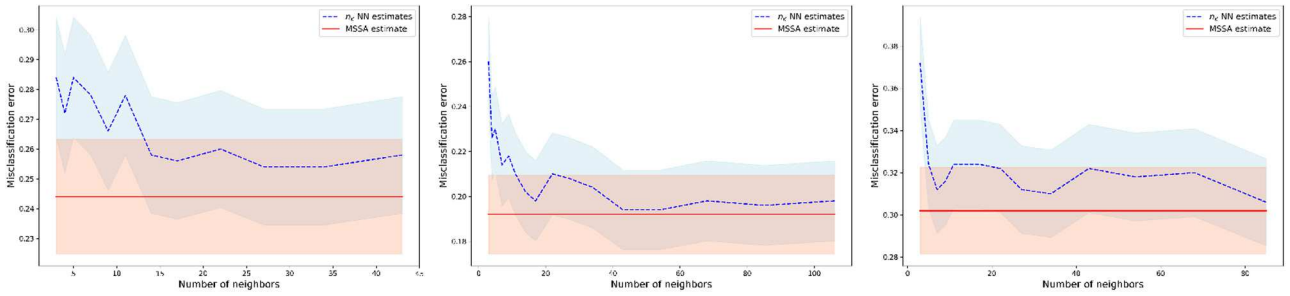

FIGURE 2. Leave-one-out cross-validation errors for the weighted nearest-neighbor classifiers (the dashed line) in the first (left), the second (center) and the third (right) experiments. The solid line corresponds to the LOO CV error of the MSSA classifier. The shaded regions reflect standard deviations.

TABLE 2. Information about datasets from the UCI repository [12].

Dataset	Train	Test	Attributes	Classes	Class distribution (in %)
Ecoli	336	–	7	8	42.6, 22.9, 15.5, 10.4, 5.9, 1.5, 0.6, 0.6
Iris	150	–	4	3	33.3, 33.3, 33.3
Glass	214	–	9	6	32.7, 35.5, 7.9, 6.1, 4.2, 13.6
Pendigits	7494	3498	16	10	10.4, 10.4, 10.4, 9.6, 10.4, 9.6, 9.6, 9.6, 10.4, 9.6, 9.6
Satimage	4435	2000	36	6	24.1, 11.1, 20.3, 9.7, 11.1, 23.7
Seeds	210	–	7	3	33.3, 33.3, 33.3
Wine	178	–	13	3	33.1, 39.8, 26.9
Yeast	1484	–	8	10	16.4, 28.1, 31.2, 2.9, 2.3, 3.4, 10.1, 2.0, 1.3, 0.3

We compare the performance of the MSSA algorithm with the oracle choice of the nearest neighbor estimate. For Pendigits and Satimage datasets, we calculated misclassification error on the test dataset, for all other datasets we used leave-one-out cross-validation. Results of our experiments are shown in Table 3, best ones are boldfaced. From Table 3, one can observe that MSSA works as good as -NN rule with the best choice of nearest neigbors and even slightly outperforms ordinary nearest neighbor rule in most situations.

TABLE 3. Leave-one-out cross-validation errors (in %) with standard deviations for datasets from the UCI repository. The best results are boldfaced.

Dataset	MSSA	Best nearest neighbor classifier
Ecoli	$\mathbf{12.8 \pm 1.8}$	13.4 ± 1.9
Iris	0.0 ± 0.0	0.0 ± 0.0
Glass	$\mathbf{27.6 \pm 3.0}$	28.0 ± 3.0
Pendigits	2.2 ± 0.2	2.2 ± 0.2
Satimage	$\mathbf{9.3 \pm 0.6}$	9.6 ± 0.7
Seeds	6.7 ± 1.7	6.7 ± 1.7
Wine	2.2 ± 1.1	2.2 ± 1.1
Yeast	$\mathbf{39.6 \pm 1.3}$	39.8 ± 1.3

APPENDIX A. PROOF OF THEOREM 4.1

The proof of Theorem 4.1 is divided into several steps. On the first one, we discuss nice properties of the MSSA estimates $\widehat{\theta}_m(x)$, $1 \leqslant m \leqslant M$. Next, we focus on the MSSA plug-in classifier $\widehat{f}(x) = \operatorname{argmax}_{1 \leqslant m \leqslant M} \widehat{\theta}_m(x)$. In Section A.2, we study the case of a bounded away from zero density and prove the first part of the upper bound (4.3) for the mean excess risk $\mathbb{E}_{S_n}\mathcal{E}(\widehat{f})$. Then, in Section A.3, we extend our analysis to the case of a general density $p(x)$, which fulfils the minimal mass assumption (A5) and the tail assumption (A6). Finally, in Section A.4, we obtain the bound (4.4) on the excess risk $\mathcal{E}(\widehat{f})$, which holds on an event with high probability.

A.1 Step 1: pointwise guarantees for MSSA estimates

Theorem 4.2, the union bound and 1-Lipschitzness of the function $\varphi(\cdot)$ immediately yield

Corollary A.1. *Under assumptions of Theorem 4.2, we have*

$$|\theta_m(x) - \widetilde{\theta}_m^{(k)}(x)| \leqslant \frac{L}{(n\varkappa p(x))^{\alpha/d}} \big(2n_k + 4\log(6KM/\delta)\big)^{\alpha/d} + \sqrt{\frac{\log(12KM/\delta)}{n_k}}$$

simultaneously for all $1 \leqslant m \leqslant M$ *and* $1 \leqslant k \leqslant K$ *on an event with probability at least* $1 - \delta/3$.

Next, the MSSA procedure comes into the play. Denote

$$\overline{\eta}_m^{(k)}(x) = \frac{1}{N_k(x)} \sum_{i=1}^n w_i^{(k)}(X_i, x)\eta_m(X_i),$$

where $N_k(x) = \sum_{i=1}^n w_i^{(k)}(X_i, x)$, and for any $\delta \in (0,1)$ and any $x \in \mathcal{X}$ define

$$k^* = k^*(\delta, x) = \max\Big\{k' : |\overline{\eta}_m^{(k)}(x) - \overline{\eta}_m^{(k-1)}(x)| \leqslant \sqrt{\frac{2\log(12KM/\delta)}{u_0 N_k(x)}} \quad \forall\, 1 \leqslant m \leqslant M, \forall\, 2 \leqslant k \leqslant k'\Big\}. \tag{A.1}$$

We call the set $\{k : 1 \leqslant k \leqslant k^*\}$ *the small bias region*. In this region, MSSA has the following oracle property.

Lemma A.2. *Let (A1) and (A2) be fulfilled. Fix any $\delta \in (0,1)$ and $x \in \mathcal{X}$ and choose*

$$z_k = \frac{8M^2}{u_0} \log \frac{12KM}{\delta}. \tag{A.2}$$

Then there exists a universal constant C_1 (depending only on u_0 and u from (A2)) such that, with probability at least $1-\delta/3$ over training samples, it holds

$$|\widehat{\theta}_m(x) - \widetilde{\theta}_m^{(k)}(x)| \leqslant C_1 M^{3/2} \sqrt{\frac{\log(12KM/\delta)}{n_k}}$$

simultaneously for all $1 \leqslant m \leqslant M$ and $1 \leqslant k \leqslant k^*$ with $k^* = k^*(\delta, x)$ given by (A.1).

The proof of Lemma A.2 is given in Appendix C.2. A natural question arises: how large is the small bias region? The answer is given in the following lemma.

Lemma A.3. *Assume (A1), (A2) and (A5). Fix any $x \in \mathcal{X}$ and $\delta \in (0,1)$. Then*

$$n_{k^*} = n_{k^*(\delta,x)} \gtrsim \left((n \varkappa p(x))^{2\alpha/(2\alpha+d)} \left(\log(12KM/\delta)\right)^{d/(2\alpha+d)} \right) \vee \log(12KM/\delta)$$

with probability at least $(1-\delta/3)$ over training samples.

The proof is moved to Appendix C.3. We will show later that an optimal value of n_k is less than n_{k^*}, so the MSSA classifier enjoys a minimax rate of convergence up to a logarithmic factor.

A.2 Step 2: the case of a bounded away from zero density

Corollary A.1, Lemma A.3, and Lemma A.2 imply that, given $x \in \mathcal{X}$, with probability at least $1-\delta$, simultaneously for all m, $1 \leqslant m \leqslant M$, and $k \leqslant k^* = k^*(\delta, x)$ (*i.e.* $n_k \lesssim (n \varkappa p(x))^{2\alpha/(2\alpha+d)} \left(\log(12KM/\delta)\right)^{d/(2\alpha+d)} \vee \log(12KM/\delta)$), it holds

$$\begin{aligned} |\widehat{\theta}_m(x) - \theta_m(x)| \leqslant{}& \frac{L}{(n\varkappa p(x))^{\alpha/d}} \left(2n_k + 4\log(6KM/\delta)\right)^{\alpha/d} \\ &+ \sqrt{\frac{\log(12KM/\delta)}{n_k}} + C_1 M^{3/2} \sqrt{\frac{\log(12KM/\delta)}{n_k}}. \end{aligned} \tag{A.3}$$

Let $r = 2+\beta$. Since $|\widehat{\theta}_m(x) - \theta_m(x)| \leqslant 1$ almost surely, the expectation of $\max\limits_{1 \leqslant m \leqslant M} |\widehat{\theta}_m(x) - \theta_m(x)|^r$ with respect to training samples can be bounded by

$$\begin{aligned} &\mathbb{E}_{S_n} \max_{1 \leqslant m \leqslant M} |\widehat{\theta}_m(x) - \theta_m(x)|^r \\ &\leqslant \delta + \min_{1 \leqslant k \leqslant k^*} \left[\frac{L}{(n\varkappa p(x))^{\alpha/d}} \left(2n_k + 4\log(6KM/\delta)\right)^{\alpha/d} \right. \\ &\left. + \sqrt{\frac{\log(12KM/\delta)}{n_k}} + C_1 M^{3/2} \sqrt{\frac{\log(12KM/\delta)}{n_k}} \right]^r. \end{aligned} \tag{A.4}$$

Choose any $k \leqslant k^*$, fulfilling

$$n_k \asymp M^{\frac{3d}{2\alpha+d}} (\varkappa np(x))^{\frac{2\alpha}{2\alpha+d}} (\log(12KM/\delta))^{\frac{d}{2\alpha+d}}.$$

Existence of such k is guaranteed by Lemma A.3. Here and further in this paper, $g(n) \asymp h(n)$ means $g(n) \lesssim h(n) \lesssim g(n)$. Then we have

$$\mathbb{E}_{S_n} \max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(x) - \theta_m(x)|^r \lesssim \delta + \left(\frac{M^3 \log(12KM/\delta)}{np(x)}\right)^{\alpha r/(2\alpha+d)}. \tag{A.5}$$

In the case, when there exists $p_0 > 0$, such that $p(x) \geqslant p_0$ for all $x \in \operatorname{supp}(\mathbb{P}_X)$, taking $\delta = \delta_*$ according to (4.2), we obtain

$$\mathbb{E}_{S_n} \max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(x) - \theta_m(x)|^r \lesssim \left(\frac{M^3 \log n}{np_0}\right)^{\alpha r/(2\alpha+d)}. \tag{A.6}$$

Here we used the fact that, due to (A2), $K \lesssim \log n$. Thus, $\log(12KM/\delta) \lesssim \log\log n + \log M + \log n \lesssim \log n$. The next lemma helps to transform the bound on moments (A.6) into the bound on the mean excess risk $\mathbb{E}_{S_n}(\widehat{f})$ of the classifier $\widehat{f}$.

Lemma A.4. *Let the low noise condition (A4) be fulfilled. Let $\widehat{\theta}_m(x)$ be any estimator of $\theta_m(x)$ at the point $x \in \mathcal{X}$. Suppose that for some $r > 1+\beta$, for all m from 1 to M and for almost all x with respect to $\mathbb{P}_X$, it holds*

$$\mathbb{E}_{S_n} \max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(x) - \theta_m(x)|^r \leqslant \chi_r,$$

with a function χ_r, which does not depend on x. Denote a plug-in classifier, associated with the estimates $\widehat{\theta}_1(x), \dots, \widehat{\theta}_M(x)$, by $\widehat{f}(x) = \operatorname{argmax}_{1\leqslant m\leqslant M} \widehat{\theta}_m(x)$. Then for the excess risk $\mathcal{E}(\widehat{f})$ it holds

$$\mathbb{E}_{S_n} \mathcal{E}(\widehat{f}) \leqslant B\left(1 + \frac{6(r+\beta+2)}{r-\beta-1}\right) \chi_r^{\frac{1+\beta}{r}}.$$

Proof of Lemma A.4 is given in Appendix C.4. The inequality (A.6) and Lemma A.4 immediately yield

$$\mathbb{E}_{S_n} \mathcal{E}(\widehat{f}) \lesssim \left(\frac{M^3 \log n}{np_0}\right)^{\alpha(1+\beta)/(2\alpha+d)},$$

which finishes the proof of the first part of the bound (4.3).

A.3 Step 3: extension to the case of a general density

Now, consider a density $p(x)$, which fulfils (A5) and (A6). Let

$$p_* = \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha\beta}{\alpha\beta+p(2\alpha+d)}}. \tag{A.7}$$

Choose $\delta_* = \psi_*^{r_*}$, $r_* = \log \psi_*^{-1}$, and

$$\psi_* = \left(\frac{M^3 \log^2 n}{np_*}\right)^{\alpha/(2\alpha+d)} = \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha}{\alpha\beta/p+2\alpha+d}}.$$

Define events $B_0 = \{p(X) \geqslant p_*\}$ and $B_j = \{2^{-j} p_* \leqslant p(X) < 2^{-j+1} p_*\}$, $1 \leqslant j \leqslant J$, where

$$J = \left\lfloor \frac{\log \psi_*^{-1} - 1}{\alpha/(2\alpha+d)\log 2} \right\rfloor.$$

Note that such choice of J implies

$$2^{j\alpha/(2\alpha+d)} \psi_* \leqslant e^{-1} \quad \forall\, j \leqslant J. \tag{A.8}$$

Then, using Lemma B.2, we have

$$\begin{aligned}
&\mathbb{E}_{S_n} \mathcal{E}(\widehat{f}) = \mathbb{E}_{S_n} \mathbb{E}_X \left(\eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X)\right) \leqslant 2\mathbb{E}_{S_n} \mathbb{E}_X \left(\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right) \\
&= 2\mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \left(\mathbb{1}(B_0) + \sum_{j=1}^{J} \mathbb{1}(B_j) + \mathbb{1}(B_{J+1})\right) \\
&\leqslant 2\mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \mathbb{1}(B_0) \\
&\quad + 2\sum_{j=1}^{J} \mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \mathbb{1}(B_j) + 2\mathbb{P}_X(B_{J+1}).
\end{aligned}$$

For the latter term $\mathbb{P}_X(B_{J+1})$ we simply have

$$\mathbb{P}_X(B_{J+1}) \leqslant C 2^{-Jp} p_*^p. \tag{A.9}$$

Consider $\mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \mathbb{1}(B_0)$. On B_0 we have $p \geqslant p_*$ and, again, applying the argument we used in the case of a bounded away from zero density, we obtain

$$\mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \mathbb{1}(B_0) \lesssim \psi_*^{1+\beta} = \left(\frac{M^3 \log^2 n}{np_*}\right)^{\alpha(1+\beta)/(2\alpha+d)}. \tag{A.10}$$

Now, consider $\mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \mathbb{1}(B_j)$, $j \in \mathbb{N}$. Let $\{t_j : j \in \mathbb{N}\}$ be a sequence of integers, which will be specified later. Then

$$\begin{aligned}
&\mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \mathbb{1}(B_j) \\
&= \mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(0 < \theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X) < 2t_j\right) \mathbb{1}(B_j) \\
&+ \mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right] \mathbb{1}\left(\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X) \geqslant 2t_j\right) \mathbb{1}(B_j).
\end{aligned} \tag{A.11}$$

Due to the tail assumption (A6), we have

$$\begin{aligned}&\mathbb{E}_X\mathbb{E}_{S_n}\left[\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)\right]\mathbb{1}\left(0<\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)<2t_j\right)\mathbb{1}(B_j)\\&\leqslant 2t_j\mathbb{P}(B_j)\leqslant 2Ct_j\left(2^{-j+1}p_*\right)^p.\end{aligned}\tag{A.12}$$

For the second term, again, using the inequality

$$\begin{aligned}&\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)=\left(\theta_{f^*(X)}(X)-\widehat{\theta}_{f^*(X)}(X)\right)\\&+\left(\widehat{\theta}_{f^*(X)}(X)-\widehat{\theta}_{\widehat{f}(X)}(X)\right)+\left(\widehat{\theta}_{\widehat{f}(X)}(X)-\theta_{\widehat{f}(X)}(X)\right)\\&\leqslant\left(\theta_{f^*(X)}(X)-\widehat{\theta}_{f^*(X)}(X)\right)+\left(\widehat{\theta}_{\widehat{f}(X)}(X)-\theta_{\widehat{f}(X)}(X)\right)\\&\leqslant 2\max_{1\leqslant m\leqslant M}|\widehat{\theta}_m(X)-\theta_m(X)|,\end{aligned}$$

one obtains

$$\begin{aligned}&\mathbb{E}_X\mathbb{E}_{S_n}\left[\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)\right]\mathbb{1}\left(\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)\geqslant 2t_j\right)\mathbb{1}(B_j)\\&\leqslant\mathbb{E}_X\mathbb{E}_{S_n}\left[\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)\right]\mathbb{1}\left(\max_{1\leqslant m\leqslant M}|\widehat{\theta}_m(X)-\theta_m(X)|\geqslant t_j\right)\mathbb{1}(B_j)\\&\leqslant\mathbb{E}_X\mathbb{E}_{S_n}\mathbb{1}\left(\max_{1\leqslant m\leqslant M}|\widehat{\theta}_m(X)-\theta_m(X)|\geqslant t_j\right)\mathbb{1}(B_j).\end{aligned}$$

With probability at least $1-\delta_*$ we have

$$\max_{1\leqslant m\leqslant M}|\widehat{\theta}_m(X)-\theta_m(X)|\mathbb{1}(B_j)\lesssim\left(\frac{M^3\log(12KM/\delta_*)}{2^{-j}np_*}\right)^{\alpha/(2\alpha+d)}\lesssim 2^{j\alpha/(2\alpha+d)}\psi_*.$$

Here we used that, due to (A2), $K\lesssim\log n$, so

$$\log(12KM/\delta_*)\lesssim\log\log n+\log M+\log^2\psi_*^{-1}\lesssim\log^2 n.$$

Denote $\psi_j=2^{j\alpha/(2\alpha+d)}\psi_*$. The Markov inequality yields

$$\mathbb{E}_{S_n}\mathbb{1}\left(\max_{1\leqslant m\leqslant M}|\widehat{\theta}_m(X)-\theta_m(X)|\geqslant t_j\right)\mathbb{1}(B_j)\leqslant\frac{\delta_*+\psi_j^{r_*}}{t_j^{r_*}}=\frac{\psi_*^{r_*}+\psi_j^{r_*}}{t_j^{r_*}}\leqslant\frac{2\psi_j^{r_*}}{t_j^{r_*}},$$

and therefore,

$$\mathbb{E}_X\mathbb{E}_{S_n}\left[\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)\right]\mathbb{1}\left(\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)\geqslant 2t_j\right)\mathbb{1}(B_j)\leqslant 2C\left(2^{-j+1}p_*\right)^p\frac{\psi_j^{r_*}}{t_j^{r_*}}.\tag{A.13}$$

Thus, taking (A.11)–(A.13) together, one obtains

$$\mathbb{E}_X\mathbb{E}_{S_n}\left[\theta_{f^*(X)}(X)-\theta_{\widehat{f}(X)}(X)\right]\mathbb{1}\left(f^*(X)\neq\widehat{f}(X)\right)\mathbb{1}(B_j)\leqslant 2C\left(2^{-j+1}p_*\right)^p\left(t_j+\frac{\psi_j^{r_*}}{t_j^{r_*}}\right).$$

Take $t_j = \psi_j^{r_*}$ to balance the two terms. Then

$$\mathbb{E}_X \mathbb{E}_{S_n} \left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X) \right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X) \right) \mathbb{1}(B_j) \leqslant 4C \left(2^{-j+1} p_*\right)^p \psi_j^{r_*/(r_*+1)}. \tag{A.14}$$

Note that

$$\psi_j^{r_*/(r_*+1)} = 2^{\frac{\alpha j r_*}{(2\alpha+d)(r_*+1)}} \psi_*^{r_*/(r_*+1)} \leqslant 2^{\alpha j/(2\alpha+d)} \cdot e\psi_*. \tag{A.15}$$

The last inequality follows from the fact

$$\psi_*^{r_*/(r_*+1)} = \psi_* \cdot \psi_*^{-1/(r_*+1)} = \psi_* e^{\log \psi_*^{-1}/(r_*+1)} = \psi_* e^{\log \psi_*^{-1}/(\log \psi_*^{-1}+1)} \leqslant e\psi_*.$$

Inequalities (A.9), (A.10), (A.14) and (A.15) immediately imply

$$\begin{aligned}
\mathbb{E}_{S_n} \mathcal{E}(\widehat{f}) &= \mathbb{E}_X \mathbb{E}_{S_n} \left(\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X) \right) \lesssim \psi_*^{1+\beta} + \sum_{j=1}^{J} \left(2^{-j+1} p_*\right)^p 2^{\alpha j/(2\alpha+d)} \psi_* + \left(2^{-J} p_*\right)^p \\
&\leqslant \psi_*^{1+\beta} + 2^p p_*^p \psi_* \sum_{j=1}^{\infty} 2^{-j(p-\alpha/(2\alpha+d)} + \left(2^{-J} p_*\right)^p = \psi_*^{1+\beta} + \frac{2^p p_*^p \psi_* \cdot 2^{-p+\alpha/(2\alpha+d)}}{1 - 2^{-p+\alpha/(2\alpha+d)}} + \left(2^{-J} p_*\right)^p \\
&= \psi_*^{1+\beta} + \frac{p_*^p \psi_* \cdot 2^{\alpha/(2\alpha+d)}}{1 - 2^{-p+\alpha/(2\alpha+d)}} + \left(2^{-J} p_*\right)^p .
\end{aligned} \tag{A.16}$$

Note that the density level p_*, defined by (A.7), balances the first and the second terms. We also have $p_* \leqslant \varepsilon_0$ from (A6), provided that n is sufficiently large. Concerning the third term in (A.16), we have

$$2^{-J-1} p_* \leqslant (e\psi_*)^{(2\alpha+d)/\alpha} p_* = \frac{eM^3 \log^2 n}{n},$$

and, since $p > \alpha/(2\alpha + d)$, it holds

$$p > \frac{\alpha(1+\beta)}{\alpha\beta/p + 2\alpha + d},$$

and with such choice of p_*

$$\left(\frac{M^3 \log^2 n}{n}\right)^p < \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha(1+\beta)}{\alpha\beta/p+2\alpha+d}} = \left(\frac{M^3 \log^2 n}{n p_*}\right)^{\alpha(1+\beta)/(2\alpha+d)} = \psi_*^{1+\beta},$$

so the third term in (A.16) is smaller than the first and the second terms. Finally,

$$\mathbb{E}_{S_n} \mathcal{E}(\widehat{f}) \lesssim \psi_*^{1+\beta} = \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha(1+\beta)}{\alpha\beta/p+2\alpha+d}}.$$

A.4 Step 4: a bound on the excess risk with high probability

In (A.3), taking any $n_k \leqslant n_{k^*(\delta,X)}$, fulfilling

$$n_k \asymp M^{\frac{3d}{2\alpha+d}} (np(X))^{\frac{2\alpha}{2\alpha+d}} \left(\log \frac{12KM}{\delta}\right)^{\frac{d}{2\alpha+d}},$$

we have

$$\begin{aligned}\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X) &\leqslant 2 \max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(X) - \theta_m(X)| \\ &\leqslant C_2 \left(\frac{M^3 \log(12KM/\delta)}{np(X)}\right)^{\frac{\alpha}{2\alpha+d}}\end{aligned}$$

with probability at least $1-\delta$ over training samples.

Consider an event $\{p(X) \geqslant p_*\}$ with $p_* \asymp \left(\frac{\log(12KM/\delta)}{n}\right)^{\frac{\alpha\beta}{\alpha\beta+p(2\alpha+d)}}$. On this event $f^*(X) \neq \widehat{f}(X)$ only if $\theta_{(1)}(X) - \theta_{(2)}(X) < C_2 \left(\frac{M^3 \log(12KM/\delta)}{np_*}\right)^{\frac{\alpha}{2\alpha+d}}$ or if $\max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(X) - \theta_m(X)| \geqslant \frac{C_2}{2} \left(\frac{M^3 \log(12KM/\delta)}{np_*}\right)^{\frac{\alpha}{2\alpha+d}}$. This yields,

$$\begin{aligned}&\mathbb{P}_X\left(f^*(X) \neq \widehat{f}(X),\, p(X) \geqslant p_*\right) \\ &\leqslant \mathbb{P}_X\left(\theta_{(1)}(X) - \theta_{(2)}(X) < C_2 \left(\frac{M^3 \log(12KM/\delta)}{np_*}\right)^{\frac{\alpha}{2\alpha+d}}\right) \\ &+ \mathbb{P}\left(\max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(X) - \theta_m(X)| > \frac{C_2}{2} \left(\frac{M^3 \log(12KM/\delta)}{np_*}\right)^{\frac{\alpha}{2\alpha+d}}\right) \\ &\leqslant \delta + BC_2^\beta \left(\frac{M^3 \log(12KM/\delta)}{np_*}\right)^{\frac{\alpha\beta}{2\alpha+d}}.\end{aligned}$$

Then, the probability of an incorrect prediction can be bounded by

$$\begin{aligned}\mathbb{P}_X\left(f^*(X) \neq \widehat{f}(X)\right) &= \mathbb{P}_X\left(f^*(X) \neq \widehat{f}(X),\, p(X) \geqslant p_*\right) \\ &\quad + \mathbb{P}_X\left(f^*(X) \neq \widehat{f}(X) \,\middle|\, p(X) < p_*\right) \mathbb{P}_X\left(p(X) < p_*\right) \\ &\leqslant \delta + BC_2^\beta \left(\frac{M^3 \log(12KM/\delta)}{np_*}\right)^{\frac{\alpha\beta}{2\alpha+d}} + Cp_*^p \\ &\lesssim \delta + \left(\frac{M^3 \log(12KM/\delta)}{n}\right)^{\frac{\alpha\beta}{\alpha\beta/p+(2\alpha+d)}},\end{aligned}$$

and it finishes the proof of Theorem 4.1.

Appendix B. Auxiliary results

Lemma B.1. *Denote the Kullback-Leibler divergence between distributions Bernoulli(ϑ) and Bernoulli(ϑ') by $\mathcal{K}(\vartheta,\vartheta')$. Then for any $\vartheta,\vartheta' \in [1/(2M), 1-1/(2M)]$ it holds*

$$\frac{3}{M}(\vartheta-\vartheta')^2 \leqslant \mathcal{K}(\vartheta,\vartheta') \leqslant M^2(\vartheta-\vartheta')^2.$$

Proof. The proof relies on some properties of the exponential family of distributions. For a random variable $Z \sim \text{Bernoulli}(\vartheta)$ the log-density $\log p(z,\vartheta)$ can be written in the following form

$$\log p(z,\vartheta) = z \log \frac{\vartheta}{1-\vartheta} + \log(1-\vartheta).$$

Denote $\nu = \nu(\vartheta) = \log \frac{\vartheta}{1-\vartheta}$ and $D(\nu) = \log(1+e^{\nu})$. ν is called a canonical parameter of the Bernoulli distribution. The direct computation shows that

$$\begin{aligned}
&\log p(z,\vartheta) = z\nu(\vartheta) - D(\nu(\vartheta)),\\
&\vartheta \equiv D'(\nu(\vartheta)),\\
&\text{Var}(Z) = D''(\nu(\vartheta)),\\
&\mathcal{K}(\vartheta,\vartheta') = D'(\nu)(\nu-\nu') - D(\nu) + D(\nu') = \frac{D''(\xi)}{2}(\nu-\nu')^2.
\end{aligned}$$

In the last formula we used a notation $\nu = \nu(\vartheta)$, $\nu' = \nu(\vartheta')$ and ξ is a number between ν and ν'. The Lagrange theorem yields

$$\vartheta - \vartheta' = D'(\nu) - D'(\nu') = D''(\zeta)(\nu-\nu'),$$

for some ζ between ϑ and ϑ'. Thus, we obtain the equality

$$\mathcal{K}(\vartheta,\vartheta') = \frac{D''(\xi)}{2\left(D''(\zeta)\right)^2}(\vartheta-\vartheta')^2,$$

which implies

$$\frac{D_0}{2D_1^2}(\vartheta-\vartheta')^2 \leqslant \mathcal{K}(\vartheta,\vartheta') \leqslant \frac{D_1}{2D_0^2}(\vartheta-\vartheta')^2$$

with $D_0 = \min\limits_{\xi\in[\nu,\nu']} D''(\xi)$ and $D_1 = \max\limits_{\xi\in[\nu,\nu']} D''(\xi)$.

Now, we use the formula $\text{Var}(Z) = D''(\nu(\vartheta))$ and obtain

$$D''(\nu(\vartheta)) = \vartheta(1-\vartheta).$$

If $(\vartheta,\vartheta') \in [1/(2M), 1-1/(2M)]^2$ then, taking into account the fact that $M \geqslant 2$, one has $D_0 = 1/(2M)\big(1-1/(2M)\big)$, $D_1 = 1/4$ and

$$\frac{D_0}{2D_1^2} = \frac{4}{M}\left(1-\frac{1}{2M}\right) \geqslant \frac{3}{M},$$

$$\frac{D_1}{2D_0^2} = \frac{M^2}{2}\left(1 - \frac{1}{2M}\right)^{-2} \leqslant M^2,$$

and the proof of Lemma B.1 is finished. □

Lemma B.2. *Fix a point $x \in \mathcal{X}$ and denote $m^* \in \operatorname*{argmax}_{1 \leqslant m \leqslant M} \eta_m(x)$. Then for any $m \neq m^*$ it holds*

$$\eta_{m^*}(x) - \eta_m(x) \leqslant 2\left(\theta_{m^*}(x) - \theta_m(x)\right).$$

Proof. There are three cases we have to consider: (i) $\eta_{m^*}(x) > 1 - 1/(2M)$, (ii) $\eta_{m^*}(x) \leqslant 1 - 1/(2M)$, $\eta_m(x) > 1/(2M)$ and (iii) $\eta_{m^*}(x) \leqslant 1 - 1/(2M)$, $\eta_m(x) \leqslant 1/(2M)$.

Consider the case (i). Note that in this case the condition $\eta_{m^*}(x) > 1 - 1/(2M)$ immediately yields $\eta_m(x) < 1/(2M)$ for all $m \neq m^*$. Then $\theta_{m^*}(x) = 1 - 1/(2M)$, $\theta_m(x) = 1/(2M)$ for all $m \neq m^*$ and one has

$$\theta_{m^*}(x) - \theta_m(x) = 1 - 1/M \geqslant \frac{1}{2} \geqslant \frac{1}{2}\left(\eta_{m^*}(x) - \eta_m(x)\right),$$

where we used $M \geqslant 2$.

Consider the case (ii). In this case, for all $m \neq m^*$, it holds

$$\eta_{m^*}(x) - \eta_m(x) = \theta_{m^*}(x) - \theta_m(x) \leqslant 2\left(\theta_{m^*}(x) - \theta_m(x)\right).$$

Finally, consider the case (iii). Since, $\eta_{m^*}(x) \geqslant \frac{1}{M}$ (otherwise, one gets a contradiction with the fact that $m^* \in \operatorname*{argmax}_{1 \leqslant m \leqslant M} \eta_m(x)$), it holds

$$\frac{\eta_{m^*}(x) + \eta_m(x)}{2} \geqslant \frac{1}{2M},$$

and we have for all $m \neq m^*$

$$\theta_{m^*}(x) - \theta_m(x) = \eta_{m^*}(x) - \frac{1}{2M} \geqslant \frac{1}{2}\big(\eta_m^*(x) - \eta_m(x)\big).$$

□

Lemma B.3. *Fix a point $x \in \mathcal{X}$, an integer $1 \leqslant m \leqslant M$ and a set of weights $\{w_i(X_i, x) : 1 \leqslant i \leqslant n\}$. Denote*

$$\widetilde{\eta}_m(x) = \frac{1}{N(x)} \sum_{i=1}^{n} w_i(X_i, x) \mathbb{1}(Y_i = m),$$
$$\overline{\eta}_m(x) = \frac{1}{N(x)} \sum_{i=1}^{n} w_i(X_i, x) \eta_m(X_i),$$

where

$$N(x) = \sum_{i=1}^{n} w_i(X_i, x).$$

Assume that $0 \leqslant w_i(x) \leqslant 1$ *for* $1 \leqslant i \leqslant n$. *Then, for any* $t > 0$, *it holds*

$$\mathbb{P}_{S_n}\left(|\widetilde{\eta}_m(x) - \overline{\eta}_m(x)| > t \,|\, X_1, \ldots, X_n\right) \leqslant 2e^{-2N(x)t^2}.$$

Proof.

$$\begin{aligned}
&\mathbb{P}_{S_n}\left(|\widetilde{\eta}_m(x) - \overline{\eta}_m(x)| > t \,|\, X_1, \ldots, X_n\right) \\
&= \mathbb{P}_{S_n}\left(\Big|\sum_{i=1}^{n} w_i(X_i, x)\left(\mathbb{1}(Y_i = m) - \eta_m(X_i)\right)\Big| > N(x)t \,\Big|\, X_1, \ldots, X_n\right).
\end{aligned}$$

The Hoeffding inequality yields

$$\begin{aligned}
\mathbb{P}_{S_n}\left(|\widetilde{\eta}_m(x) - \overline{\eta}_m(x)| > t \,|\, X_1, \ldots, X_n\right) &\leqslant 2\exp\left\{-\frac{2N^2(x)t^2}{\sum\limits_{i=1}^{n} w_i^2(X_i, x)}\right\} \\
&\leqslant 2\exp\left\{-\frac{2N^2(x)t^2}{\sum\limits_{i=1}^{n} w_i(X_i, x)}\right\} = 2e^{-2N(x)t^2}.
\end{aligned}$$

□

Appendix C. Additional proofs

C.1 Proof of Theorem 4.2

Proof. Fix any $x \in \mathcal{X}$ and $1 \leqslant m \leqslant M$. Denote

$$\overline{\eta}_m^w(x) = N_w^{-1}(x)\sum_{i=1}^{n} w_i(X_i, x)\eta_m(X_i),$$

where

$$N_w(x) = \sum_{i=1}^{n} w_i(X_i, x).$$

The triangle inequality yields

$$|\eta_m(x) - \widetilde{\eta}_m^w(x)| \leqslant |\eta_m(x) - \overline{\eta}_m^w(x)| + |\widetilde{\eta}_m^w(x) - \overline{\eta}_m^w(x)|.$$

Consider $|\eta_m(x) - \overline{\eta}_m^w(x)|$. Since, according to (A3), $\eta_m(\cdot)$ is (L, α)-Holder continuous, it holds

$$\begin{aligned}
|\eta_m(x) - \overline{\eta}_m^w(x)| &= \left|\eta_m(x) - \frac{1}{N_w(x)}\sum_{i:w_i(X_i,x)>0} w_i(X_i, x)\eta_m(X_i)\right| \\
&\leqslant \frac{1}{N_w(x)}\sum_{i:w_i(X_i,x)>0} w_i(X_i, x)\,|\eta_m(x) - \eta_m(X_i)|
\end{aligned}$$

$$\leqslant L \max_{i:w_i(X_i,x)>0} \|X_i - x\|^\alpha = L\|X_{(\mathsf{k})}(x) - x\|^\alpha,$$

where $X_{(\mathsf{k})}(x)$ is the k-th nearest neighbor of x. The last equality holds, because $w_i(X_i, x) = 0$ if the point X_i is not amongst n_k nearest neighbors of x.

For any $t \in (0, r_0]$, it holds

$$\begin{aligned}\mathbb{P}\left(\|X_{(\mathsf{k})}(x) - x\| > t\right) &= \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}(X_i \in B(x,t)) < \mathsf{k}\right)\\ &\leqslant \mathbb{P}\left(\mathrm{Binom}(n, \varkappa p(x) t^d) < \mathsf{k}\right),\end{aligned}$$

where $\mathrm{Binom}(n, \varkappa p(x)t^d)$ stands for the Binomial random variable with parameters n and $\varkappa p(x)t^d$, and the last inequality follows from (A5). Next, the Bernstein's inequality yields

$$\begin{aligned}&\mathbb{P}\left(\|X_{(\mathsf{k})}(x) - x\| > t\right)\\ &\leqslant \exp\left\{-\frac{(n\varkappa p(x)t^d - \mathsf{k})^2}{2n\varkappa p(x)t^d(1 - \varkappa p(x)t^d) + 2(n\varkappa p(x)t^d - n_k)/3}\right\}\\ &\leqslant \exp\left\{-\frac{(n\varkappa p(x)t^d - \mathsf{k})^2}{2n\varkappa p(x)t^d + 2(n\varkappa p(x)t^d - \mathsf{k})/3}\right\},\end{aligned}$$

provided that $n\varkappa p(x)t^d > \mathsf{k}$. Denote $u = n\varkappa p(x)t^d - \mathsf{k}$. We want to choose $u > 0$ in such a way that

$$\exp\left\{-\frac{u^2}{2(u+\mathsf{k}) + 2u/3}\right\} \leqslant \frac{\delta}{2}.$$

This is equivalent to choose $u > 0$ satisfying

$$u^2 \geqslant \frac{8u}{3}\log\frac{2}{\delta} + 2\mathsf{k}\log\frac{2}{\delta}. \tag{C.1}$$

Take $u = \mathsf{k} + 4\log(2/\delta)$. The chain of inequalities

$$\begin{aligned}\mathsf{k} + 4\log\frac{2}{\delta} &\geqslant \frac{8}{3}\log\frac{2}{\delta} + \mathsf{k} + \frac{1}{2}\log\frac{2}{\delta}\\ &\geqslant \frac{8}{3}\log\frac{2}{\delta} + \sqrt{2\mathsf{k}\log\frac{2}{\delta}} \geqslant \frac{4}{3}\log\frac{2}{\delta} + \sqrt{\left(\frac{4}{3}\log\frac{2}{\delta}\right)^2 + 2\mathsf{k}\log\frac{2}{\delta}}\end{aligned}$$

ensures that such choice of u fulfils (C.1). Thus, the choice

$$t^d = \frac{1}{n\varkappa p(x)}\left(2\mathsf{k} + 4\log(2/\delta)\right)$$

yields

$$\exp\left\{-\frac{(n\varkappa p(x)t^d - \mathsf{k})^2}{2n\varkappa p(x)t^d + 2(n\varkappa p(x)t^d - \mathsf{k})/3}\right\} \leqslant \frac{\delta}{2}.$$

Thus, we proved that, on an event with probability at least $1-\delta/2$, it holds

$$\|X_{(\mathsf{k})}(x) - x\| \leqslant \frac{1}{n\varkappa p(x)}\left(2\mathsf{k} + 4\log(2/\delta)\right).$$

It remains to bound $|\widetilde{\eta}_m^w(x) - \overline{\eta}_m^w(x)|$. Lemma B.3 implies

$$\mathbb{P}^{\otimes n}\left(|\widetilde{\eta}_m^w(x) - \overline{\eta}_m^w(x)| > s \,|\, X_1, \dots, X_n\right) \leqslant 2e^{-2N_k s^2} \leqslant 2e^{-n_k s^2}.$$

Then, taking the expectation with respect to $X_1, \dots, X_n$, one obtains

$$\mathbb{P}^{\otimes n}\left(|\widetilde{\eta}_m^w(x) - \overline{\eta}_m^w(x)| > s\right) \leqslant 2e^{-n_k s^2}.$$

Bringing the two bounds together, one obtains that, with probability at least $1-\delta$ over training samples, it holds

$$\begin{aligned}|\eta_m(x) - \widetilde{\eta}_m^w(x)| &\leqslant L\|X_{(n_k)}(x) - x\|^\alpha + |\widetilde{\eta}_m^w(x) - \overline{\eta}_m^w(x)| \\ &\leqslant \frac{L}{(n\varkappa p(x))^{\alpha/d}}\left(2\mathsf{k} + 4\log(2/\delta)\right)^{\alpha/d} + \sqrt{\frac{\log(4/\delta)}{\mathsf{k}}}.\end{aligned}$$

□

C.2 Proof of Lemma A.2

Proof. Equations (A.1) and (A.2) and 1-Lipschitzness of the funciton $\varphi(\cdot)$ imply

$$M|\overline{\theta}_m^{(k)}(x) - \overline{\theta}_m^{(k-1)}(x)| \leqslant \sqrt{\frac{z_k}{4N_k(x)}}, \quad 1 \leqslant m \leqslant M, 1 \leqslant k \leqslant k^*,$$

where $\overline{\theta}_m^{(k)}(x) = \varphi\big(\overline{\eta}_m^{(k)}(x)\big)$ and $\overline{\theta}_m^{(k-1)}(x) = \varphi\big(\overline{\eta}_m^{(k-1)}(x)\big)$. Next, we need an auxiliary result, which is formulated in Lemma B.1. It claims that

$$\sqrt{\frac{3}{M}}|\widetilde{\theta}_m^{(k)}(x) - \widetilde{\theta}_m^{(k-1)}(x)| \leqslant \mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}(x), \widetilde{\theta}_m^{(k-1)}(x)\right) \leqslant M|\widetilde{\theta}_m^{(k)}(x) - \widetilde{\theta}_m^{(k-1)}(x)|.$$

Then for any fixed $1 \leqslant m \leqslant M$ and $1 \leqslant k \leqslant k^*$ it holds

$$\begin{aligned}&\mathbb{P}_{S_n}\left(N_k(x)\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widetilde{\theta}_m^{(k-1)}(x)\right) > z_k \,\middle|\, X_1, \dots, X_n\right) \\ &= \mathbb{P}_{S_n}\left(\mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}(x), \widetilde{\theta}_m^{(k-1)}(x)\right) > \sqrt{\frac{z_k}{N_k(x)}} \,\middle|\, X_1, \dots, X_n\right) \\ &\leqslant \mathbb{P}_{S_n}\left(M|\widetilde{\theta}_m^{(k)}(x) - \widetilde{\theta}_m^{(k-1)}(x)| > \sqrt{\frac{z_k}{N_k(x)}} \,\middle|\, X_1, \dots, X_n\right) \\ &\leqslant \mathbb{P}_{S_n}\Big(M|\widetilde{\theta}_m^{(k)}(x) - \overline{\theta}_m^{(k)}(x)| + M|\overline{\theta}_m^{(k)}(x) - \overline{\theta}_m^{(k-1)}(x)| \\ &\qquad + M|\widetilde{\theta}_m^{(k-1)}(x) - \overline{\theta}_m^{(k-1)}(x)| > \sqrt{\frac{z_k}{N_k(x)}} \,\middle|\, X_1, \dots, X_n\Big) \\ &\leqslant \mathbb{P}_{S_n}\Big(M|\widetilde{\theta}_m^{(k)}(x) - \overline{\theta}_m^{(k)}(x)| + M|\widetilde{\theta}_m^{(k-1)}(x) - \overline{\theta}_m^{(k-1)}(x)|\end{aligned}$$

$$
\begin{aligned}
&> \sqrt{\frac{z_k}{4N_k(x)}} \,\Big| X_1, \dots, X_n \Big) \\
&\leqslant \mathbb{P}_{S_n}\left(M|\widetilde{\theta}_m^{(k)}(x) - \overline{\theta}_m^{(k)}(x)| > \sqrt{\frac{z_k}{16N_k(x)}} \,\Big| X_1, \dots, X_n \right) \\
&\quad + \mathbb{P}_{S_n}\left(M|\widetilde{\theta}_m^{(k-1)}(x) - \overline{\theta}_m^{(k-1)}(x)| > \sqrt{\frac{z_k}{16N_k(x)}} \,\Big| X_1, \dots, X_n \right) \\
&\leqslant \mathbb{P}_{S_n}\left(M|\widetilde{\eta}_m^{(k)}(x) - \overline{\eta}_m^{(k)}(x)| > \sqrt{\frac{z_k}{16N_k(x)}} \,\Big| X_1, \dots, X_n \right) \\
&\quad + \mathbb{P}_{S_n}\left(M|\widetilde{\eta}_m^{(k-1)}(x) - \overline{\eta}_m^{(k-1)}(x)| > \sqrt{\frac{u_0 z_k}{16N_{k-1}(x)}} \,\Big| X_1, \dots, X_n \right).
\end{aligned}
$$

Lemma B.3 yields

$$
\begin{aligned}
&\mathbb{P}_{S_n}\left(M|\widetilde{\eta}_m^{(k)}(x) - \overline{\eta}_m^{(k)}(x)| > \sqrt{\frac{z_k}{16N_k(x)}} \,\Big| X_1, \dots, X_n \right) \leqslant 2e^{-z_k/(8M^2)}, \\
&\mathbb{P}_{S_n}\left(M|\widetilde{\eta}_m^{(k-1)}(x) - \overline{\eta}_m^{(k-1)}(x)| > \sqrt{\frac{u_0 z_k}{16N_{k-1}(x)}} \,\Big| X_1, \dots, X_n \right) \leqslant 2e^{-u_0 z_k/(8M^2)},
\end{aligned}
$$

and then

$$
\begin{aligned}
&\mathbb{P}_{S_n}\left(N_k(x)\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widetilde{\theta}_m^{(k-1)}(x)\right) > z_k \,\Big| X_1, \dots, X_n \right) \\
&\leqslant 2e^{-z_k/(8M^2)} + 2e^{-u_0 z_k/(8M^2)} \leqslant 4e^{-u_0 z_k/(8M^2)}.
\end{aligned}
$$

The union bound implies that the next inequality holds simultaneously for all $1 \leqslant m \leqslant M$ and $1 \leqslant k \leqslant k^*$:

$$
\mathbb{P}_{S_n}\left(N_k(x)\mathcal{K}\left(\widetilde{\theta}_m^{(k)}(x), \widetilde{\theta}_m^{(k-1)}(x)\right) > z_k \,\Big| X_1, \dots, X_n \right) \leqslant 4KMe^{-u_0 z_k/(8M^2)}.
$$

Now, it is easy to observe that, given $\delta \in (0, 1/3)$, the choice

$$
z_k = \frac{8M^2}{u_0} \log \frac{12KM}{\delta}
$$

ensures that $\widehat{\theta}_m^{(k)}(x) = \widetilde{\theta}_m^{(k)}(x)$ simultaneously for all m, $1 \leqslant m \leqslant M$, and k, $1 \leqslant k \leqslant k^*$, with probability at least $1 - \delta/3$ over training samples.

Next, following the proof of Theorem 5.3 in [5], one can easily obtain that it holds almost surely

$$
\mathcal{K}\left(\widehat{\theta}_m^{(k)}(x), \widehat{\theta}_m^{(k-1)}(x)\right) \leqslant \frac{z_k}{N_k(x)}, \quad \forall\, 1 \leqslant m \leqslant M, \forall\, 1 \leqslant k \leqslant k^*.
$$

This and Lemma B.1 imply

$$
|\widehat{\theta}_m^{(k)}(x) - \widehat{\theta}_m^{(k-1)}(x)| \leqslant \sqrt{\frac{Mz_k}{3N_k(x)}}, \quad \forall\, 1 \leqslant m \leqslant M, \forall\, 1 \leqslant k \leqslant k^*.
$$

Due to (A1) and (A2), $u_0 \leqslant N_{k-1}(x)/N_k(x) \leqslant u$ holds almost surely. Then

$$\begin{aligned}
|\widehat{\theta}_m(x) - \widehat{\theta}_m^{(k)}(x)| &\leqslant \sum_{j=k+1}^{K} |\widehat{\theta}_m^{(j)}(x) - \widehat{\theta}_m^{(j-1)}(x)| \leqslant \sum_{j=k+1}^{K} \sqrt{\frac{Mz_j}{3N_j(x)}} \\
&= \frac{2M^{3/2}\sqrt{2\log(12KM/\delta)}}{\sqrt{3u_0}} \sum_{j=k+1}^{K} \frac{1}{\sqrt{N_j(x)}} \\
&\leqslant \frac{2M^{3/2}\sqrt{2\log(12KM/\delta)}}{\sqrt{3u_0}} \sum_{j=1}^{K-k} \sqrt{\frac{u^j}{N_k(x)}} \\
&\leqslant \frac{2M^{3/2}\sqrt{2\log(12KM/\delta)}}{\sqrt{3u_0}} \sum_{j=1}^{\infty} \sqrt{\frac{u^j}{N_k(x)}} \\
&\leqslant \frac{2M^{3/2}\sqrt{2\log(12KM/\delta)}}{\sqrt{3u_0}} \cdot \frac{\sqrt{u}}{1-\sqrt{u}} \cdot \frac{1}{\sqrt{N_k(x)}} \\
&\leqslant \frac{2M^{3/2}\sqrt{2\log(12KM/\delta)}}{\sqrt{3u_0}} \cdot \frac{\sqrt{u}}{1-\sqrt{u}} \cdot \sqrt{\frac{2}{n_k}}.
\end{aligned}$$

Here we used that, due to (A1), $N_k(x) \leqslant n_k \leqslant 2N_k(x)$ for any x. Thus, with probability at least $1-\delta$ over learning samples, simultaneously for all $1 \leqslant m \leqslant M$ and $1 \leqslant k \leqslant k^*$ it holds

$$|\widehat{\theta}_m(x) - \widetilde{\theta}_m^{(k)}(x)| \leqslant C_1 M^{3/2} \sqrt{\frac{\log(12KM/\delta)}{n_k}}$$

with the constant $C_1 = 4\sqrt{\frac{u}{3u_0}}(1-\sqrt{u})^{-1}$. □

C.3 Proof of Lemma A.3

Proof. Fix some $1 \leqslant m \leqslant M$ and $1 \leqslant k \leqslant K$. Let h_{k-1} and h_k stand for the distance from x to its n_{k-1}-th and n_k-th nearest neighbors respectively. Then

$$\begin{aligned}
|\overline{\eta}_m^{(k)}(x) - \overline{\eta}_m^{(k-1)}(x)| &\leqslant |\overline{\eta}_m^{(k)}(x) - \eta_m(x)| + |\eta_m(x) - \overline{\eta}_m^{(k-1)}(x)| \\
&\leqslant \sum_{i:\|X_i - x\| \leqslant h_k} \left| \frac{w_i^{(k)}(X_i, x)\eta_m(X_i)}{N_k(x)} - \eta_m(x) \right| \\
&+ \sum_{i:\|X_i - x\| \leqslant h_{k-1}} \left| \frac{w_i^{(k-1)}(X_i, x)\eta_m(X_i)}{N_{k-1}(x)} - \eta_m(x) \right| \\
&= \frac{1}{N_k(x)} \sum_{i:\|X_i - x\| \leqslant h_k} \left| w_i^{(k)}(X_i, x)\,(\eta_m(X_i) - \eta_m(x)) \right| \\
&+ \frac{1}{N_{k-1}(x)} \sum_{i:\|X_i - x\| \leqslant h_{k-1}} \left| w_i^{(k-1)}(X_i, x)\,(\eta_m(X_i) - \eta_m(x)) \right| \\
&\leqslant \frac{1}{N_k(x)} \sum_{i:\|X_i - x\| \leqslant h_k} w_i^{(k)}(X_i, x)\,|\eta_m(X_i) - \eta_m(x)|
\end{aligned}$$

$$+\frac{1}{N_{k-1}(x)}\sum_{i:\|X_i-x\|\leqslant h_{k-1}} w_i^{(k-1)}(X_i,x)\left|\eta_m(X_i)-\eta_m(x)\right|.$$

Since $\eta_m(x)$ is (L,α)-Holder function, the last expression does not exceed

$$\begin{aligned}&\frac{1}{N_k(x)}\sum_{i:\|X_i-x\|\leqslant h_k} w_i^{(k)}(X_i,x)L\|X_i-x\|^\alpha\\&+\frac{1}{N_{k-1}(x)}\sum_{i:\|X_i-x\|\leqslant h_{k-1}} w_i^{(k-1)}(X_i,x)L\|X_i-x\|^\alpha\\&\leqslant Lh_k^\alpha+Lh_{k-1}^\alpha\leqslant 2Lh_k^\alpha.\end{aligned}$$

Thus,

$$|\overline{\eta}_m^{(k)}(x)-\overline{\eta}_m^{(k-1)}(x)|\leqslant 2Lh_k^\alpha.$$

For any $t\in(0,r_0]$ it holds

$$\mathbb{P}(h_k>t)=\mathbb{P}\left(\sum_{i=1}^n \mathbb{1}(X_i\in B(x,t))\leqslant n_k\right)\leqslant \mathbb{P}\left(\mathrm{Binom}(n,\varkappa p(x)t^d)\leqslant n_k\right),$$

where $\mathrm{Binom}(n,\varkappa p(x)t^d)$ stands for the Binomial random variable with parameters n and $\varkappa p(x)t^d$, and the last inequality holds since the condition (A5) implies $\mathbb{P}_X(B(x,t))\geqslant \varkappa p(x)t^d$. Next, the Bernstein's inequality yields

$$\begin{aligned}\mathbb{P}\left(h_k>t\right)&\leqslant\exp\left\{-\frac{(n\varkappa p(x)t^d-n_k)^2}{2n\varkappa p(x)t^d(1-\varkappa p(x)t^d)+2(n\varkappa p(x)t^d-n_k)/3}\right\}\\&\leqslant\exp\left\{-\frac{(n\varkappa p(x)t^d-n_k)^2}{2n\varkappa p(x)t^d+2(n\varkappa p(x)t^d-n_k)/3}\right\},\end{aligned}$$

provided that $n\varkappa p(x)t^d>n_k$. Let $u=n\varkappa p(x)t^d-n_k$. We want to choose $u>0$ such that

$$\exp\left\{-\frac{u^2}{2(u+n_k)+2u/3}\right\}\leqslant\frac{\delta}{3KM}.$$

This is equivalent

$$u^2\geqslant\frac{8u}{3}\log\frac{3}{\delta}+2n_k\log\frac{3KM}{\delta}. \tag{C.2}$$

Take $u=n_k+4\log(3KM/\delta)$. Again, as in the proof of Theorem 4.2, the chain of inequalities

$$\begin{aligned}&n_k+4\log\frac{3KM}{\delta}\geqslant\frac{8}{3}\log\frac{3KM}{\delta}+n_k+\frac{1}{2}\log\frac{3KM}{\delta}\\&\geqslant\frac{8}{3}\log\frac{3KM}{\delta}+\sqrt{2n_k\log\frac{3KM}{\delta}}\geqslant\frac{4}{3}\log\frac{3KM}{\delta}+\sqrt{\left(\frac{4}{3}\log\frac{3KM}{\delta}\right)^2+2n_k\log\frac{3KM}{\delta}}\end{aligned}$$

ensures that such u fulfils (C.2). Thus, the choice

$$t^d = \frac{1}{n\varkappa p(x)} \left(2n_k + 4\log(3KM/\delta)\right)$$

yields

$$\exp\left\{-\frac{(n\varkappa p(x)t^d - n_k)^2}{2n\varkappa p(x)t^d + 2(n\varkappa p(x)t^d - n_k)/3}\right\} \leqslant \frac{\delta}{3KM}.$$

Thus, with probability at least $(1 - \delta/(3KM))$ over training samples, one has

$$h_k^d \leqslant \frac{1}{n\varkappa p(x)} \left(2n_k + 4\log(3KM/\delta)\right)$$

and

$$|\overline{\eta}_m^{(k)} - \overline{\eta}_m^{(k-1)}| \leqslant 2L \left(n\varkappa p(x)\right)^{-\alpha/d} \left(2n_k + 4\log(3KM/\delta)\right)^{\alpha/d}.$$

Now, fix any $1 \leqslant k' \leqslant K$. The union bound implies that, with probability at least $1 - \delta/3$, for all k, $1 \leqslant k \leqslant k'$, and all m, $1 \leqslant m \leqslant M$, it holds

$$|\overline{\eta}_m^{(k)}(x) - \overline{\eta}_m^{(k-1)}(x)| \leqslant 2L \left(n\varkappa p(x)\right)^{-\alpha/d} \left(2n_k + 4\log(3KM/\delta)\right)^{\alpha/d}.$$

It remains to find values of n_k when

$$2L \left(n\varkappa p(x)\right)^{-\alpha/d} \left(2n_k + 4\log(3KM/\delta)\right)^{\alpha/d} \leqslant \sqrt{\frac{2\log(12KM/\delta)}{u_0 n_k}}.$$

Let $n_k = c^2 \cdot \log(12KM/\delta)$ and find such values of c that

$$2L \left(n\varkappa p(x)\right)^{-\alpha/d} (2c^2 + 4)^{\alpha/d} \left(\log(12KM/\delta)\right)^{\alpha/d} \leqslant \sqrt{\frac{2}{u_0}} \cdot \frac{1}{c}.$$

It is equivalent to

$$c(2c^2 + 4)^{\alpha/d} \leqslant \frac{\sqrt{2}}{2L\sqrt{u_0}} \left(\frac{n\varkappa p(x)}{\log(12KM/\delta)}\right)^{\alpha/d}.$$

Denote $c_1 = c \vee 1$. One can easily observe that any c_1 fulfilling

$$c_1^{(2\alpha+d)/d} \leqslant \frac{\sqrt{2}}{2L\sqrt{u_0}} \left(\frac{n\varkappa p(x)}{5\log(12KM/\delta)}\right)^{\alpha/d}$$

ensures the previous inequality. Thus, we finally obtain that if

$$n_k \lesssim \left(\left(n\varkappa p(x)\right)^{2\alpha/(2\alpha+d)} \left(\log(12KM/\delta)\right)^{d/(2\alpha+d)}\right) \vee \log(12KM/\delta), \quad 1 \leqslant k \leqslant k',$$

then

$$|\overline{\eta}_m^{(k)}(x) - \overline{\eta}_m^{(k-1)}(x)| \leqslant \sqrt{\frac{8\log(12KM/\delta)}{u_0 n_k}}$$

holds simultaneously for $1 \leqslant m \leqslant M$ and $1 \leqslant k \leqslant k'$ with probability at least $1 - \delta/3$. This yields

$$n_{k^*} \geqslant n_{k'} \asymp \left((n\varkappa p(x))^{2\alpha/(2\alpha+d)} \left(\log(12KM/\delta)\right)^{d/(2\alpha+d)}\right) \vee \log(12KM/\delta)\,.$$

□

C.4 Proof of Lemma A.4

Proof. Define $q = 1 + \frac{1}{r+\beta+2}$. Fix an arbitrary $t > 0$ and denote

$$A_i = \left\{q^{i-1}t < \eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X) \leqslant q^i t\right\}, \quad i \geqslant 1.$$

Then, due to (2.7) and Lemma B.2, we have

$$\begin{aligned}
\mathbb{E}_{S_n}\mathcal{E}(\widehat{f}) &= \mathbb{E}_{S_n} R(\widehat{f}) - R(f^*) \\
&= \mathbb{E}_{S_n}\mathbb{E}_X\left[\eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X)\right] \\
&= \mathbb{E}_{S_n}\mathbb{E}_X\left[\eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X)\right]\mathbb{1}\left(f^*(X) \neq \widehat{f}(X)\right) \\
&= \mathbb{E}_{S_n}\mathbb{E}_X\left[\eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X)\right]\mathbb{1}\left(0 < \eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X) \leqslant t\right) \\
&\quad + \sum_{i=1}^{\infty}\mathbb{E}_{S_n}\mathbb{E}_X\left[\eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X)\right]\mathbb{1}(f^*(X) \neq \widehat{f}(X))\mathbb{1}(A_i) \\
&\leqslant t\mathbb{E}_{S_n}\mathbb{P}_X\left(0 < \eta_{f^*(X)}(X) - \eta_{\widehat{f}(X)}(X) \leqslant t\right) \\
&\quad + 2\sum_{i=1}^{\infty}\mathbb{E}_{S_n}\mathbb{E}_X\left[\theta_{f^*(X)}(X) - \theta_{\widehat{f}(X)}(X)\right]\mathbb{1}(f^*(X) \neq \widehat{f}(X))\mathbb{1}(A_i) \\
&\leqslant t\mathbb{P}_X\left(\eta_{(1)}(X) - \eta_{(2)}(X) \leqslant t\right) \\
&\quad + 2\sum_{i=1}^{\infty} q^i t\mathbb{E}_{S_n}\mathbb{E}_X\left[\mathbb{1}(f^*(X) \neq \widehat{f}(X))\mathbb{1}(A_i)\right] \\
&\leqslant Bt^{1+\beta} + 2\sum_{i=1}^{\infty} q^i t\mathbb{E}_{S_n}\mathbb{E}_X\left[\mathbb{1}(f^*(X) \neq \widehat{f}(X))\mathbb{1}(A_i)\right].
\end{aligned}$$

Note that $\widehat{f}(X) \neq f^*(X)$ if and only if $\widehat{\theta}_{\widehat{f}(X)}(X) \geqslant \widehat{\theta}_{f^*(X)}(X)$. Then

$$\begin{aligned}
\theta_{f^*(X)}(X) &\leqslant \widehat{\theta}_{f^*(X)}(X) + |\widehat{\theta}_{f^*(X)}(X) - \theta_{f^*(X)}(X)| \\
&\leqslant \widehat{\theta}_{\widehat{f}(X)}(X) + |\widehat{\theta}_{f^*(X)}(X) - \theta_{f^*(X)}(X)| \\
&\leqslant \theta_{\widehat{f}(X)}(X) + |\widehat{\theta}_{f^*(X)}(X) - \theta_{f^*(X)}(X)| + |\widehat{\theta}_{\widehat{f}(X)}(X) - \theta_{\widehat{f}(X)}(X)|.
\end{aligned}$$

For each $i \in \mathbb{N}$ we have

$$\begin{aligned}
\mathbb{E}_{S_n}\mathbb{E}_X \mathbb{1}\Big(f^*(X) \neq \widehat{f}(X)\Big)\mathbb{1}(A_i) &\leqslant \mathbb{E}_{S_n}\mathbb{E}_X \mathbb{1}\Big(\theta_{f^*(X)}(X) \leqslant \theta_{\widehat{f}(X)}(X) \\
&\quad + |\widehat{\theta}_{f^*(X)}(X) - \theta_{f^*(X)}(X)| + |\widehat{\theta}_{\widehat{f}(X)}(X) - \theta_{\widehat{f}(X)}(X)|\Big)\mathbb{1}(A_i) \\
&\leqslant \mathbb{E}_{S_n}\mathbb{E}_X \mathbb{1}\Big(|\widehat{\theta}_{f^*(X)}(X) - \theta_{f^*(X)}(X)| + |\widehat{\theta}_{\widehat{f}(X)}(X) - \theta_{\widehat{f}(X)}(X)| \geqslant q^{i-1}t\Big)\mathbb{1}(A_i) \\
&\leqslant \mathbb{E}_{S_n}\mathbb{E}_X \mathbb{1}\Big(|\widehat{\theta}_{f^*(X)}(X) - \theta_{f^*(X)}(X)| \geqslant q^{i-2}t) \\
&\quad + \mathbb{1}(|\widehat{\theta}_{\widehat{f}(X)}(X) - \theta_{\widehat{f}(X)}(X)| \geqslant q^{i-2}t\Big)\mathbb{1}(A_i) \\
&\leqslant 2\mathbb{E}_{S_n}\mathbb{E}_X \mathbb{1}\left(\max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(X) - \theta_m(X)| \geqslant q^{i-2}t\right)\mathbb{1}(A_i) \\
&\leqslant 2\mathbb{E}_{S_n}\mathbb{E}_X \mathbb{1}\left(\max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(X) - \theta_m(X)| \geqslant q^{i-2}t\right)\mathbb{1}\left(\eta_{(1)}(X) - \eta_{(2)}(X) \leqslant q^i t\right) \\
&= 2\mathbb{E}_X \mathbb{P}_{S_n}\left(\max_{1\leqslant m\leqslant M} |\widehat{\theta}_m(X) - \theta_m(X)| \geqslant q^{i-2}t\right)\mathbb{1}\left(\eta_{(1)}(X) - \eta_{(2)}(X) \leqslant q^i t\right) \\
&\leqslant 2\mathbb{E}_X \frac{\chi_r}{q^{r(i-2)}t^r}\mathbb{1}\left(\theta_{(1)}(X) - \theta_{(2)}(X) \leqslant q^i t\right) \\
&\leqslant \frac{2\chi_r}{q^{r(i-2)}t^r} \cdot Bq^{\beta i}t^\beta = 2Bq^{2r} \cdot \frac{\chi_r}{(q^i t)^{r-\beta}}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_{S_n}\mathcal{E}(\widehat{f}) &\leqslant Bt^{1+\beta} + \sum_{i=1}^{\infty} 2Bq^{2r} \cdot \frac{\chi_r}{(q^i t)^{r-\beta-1}} \\
&= Bt^{1+\beta}\left(1 + \sum_{i=1}^{\infty} 2q^{2r} \cdot \frac{\chi_r}{q^{i(r-\beta-1)}t^r}\right) \\
&= Bt^{1+\beta}\left(1 + \sum_{i=1}^{\infty} 2q^{2r} \cdot \frac{\chi_r}{q^{i(r-\beta-1)}t^r}\right) \\
&= Bt^{1+\beta}\left(1 + \frac{2q^{2r}\chi_r}{(q^{r-\beta-1}-1)t^r}\right) \\
&\leqslant Bt^{1+\beta}\left(1 + \frac{2(r+\beta+2)q^{r+\beta+1}\chi_r}{(r-\beta-1)t^r}\right).
\end{aligned}$$

Note that

$$q^{r+\beta+1} = \left(1 + \frac{1}{r+\beta+2}\right)^{r+\beta+1} \leqslant e < 3.$$

Now, the choice $t = \chi_r^{1/r}$ implies the assertion of Lemma A.4. □

References

[1] A. Agarwal, Selective sampling algorithms for cost-sensitive multiclass prediction, in Proceedings of the 30th International Conference on Machine Learning, edited by S. Dasgupta and D. McAllester, Vol. 28 of *Proceedings of Machine Learning Research*. Atlanta, Georgia, USA, (2013) 1220–1228.

[2] H. Ahn and K.-J. Kim, Corporate credit rating using multiclass classification models with order information. *World Acad. Sci. Eng. Technol.* **60** (2011) 95–100.
[3] E.L. Allwein, R.E. Schapire and Y. Singer, Reducing multiclass to binary: a unifying approach for margin classifiers. *J. Mach. Learn. Res.* **1** (2001) 113–141.
[4] J.-Y. Audibert and A.B. Tsybakov, Fast learning rates for plug-in classifiers. *Ann. Stat.* **35** (2007) 608–633.
[5] D. Belomestny and V. Spokoiny, Spatial aggregation of local likelihood estimates with applications to classification. *Ann. Stat.* **35** (2007) 2287–2311.
[6] C. Butucea, J.-F. Delmas, A. Dutfoy and R. Fischer, Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Electron. J. Stat.* **11** (2017) 2258–2294.
[7] T.I. Cannings, T.B. Berrett and R.J. Samworth, Local nearest neighbour classification with applications to semi-supervised learning (2017), arxiv.org/abs/1704.00642
[8] K. Chaudhuri and S. Dasgupta, Rates of convergence for nearest neighbor classification, in Vol. 2 of *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, Cambridge, MA, USA (2014) 3437–3445.
[9] K. Crammer and Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.* **2** (2002) 265–292.
[10] D. Dai, P. Rigollet and T. Zhang, Deviation optimal learning using greedy Q-aggregation. *Ann. Stat.* **40** (2012) 1878–1905.
[11] A. Daniely, S. Sabato and S.S. Shwartz, Multiclass learning approaches: A theoretical comparison with implications, in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Curran Associates, Inc. (2012), 485–493.
[12] D. Dheeru and E. Karra Taniskidou, UCI machine learning repository, 2017.
[13] T.G. Dietterich and G. Bakiri, Solving multiclass learning problems via error-correcting output codes. *J. Artif. Int. Res.* **2** (1995) 263–286.
[14] C.H.Q. Ding and I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17** (2001) 349–358.
[15] V. Dinh, L.S.T. Ho, N.V. Cuong, D. Nguyen and B.T. Nguyen, in Theory and Applications of Models of Computation, Vol. 9076 of *Lecture Notes in Computer Sciences*. Springer, Cham (2015) 375–387.
[16] M. Döring, L. Györfi and H. Walk, Rate of convergence of k-nearest-neighbor classification rule. *J. Mach. Learn. Res.*, **18** (2017) 16.
[17] S. Gadat, T. Klein and C. Marteau, Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Stat.* **44** (2016) 982–1009.
[18] A. Ganapathiraju, J.E. Hamaker and J. Picone, Application of support vector machines to speech recognition. *IEEE Trans. Signal Process.* **52** (2004) 2348–2355.
[19] A. Juditsky, P. Rigollet and A.B. Tsybakov, Learning by mirror averaging. *Ann. Stat.* **36** (2008) 2183–2206.
[20] J. Kittler, R. Ghaderi, T. Windeatt and J. Matas. Face verification via error correcting output codes. *Image Vis. Comput.* **21** (2003) 1163–1169.
[21] G. Lecué, Optimal rates of aggregation in classification under low noise assumption. *Bernoulli* **13** (2007) 1000–1022.
[22] G. Lecué, Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli* **19** (2013) 2153–2166.
[23] G. Lecué and P. Rigollet, Optimal learning with Q-aggregation. *Ann. Stat.* **42** (2014) 211–224.
[24] E. Mammen and A.B. Tsybakov, Smooth discrimination analysis. *Ann. Stat.* **27** (1999) 1808–1829.
[25] V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Ann. Stat.* **41** (2013) 693–721.
[26] R. Rifkin and A. Klautau, In defense of one-vs-all classification. *J. Mach. Learn. Res.* **5** (2003/04) 101–141.
[27] P. Rigollet, Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Stat.* **40** (2012) 639–665.
[28] P. Rigollet and A.B. Tsybakov, Sparse estimation by exponential weighting. *Stat. Sci.* **27** (2012) 558–575.
[29] B.I. Rubinstein, P.L. Bartlett and J.H. Rubinstein, Shifting, one-inclusion mistake bounds and tight multiclass expected risk bounds, in *Advances in Neural Information Processing Systems 19*, edited by B. Schölkopf, J.C. Platt, and T. Hoffman, MIT Press, Cambridge (2007) 1193–1200.
[30] R.J. Samworth, Optimal weighted nearest neighbour classifiers. *Ann. Stat.* **40** (2012) 2733–2763.
[31] V. Spokoiny and C. Vial, Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Stat.* **37** (2009) 2783–2807.
[32] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Stat. Sci.* **18** (2003) 02.
[33] A.B. Tsybakov, Optimal Rates of Aggregation, Springer, Berlin (2003) 303–313.
[34] A.B. Yuditskiĭ, A.V. Nazin, A.B. Tsybakov and N. Vayatis, Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii* **41** (2005) 78–96.