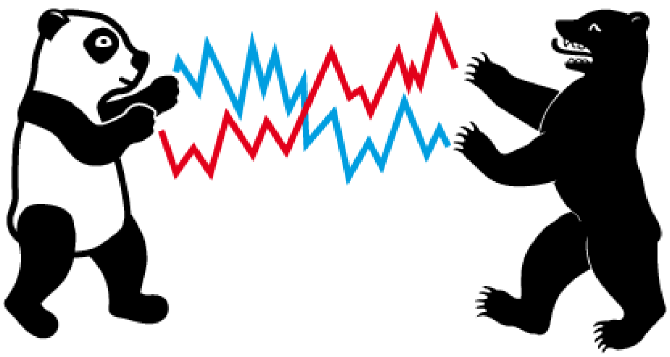
IRTG 1792 Discussion Paper 2018-029

Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification

Nikita Puchkin *
Vladimir Spokoiny *[2]



* National Research University Higher School of Economics, Russia
*[2] WIAS, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification*

Nikita Puchkin [1,3,4] **and Vladimir Spokoiny** [1,2,3,5]

[1] *Institute for Information Transmission Problems RAS, Bolshoy Karetny per. 19, 127051, Moscow, RF.*
e-mail: npuchkin@gmail.com

[2] *Weierstrass Institute and Humboldt University, Mohrenstr. 39, 10117 Berlin, Germany.*
e-mail: spokoiny@wias-berlin.de

[3] *National Research University Higher School of Economics, 20 Myasnitskaya ulitsa, 101000, Moscow, RF.*

[4] *Moscow Institute of Physics and Technology, 141701, Dolgoprudny, RF.*

[5] *Skolkovo Institute of Science and Technology (Skoltech), 143026, Moscow, RF.*

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 $\mathbb{P}(Y = m | X = x) = \theta_m(x)$, $1 \leqslant m \leqslant M$, and $\theta_1(x), \dots, \theta_M(x)$ are unknown Lipschitz functions. Given a test point X, our goal is to estimate $\theta_1(X), \dots, \theta_M(X)$. An approach based on nonparametric smoothing uses a localization technique, i.e. the weight of observation (X_i, Y_i) depends on the distance between X_i and X. However, local estimates strongly depend on localizing scheme. In our solution we fix several schemes $W_1, \dots, W_K$, compute corresponding local estimates $\widetilde{\theta}^{(1)}, \dots, \widetilde{\theta}^{(K)}$ for each of them and apply an aggregation procedure. We propose an algorithm, which constructs a convex combination of the estimates $\widetilde{\theta}^{(1)}, \dots, \widetilde{\theta}^{(K)}$ such that the aggregated estimate behaves approximately as well as the best one from the collection $\widetilde{\theta}^{(1)}, \dots, \widetilde{\theta}^{(K)}$. We also study theoretical properties of the procedure, prove oracle results and establish rates of convergence under mild assumptions.

MSC 2010 subject classifications: Primary 62H30; secondary 62G08, 62G10.

Contents

*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.

1. Introduction

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

$$\mathbb{P}\left(Y \neq f(X)\right)$$

as small as possible. For a given class of admissible functions $\mathcal{F}$, one is often interested in the excess risk

$$\mathcal{E}(f) = \mathbb{P}\left(Y \neq f(X)\right) - \min_{f' \in \mathcal{F}} \mathbb{P}\left(Y \neq f'(X)\right),$$

which shows, how far the classifier f from the best one in the class $\mathcal{F}$. Note that in this setting f may be chosen outside of $\mathcal{F}$.

Concerning the multiclass learning problem, one can distinguish between two main approaches. The first one is by reducing to 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 [2] 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 [8] and multiclass one-inclusion graph strategy [27]. One can refer to [3] for brief overview of multiclass classification methods. Daniely, Sabato and Shalev-Shwartz in [10] compared OvA, OvO, ECOC, tree-based classifiers and multiclass SVM for linear discrimination rules in a finite-dimensional space. From their theoretical study, multiclass SVM outperforms the OvA method. In [8] 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 [10] do not cover this case. Second, in [23] Rifkin and Klautau showed that OvA behaves comparably to multiclass SVM if binary classifier in OvA is strong.

For each class m, we construct binary labels $Y'_{m,i} = \mathbb{1}(Y = m)$ and assume that, given X, a conditional distribution of $(Y'_m | X)$ is $\mathrm{Be}(\theta^*_m(X))$, where

$\theta^*_m(x) \geqslant 0$, $1 \leqslant m \leqslant M$ and $\sum\limits_{m=1}^{M} \theta^*_m(x) \equiv 1$. This model is very general and covers all possible distributions of $(Y|X)$ on M points. We must put some restrictions on the functions $\theta^*_m(x)$, $1 \leqslant m \leqslant M$. We will provide learning guarantees for the class of Lipschitz functions, i. e. we assume, there exists a constant L such that for all $x, x' \in \mathcal{X}$ and for all m from 1 to M it holds

$$|\theta^*_m(x) - \theta^*_m(x')| \leqslant L|x - x'|.$$

For this model, the optimal classifier f^* can be found analytically

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

Unfortunately, true values $\theta^*_1(X), \dots, \theta^*_M(X)$ are unknown, therefore we study a plug-in rule

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

where $\widehat{\theta}_m(X)$ stands for an estimate of $\theta^*_m(X)$, $1 \leqslant m \leqslant M$. This reduces the problem of classification to a regression problem. In [11] it was shown that in general the regression problem is more difficult than classification. Fortunately, for some classes (including the class of Lipschitz functions), classification and regression have similar complexities as it was shown in [30].

For problems of nonparametric regression, different localization techniques are often used. Namely, one considers an estimate $\widetilde{\theta}_m(X)$ defined by maximization of localized log-likelihood

$$\widetilde{\theta}_m(X) = \operatorname*{argmax}_{\theta_m} \mathcal{L}_m(W, \theta_m) = \operatorname*{argmax}_{\theta_m} \sum_{i=1}^{n} w_i \ell_i(Y'_{m,i}, \theta),$$

where $\ell_i(Y'_{m,i}, \theta)$ is a log-likelihood of the i-th observation, $W = \{w_i\}_{1\leqslant i\leqslant n}$ is called a localizing scheme and localizing weights w_i depend on X_i and X. Particular examples of such technique are Nadaraya-Watson estimator, local polynomial estimators and nearest-neighbor-based estimators.

Note that the estimate $\widetilde{\theta}_m(X)$ strongly depends on the localizing scheme W and its choice determines the performance of the classifier $\widehat{f}$. Moreover, in multiclass learning there is a common problem of class imbalance, i. e. some classes may be not presented in a small vicinity of a distinct point. Obviously, one localizing scheme is not enough for such situation. To solve this problem, we consider several localizing schemes $W_1, \dots, W_K$, compute local likelihood estimates (they are also called weak) $\widetilde{\theta}^{(1)}_m, \dots, \widetilde{\theta}^{(K)}_m$, $1 \leqslant m \leqslant M$, for each of them and use a plug-in classifier based on a convex combination of these estimates. An aggregation of weak estimates is a key feature of our procedure.

The aggregation of estimators takes its origins in model selection and it was generalized to convex and linear aggregation in [15]. In [28] and [31] optimal rates of aggregation were derived. Aggregation procedures have a wide range of applications and can be used in regression problems ([28], [31]), density estimation ([21], [25], [17]) and classification problems ([29], [32], [21]). They often solve an optimization problem in order to find aggregating coefficients ([16], [9], [19], [20]). In some cases such as exponential weighting ([18], [26]), a solution of the optimization problem can be written explicitly. An aggregation under the KL-loss was also studied in [24] and [7], where optimal rates of aggregation and exponential bounds were obtained. However, most of the existing aggregation procedures and results concern with a *global* aggregation. This means that the aggregating coefficients are universal and do not depend on the point X where the classification rule is applied.

Our approach is based on *local* aggregation yielding a point dependent aggregation scheme. However, the proposed procedure does not require to solve an optimization problem. Instead, our procedure and sequentially finds a convex combination of weak estimates, which mimics the best possible choice of a model under the Kullback-Leibler loss for a given test point X. The idea of the approach originates from [6], where an aggregation of binary classifiers was studied.

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 [5] and then in [14] 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 following:

- we propose an algorithm of multiclass classification, based on aggregation of local likelihood estimates, which works for a broad class of admissible functions;
- the procedure is robust against class imbalance and outliers;
- computational time of the procedure is $\widetilde{O}(n_{test} \cdot n_{train})$, where n_{train} and n_{test} stand for the size of train and test datasets respectively, which makes it scalable for large problems;
- theoretical guarantees claim optimal accuracy of classification with only a logarithmic payment for the number of classes and aggregated estimates.

The paper is organized as follows. In Section 2 we introduce definitions and notations. In Section 3 we formulate the multiclass classification procedure and demonstrate its performance on both artificial and real-world datasets. Finally, in Section 4 we study theoretical properties of the procedure. In particular, we derive oracle results for model selection, establish rates of convergence for the problem of nonparametric estimation and provide bounds for the excess risk $\mathcal{E}(\widehat{f})$.

2. Setup and notations

Given a training sample $S_n = \{(X_i, Y_i)\}_{i=1}^n$, we apply the following probabilistic model. Suppose that, given X, labels have a conditional distribution

$$\mathbb{P}\left(Y = m | X\right) = \theta_m^*(X), \quad 1 \leqslant m \leqslant M, \tag{1}$$

$\theta_m^*(x) \geqslant 0$, $1 \leqslant m \leqslant M$ and $\sum\limits_{m=1}^{M} \theta_m^*(x) \equiv 1$. The optimal classifier is the Bayes rule defined by

$$f^*(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \theta_m^*(X) \tag{2}$$

Unfortunately, true values $\theta_1^*(X), \dots, \theta_M^*(X)$ are unknown, therefore we fix a test point $X \in \mathcal{X}$ and consider a plug-in classifier

$$\widehat{f}(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \widehat{\theta}_m(X), \tag{3}$$

where $\widehat{\theta}_m(X)$ stands for an estimate of $\theta_m^*(X)$, $1 \leqslant m \leqslant M$.

Now, the problem is how to estimate $\theta_m^*(X)$, $1 \leqslant m \leqslant M$. Fix some m and transform labels to binary:

$$Y'_{m,i} = \mathbb{1}\left(Y_i = m\right) \tag{4}$$

It is clear that

$$\left(Y'_{m,i} | X_i\right) \sim \mathrm{Be}(\theta_m^*(X_i)), \tag{5}$$

where $\theta_m^* : \mathcal{X} \to \Theta \subseteq \mathbb{R}$. This approach is nothing but the One-vs-All procedure for multiclass classification.

For fixed m and $X \in \mathbb{R}^d$, denote

$$\theta_{m,i}^* = \theta_m^*(X_i), \quad 1 \leqslant i \leqslant n,$$
$$\widehat{\theta}_{m,i} = \widehat{\theta}_m(X_i), \quad 1 \leqslant i \leqslant n,$$

and

$$\theta_m^* = \theta_m^*(X),$$
$$\widehat{\theta}_m = \widehat{\theta}_m(X).$$

One of the ways to estimate θ_m^* is to consider a localized log-likelihood

$$\mathcal{L}_m(\theta_m) = \sum_{i=1}^{n} w_i \ell(Y'_{m,i}, \theta_m), \tag{6}$$

where $\ell(Y'_m, \theta_m) = Y'_m \log \theta_m + (1 - Y'_m)\log(1 - \theta_m)$ is a log-likelihood of one observation and $\{w_i\}_{i=1}^n$ are some non-negative localizing weights. The local maximum likelihood estimate can be found explicitly.

Proposition 1. *For the log-likelihood function of the form (6) the estimate* $\widetilde{\theta}_m = \operatorname*{argmax}_{\theta_m} \mathcal{L}_m(W, \theta_m)$ *is given by the formula*

$$\widetilde{\theta}_m = \frac{S_m}{N}, \tag{7}$$

where $S_m = \sum\limits_{i=1}^n w_i Y'_{m,i}$, $N = \sum\limits_{i=1}^n w_i$. *Moreover, for any* θ_m *it holds*

$$\mathcal{L}_m(W, \widetilde{\theta}_m) - \mathcal{L}_m(W, \theta_m) = N\mathcal{K}\left(\widetilde{\theta}_m, \theta_m\right),$$

where $\mathcal{K}\left(\widetilde{\theta}_m, \theta_m\right) = \widetilde{\theta}_m \log \frac{\widetilde{\theta}_m}{\theta_m} + (1 - \widetilde{\theta}_m) \log \frac{1-\widetilde{\theta}_m}{1-\theta_m}$.

Proof of the proposition is straightforward and requires to compute the derivative $(\partial \mathcal{L}_m(W, \theta_m)/\partial \theta_m)$ and put it to zero.

We proceed with two examples of such weights.

Example 2.1: k nearest neighbors

For k-NN estimates we have

$$w_i = \mathbb{1}(X_i \in \mathscr{D}_k(X)),$$

where $\mathscr{D}_k(X)$ is a set of k nearest to X points over $\{X_i\}_{i=1}^n$. Then

$$\widetilde{\theta}_m = \frac{1}{k} \sum_{i : X_i \in \mathscr{D}_k(X)} Y'_{m,i}$$

Example 2.2: bandwidth-based kernel estimates

For bandwidth-based kernel estimates localizing weights are defined by the formula

$$w_i = K\left(\frac{\|X_i - X\|}{h}\right), \quad 1 \leqslant i \leqslant n, \tag{8}$$

where $\|\cdot\|$ stands for some norm, h is called bandwidth and $K(\cdot)$ is a localizing kernel. Standard examples of such kernels are following:

- rectangular kernel: $K(t) = \mathbb{1}\left(|t| \leqslant 1\right)$
- triangular kernel: $K(t) = (1 - |t|)_+$
- Epanechnikov kernel: $K(t) = (1 - t^2)_+$
- Gaussian kernel: $K(t) = e^{-\frac{t^2}{2}}$

We will use a Euclidean norm in examples in Section 3.

Both k-NN and bandwidth-based localizing schemes require a proper choice of smoothing parameters (k and h respectively). Fortunately, Bernoulli distribution belongs to an exponential family. Such distributions are quite well studied.

In particular, in [6] an adaptive procedure of choosing the smoothing parameter was proposed. We will refer to that procedure as SSA (Spatial Stagewise Aggregation) as it called in the paper [6].

Let $\left\{W^{(k)}\right\}_{k=1,\dots,K}$ be a set of localizing schemes, i. e. for each $W^{(k)} = \left\{w_i^{(k)}, i = 1, \dots, n\right\}$. Each localizing scheme $W^{(k)}$ induces a set of estimates $\widetilde{\theta}_1^{(k)}, \dots, \widetilde{\theta}_M^{(k)}$, defined by (7). Using the SSA procedure, we can get aggregated estimates $\widehat{\theta}_1, \dots, \widehat{\theta}_M$. It was shown in [6] that for each m $\widehat{\theta}_m$ behaves like almost the best estimate from $\widetilde{\theta}_m^{(1)}, \dots, \widetilde{\theta}_m^{(K)}$. Next, we can use the plug-in rule (3). We will only require that for each k from 2 to K, it holds

$$0 \leqslant w_i^{(k-1)} \leqslant w_i^{(k)} \leqslant 1. \tag{A1}$$

For two examples considered above, this condition means that candidates for the best model should be ordered by the number of nearest neighbors or by the bandwidth. The detailed description of the procedure for multiclass classification is given in Section 3. We will refer to it as MSSA (Multiclass Spatial Stagewise Aggregation).

To show consistency of the MSSA procedure we will derive convergence rates for $\max\limits_{1 \leqslant m \leqslant M} |\widehat{\theta}_m - \theta_m^*|^2$ and $\max\limits_{1 \leqslant m \leqslant M} \mathcal{K}\left(\widehat{\theta}_m, \theta_m^*\right)$, where $\mathcal{K}(\cdot, \cdot)$ stands for the KL-divergence between two distributions, under certain assumptions. For two Bernoulli distributions with parameters θ and θ' KL-divergence is defined by

$$\mathcal{K}(\theta, \theta') = \theta \log \frac{\theta}{\theta'} + (1 - \theta) \log \frac{1 - \theta}{1 - \theta'} \tag{9}$$

and it is more informative, than the squared error. In the theoretical study, we will require a regularity of the KL-divergence. Namely, we assume that there exist constants $\kappa_1, \kappa_2 > 0$ such that

$$\kappa_1 \leqslant \theta(1 - \theta) \leqslant \kappa_2, \quad \forall \theta \in \Theta \tag{A2}$$

or equally

$$\sup_{\theta_1, \theta_2 \in \Theta} \frac{I(\theta_1)}{I(\theta_2)} \leqslant \varkappa^2 \tag{A2$'$}$$

where $\varkappa^2 = \frac{\kappa_2}{\kappa_1}$ and

$$I(\theta) = \mathbb{E}_\theta \left(\frac{\partial \log \ell(Y', \theta)}{\partial \theta}\right)^2 = \frac{1}{\theta(1 - \theta)}$$

is a Fischer information. These regularity conditions are usually used (cf. [6], [24]). One may easily notice that κ_2 is bounded above by $\frac{1}{4}$.

Assumptions (A2), (A2$'$) define metric-like properties of the KL-divergence, under these assumptions

$$\mathcal{K}^{1/2}\left(\widetilde{\theta}, \widehat{\theta}\right) \asymp |\widetilde{\theta} - \widehat{\theta}|$$

More precise statements will be discussed in Section 4.

Besides bounds on the estimation error, we also derive bounds on the classification error. Assume that pairs (X, Y) are such that X is drawn according to some distribution $\mathbb{P}_X$ and conditional distribution of Y is given by (1). For every rule $f(X)$ define the risk

$$R(f) = \mathbb{P}(Y \neq f(X))$$

and the excess risk

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

where f^* stands for the Bayes classifier.

Since we deal with the problem of nonparametric estimation, even the optimal estimator can show poor performance in the case of 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

$$R(f) = 1 - \mathbb{E}\mathbb{1}(Y = f(X)) = 1 - \mathbb{E}_X \mathbb{P}(Y = f(X)|X) = 1 - \mathbb{E}_X \theta^*_{f(X)}$$

In case of binary classification a misclassification often occurs, when $\theta^*(X)$ is close to $1/2$ with high probability. The well-known Mammen-Tsybakov noise condition ensures that such situation appears with low probability. Namely, it assumes that there exist universal non-negative constants B and α such that for all $t > 0$ it holds

$$\mathbb{P}_X \left(|2\theta^*(X) - 1| < t\right) \leqslant B t^\alpha \tag{10}$$

This assumption can be generalized to the multiclass case. Suppose, at the point X we have

$$\mathbb{P}(Y = m|X) = \theta^*_m = \theta^*_m(X), \quad 1 \leqslant m \leqslant M$$

Let $\theta^*_{(1)} \geqslant \theta^*_{(2)} \geqslant \ldots \geqslant \theta^*_{(M)}$ be ordered values of $\theta^*_1, \ldots, \theta^*_M$. Then the condition (10) for multiclass can be formulated as follows (cf. [1], [14])

$$\mathbb{P}_X \left(|\theta^*_{(1)}(X) - \theta^*_{(2)}(X)| < t \right) \leqslant B t^\alpha \tag{11}$$

for some non-negative constants B and α. We will use this assumption to establish fast rates for the built plug-in classifier $\hat{f}(X)$ in Section 4.

3. Algorithm and numerical experiments

3.1. Algorithm

Suppose, one has candidates $\left\{w_i^{(k)}\right\}_{i=1}^n$, $1 \leqslant k \leqslant K$, for an optimal localizing scheme. The set of weights $\left\{w_i^{(k)}\right\}_{i=1}^n$ induces weak estimates $\widetilde{\theta}_m^{(k)} = \widetilde{\theta}_m^{(k)}(X)$ for each class m from 1 to M. Denote

$$N_k = \sum_{i=1}^n w_i^{(k)} \tag{12}$$

We assume that any two collections of weights differ at least in one element. The idea of the procedure is simple. For each class m on the first step we choose the estimate $\widetilde{\theta}_m^{(1)}$. This estimate is very local and therefore has the smallest bias and the largest variance of order $1/N_1$. Next, we try to enlarge the vicinity of averaging under a condition that the bias does not change dramatically. For this purpose we run a likelihood-ratio test on homogeneity: if the hypothesis $\mathbb{E}\widetilde{\theta}_m^{(1)} = \mathbb{E}\widetilde{\theta}_m^{(2)}$ is correct, then the difference $\mathcal{L}_m(W_2, \widetilde{\theta}_m^{(2)}) - \mathcal{L}_m(W_2, \widetilde{\theta}_m^{(1)})$ won't be significant. Otherwise, the function $\theta^*(x)$ changes quickly in the vicinity of the test point and it is better to utilize $\widetilde{\theta}_m^{(1)}$. Fix a critical value z_2 and construct the estimate $\widehat{\theta}_m^{(2)} = \gamma_2 \widetilde{\theta}_m^{(2)} + (1-\gamma_2)\widetilde{\theta}_m^{(1)}$, where the coefficient γ_2 is close to 1 if $\mathcal{L}_m(W_2, \widetilde{\theta}_m^{(2)}) - \mathcal{L}_m(W_2, \widetilde{\theta}_m^{(1)}) = N_2 \mathcal{K}\left(\widetilde{\theta}_m^{(2)}, \widetilde{\theta}_m^{(1)}\right)$ is much less than z_2, and close to 0 if $\mathcal{L}_m(W_2, \widetilde{\theta}_m^{(2)}) - \mathcal{L}_m(W_2, \widetilde{\theta}_m^{(1)}) = N_2 \mathcal{K}\left(\widetilde{\theta}_m^{(2)}, \widetilde{\theta}_m^{(1)}\right)$ exceeds the value z_2. On the step k, $3 \leqslant k \leqslant K$, we repeat this test for a new estimate $\widetilde{\theta}_m^{(k)}$ and the estimate $\widehat{\theta}_m^{(k)}$ constructed on the previous step.

Algorithm 1 Multiclass Spatial Stagewise Aggregation (MSSA)

1: **procedure** MSSA($\{w_i^{(k)}\}_{i=\overline{1,n}}^{k=\overline{1,K}}$ – given collection of sets of weights, $\{z_k\}_{k=1}^K$ – given set of critical values)
2: **for** $m \leftarrow 1$ **to** M **do**
3: For each $k = \overline{1,K}$ calculate $S_m^{(k)} = \sum\limits_{i=1}^n Y'_{m,i} w_i^{(k)}$, $N_k = \sum\limits_{i=1}^n w_i^{(k)}$ and $\widetilde{\theta}_m^{(k)} = \frac{S_m^{(k)}}{N_k}$
4: $\widehat{\theta}_m^{(1)} \leftarrow \widetilde{\theta}_m^{(1)}$
5: **for** $k \leftarrow 2$ **to** K **do**
6: $T_m^{(k)} \leftarrow N_k \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m^{(k-1)}\right)$
7: $\gamma_k \leftarrow K_{ag}\left(\frac{T_m^{(k)}}{z_k}\right)$, $K_{ag}(t) = (1-(t-1/6)_+)_+$
8: $\widehat{\theta}_m^{(k)} \leftarrow \gamma_k \widetilde{\theta}_m^{(k)} + (1-\gamma_k)\widehat{\theta}_m^{(k)}$
$\widehat{\theta}_m = \widehat{\theta}_m^{(K)}$
return $\widehat{\theta}_1, \dots, \widehat{\theta}_M$

The choice of $K_{ag}(t) = (1-(t-b))_+$ is the same as it was in [6]. The MSSA procedure returns aggregated estimates $\widehat{\theta}_1 = \widehat{\theta}_1(X), \dots, \widehat{\theta}_M = \widehat{\theta}_M(X)$ for each class. The classification rule $\widehat{f}(X)$ is defined as

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

The choice of parameters z_k is crucial for performance of the procedure. We tune values z_k according to the propagation condition. The propagation condition means that in the homogeneous case $\theta^*(\cdot) \equiv \theta^*$ the procedure must return the estimate $\widetilde{\theta}^{(K)}$ corresponding to the most broadest localizing scheme W_K. If it does not happen such a situation is called an early stopping. The chosen values z_k must ensure that the early stopping occurs with a small probability (e.g. 0.05). In our experiments described further values z_k were tuned only at one point and then used for all test points.

In all numerical experiments, we choose localizing weights according to nearest-neighbor-based schemes. Namely, for a number of neighbors n_k we set h_k equal to the distance to the n_k-th nearest neighbor of the test point X. The weight $w_i^{(k)}$ is then defined by the formula

$$w_i^{(k)} = K\left(\frac{\|X_i - X\|}{h_k}\right),$$

where $K(\cdot)$ is either Gaussian or Epanechnikov kernel. Typically in our experiments, $K = O(\log n)$ and then it requires $\widetilde{O}(n)$ time to compute local estimates for one class and $O(\log n)$ time to aggregate them. As result, it takes $\widetilde{O}(Mn)$ time to compute estimates at one test point.

3.2. Experiments on artificial datasets

We start with presenting the performance of MSSA on artificial datasets. We generate points from a mixture model:

$$p(x|Y = m) = p_m(x)$$

and

$$\mathbb{P}(Y = m) = \pi_m$$

Then the density of X is given by the formula

$$p(x) = \sum_{m=1}^{M} \pi_m p_m(x) \tag{13}$$

The Bayes rule for this case is given by the formula

$$f^*(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \pi_m p_m(x)$$

Below we provide results for three different experiments.

Typical sample realizations in all three experiments are shown on Figure 1. In each experiment, we took a sequence of integers $n_k = \lfloor 3 \cdot 1.25^{i-1} \rfloor$, $1 \leqslant i \leqslant 23$ and considered n_k-nearest-neighbor-based localizing schemes with Epanechnikov and Gaussian kernels. We computed average leave-one-out cross-validation errors over sample realizations.

In the first experiment, we took $M = 3$ classes, $n = 500$ points, equal prior class probabilities $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$ and considered a mixture of the form (13) with

$$p_1(x) = \phi\left(x, [0, -1], 0.5 I_2\right)$$
$$p_2(x) = \phi\left(x, [\sqrt{3}/2, 0], 0.5 I_2\right)$$
$$p_3(x) = \phi\left(x, [-\sqrt{3}/2, 0], 0.5 I_2\right),$$

where $\phi(x, \mu, \Sigma)$ stands for the density of Gaussian random vector with mean μ and variance Σ.

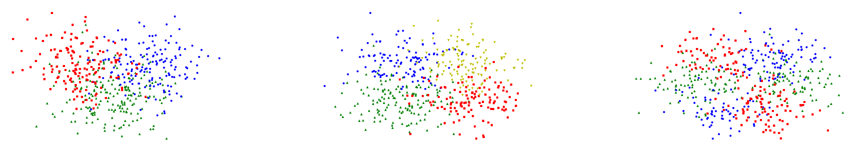


Fig 1. *Sample realizations in the first (left, $M = 3$ classes, $n = 500$ points), second (center, $M = 4$ classes, $n = 500$ points and third (right, $M = 3$ classes, $n = 500$ points) experiments with artificial datasets.*

Misclassification errors for each weak estimate and SSA estimate for both k-NN and bandwidth-based localizing schemes are shown on Figure 2.

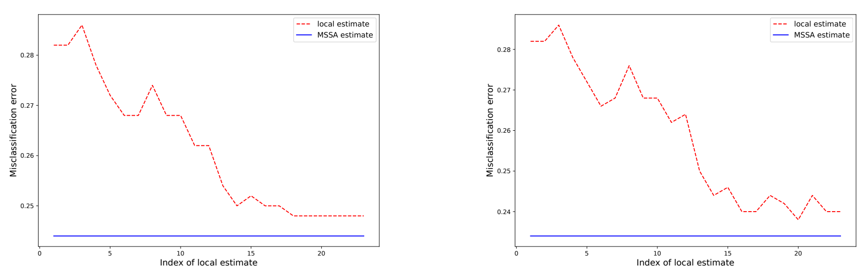


Fig 2. *Misclassification errors for nearest-neighbor-based classifiers with Epanechnikov (left) and Gaussian(right) kernels. The solid line corresponds to the MSSA misclassification error.*

In the second experiment, we took $M = 4$ classes, $n = 500$ points, equal prior class probabilities $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$ and considered a mixture (13) with

$$p_1(x) = \phi\left(x, [-1, -1], 0.7 I_2\right)$$
$$p_2(x) = \phi\left(x, [1, -1], 0.7 I_2\right)$$
$$p_3(x) = \phi\left(x, [-1, 1], 0.7 I_2\right)$$
$$p_4(x) = \phi\left(x, [1, 1], 0.7 I_2\right),$$

where $\phi(x, \mu, \Sigma)$ stands for the density of Gaussian random vector with mean μ and variance Σ. Misclassification errors for each weak estimate and SSA estimate for both k-NN and bandwidth-based localizing schemes are shown on Figure 3.

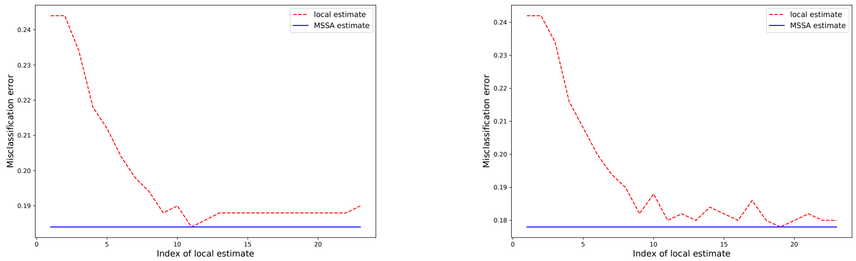


Fig 3. *Misclassification errors for nearest-neighbor-based classifiers with Epanechnikov (left) and Gaussian(right) kernels. The solid line corresponds to the MSSA misclassification error.*

Finally, in the third experiment, we took $M = 3$ classes, $n = 500$ points, equal prior class probabilities $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$ and considered a mixture (13) with

$$\begin{aligned} p_1(x) &= 0.5\phi\left(x, [-1,0], 0.5I_2\right) + 0.5\phi\left(x, [1,0], 0.5I_2\right) \\ p_2(x) &= 0.5\phi\left(x, [0.5, \sqrt{3}/2], 0.5I_2\right) + 0.5\phi\left(x, [-0.5, -\sqrt{3}/2], 0.5I_2\right) \\ p_3(x) &= 0.5\phi\left(x, [-0.5, \sqrt{3}/2], 0.5I_2\right) + 0.5\phi\left(x, [0.5, -\sqrt{3}/2], 0.5I_2\right), \end{aligned}$$

where $\phi(x, \mu, \Sigma)$ stands for the density of Gaussian random vector with mean μ and variance Σ. Misclassification errors for each weak estimate and SSA estimate for both k-NN and bandwidth-based localizing schemes are shown on Figure 4.

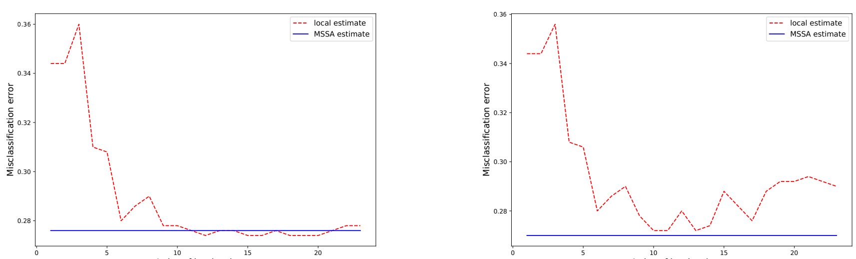


Fig 4. *Misclassification errors for nearest-neighbor-based classifiers with Epanechnikov (left) and Gaussian(right) kernels. The solid line corresponds to the MSSA misclassification error.*

3.3. Experiments on the real world 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 3.3.

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

TABLE 1
Information about datasets from the UCI repository [12]

We compare the performance of our algorithm with boosting of k-NN classifiers considered in [4] and SVM [22]. 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 3.3, best ones are boldfaced.

Dataset	EK MSSA	GK MSSA	Boost-NN, [4]	SVM, [22] (table 2)
Ecoli	12.8 ± 1.8	$\mathbf{12.5 \pm 1.8}$	–	13.0 ± 5.3
Iris	$\mathbf{0.0}$	$\mathbf{0.0}$	–	2.7 ± 2.8
Glass	27.5 ± 3.1	26.6 ± 3.0	$\mathbf{24.4 \pm 1.7}$	32.3 ± 6.6
Pendigits	2.6 ± 0.3	2.5 ± 0.3	3.9 ± 0.6	$\mathbf{0.5 \pm 0.1}$
Satimage	$\mathbf{9.6 \pm 0.7}$	$\mathbf{9.6 \pm 0.7}$	$\mathbf{9.6 \pm 0.3}$	11.0 ± 0.7
Seeds	5.7 ± 1.6	5.7 ± 1.6	–	$\mathbf{4.8 \pm 2.4}$
Wine	2.2 ± 1.1	2.2 ± 1.1	–	$\mathbf{1.7 \pm 1.5}$
Yeast	40.5 ± 1.3	$\mathbf{40.4 \pm 1.3}$	41.7 ± 0.6	40.8 ± 2.3

TABLE 2
Misclassification errors (in %) with standard deviations for datasets from the UCI repository. Best results are boldfaced.

From Table 3.3, one can observe that localizing schemes with the Gaussian kernel behave slightly better than with the Epanechnikov kernel and MSSA with both kernels is comparable with SVM.

4. Theoretical properties

4.1. Main results

Before we formulate main theoretical properties of the procedure, we introduce an additional assumption. Namely, we assume that there exist constants u_0 and u such that

$$0 < u_0 \leqslant \frac{N_k}{N_{k+1}} \leqslant u < 1, \quad 1 \leqslant k \leqslant K-1 \tag{A3}$$

The choice of models, which fulfill the assumption (A3), is up to statistician. Note that this assumption is quite reasonable in sense that if we assume that

N_K is of order $O(n)$ and N_1 is of order $O(1)$ then, K is of order $O(\log n)$, and thus, the number of models we aggregate is not huge.

Main theoretical properties of the MSSA procedure can be formulated in the following theorems. First two results concern accuracy of estimation.

Theorem 1. *Let assumptions* (A1) – (A3) *be fulfilled. Fix arbitrary* $\delta \in (0,1)$ *and define a constant* $\alpha \in (0,1)$ *from the condition*

$$N_k \mathcal{K}\left(\overline{\theta}_m^{(k)}, \overline{\theta}_m^{(k-1)}\right) \leqslant \left(\frac{1-\alpha}{\alpha}\right)^2 \cdot \frac{4\varkappa^4}{u_0} \cdot \log \frac{4KM}{\delta}, \quad m = \overline{1,M},\, k = \overline{2,K} \tag{14}$$

Set

$$z_k = \frac{4\varkappa^4}{\alpha^2 u_0 b} \log \frac{4KM}{\delta} \tag{15}$$

Then with probability at least $1-\delta$ *simultaneously for all* m *from* 1 *to* M *it holds*

$$\mathcal{K}^{1/2}\left(\theta_m^*, \widehat{\theta}_m\right) \leqslant \min_{1 \leqslant k \leqslant K} \left\{ \mathcal{K}^{1/2}\left(\theta_m^*, \widetilde{\theta}_m^{(k)}\right) + \frac{C_1}{\sqrt{N_k}} \log^{\frac{1}{2}} \frac{4KM}{\delta} \right\} \tag{16}$$

and for all $\varepsilon > 0$

$$\mathcal{K}\left(\theta_m^*, \widehat{\theta}_m\right) \leqslant \min_{1 \leqslant k \leqslant K} \left\{ (1+\varepsilon) \mathcal{K}\left(\theta_m^*, \widetilde{\theta}_m^{(k)}\right) + \frac{C_2(\varepsilon)}{N_k} \log \frac{4KM}{\delta} \right\} \tag{17}$$

$$(\theta_m^* - \widehat{\theta}_m)^2 \leqslant \min_{1 \leqslant k \leqslant K} \left\{ (1+\varepsilon)(\theta_m^* - \widetilde{\theta}_m^{(k)})^2 + \frac{C_3(\varepsilon)}{N_k} \log \frac{4KM}{\delta} \right\} \tag{18}$$

where C_1 *is a universal constant and* C_4*,* C_5 *depend only on* $\frac{1}{\varepsilon}$ *(polynomially).*

The result of Theorem 1 improves results in [6]. However, note that this theorem does not imply similar results in expectation, since the choice of parameters depends on the predetermined confidence set level δ. Note that the logarithmic term of the number of models K in (17) is usual for problems of model selection and cannot be improved.

The next result establishes rates of convergence for the procedure.

Theorem 2. *Let conditions of Theorem 1 be fulfilled and set* z_k *as in* (15)*. Let*

$$|\theta_m^*(x+t) - \theta^*(x)| \leqslant Lh, \quad \forall\, |t| \leqslant h,\, 1 \leqslant m \leqslant M$$

Select

$$h = C_4 \left(\frac{\log \frac{4KM}{\delta}}{L^2 n} \right)^{\frac{1}{d+2}} \tag{19}$$

Suppose that for each m *from* 1 *to* M *there exists a localizing scheme* $W_{k^*(m)}$*,* $1 \leqslant k \leqslant K$ *such that*

$$w_i^{(k^*(m))} = 0, \quad \forall\, i : |X_i - x| > h$$

and

$$\alpha_1 n h^d \leqslant N_{k^*(m)} \leqslant \alpha_2 n h^d \tag{20}$$

for some positive constants $0 < \alpha_1 \leqslant \alpha_2$. Then with probability at least $1 - 2\delta$ it holds

$$\max_{1 \leqslant m \leqslant M} \mathcal{K}\left(\theta^*_m, \widehat{\theta}_m\right) \leqslant C_5 \left(\frac{L^d \log \frac{4KM}{\delta}}{n}\right)^{\frac{2}{d+2}}$$

and

$$\max_{1 \leqslant m \leqslant M} (\theta^*_m - \widehat{\theta}_m)^2 \leqslant C_6 \left(\frac{L^d \log \frac{4KM}{\delta}}{n}\right)^{\frac{2}{d+2}}$$

for some absolute constants $C_5, C_6 > 0$.

The rate $O\left(n^{-2/(d+2)}\right)$ is optimal for estimation of Lipschitz functions under regularity of the design. The MSSA procedure provides the optimal rate up to a logarithmic factor, which can be considered as a payment for adaptation.

Note that the condition (A2′) implies that the KL-divergence is bounded by $\frac{\varkappa^2}{2\kappa_1}$. It allows obtaining bounds in expectation for the r-th moment of the KL-loss. Indeed, fix arbitrary $r > 0$ and choose $\delta \sim \left(\frac{K}{n}\right)^{2r/(d+2)} \sim \left(\frac{\log n}{n}\right)^{2r/(d+2)}$. Using the result of Theorem 2 we immediately obtain

$$\mathbb{E}\mathcal{K}^r\left(\theta^*_m, \widehat{\theta}_m\right) \leqslant \frac{\varkappa^{2r}}{2^r \kappa_1^r}\delta + C_5^r \left(\frac{L^d \log \frac{8KM}{\delta}}{n}\right)^{2r/(d+2)} \lesssim \left(\frac{\log Mn}{n}\right)^{2r/(d+2)}$$

Bounds in expectation can be easily improved by a simple modification of the procedure. Namely, fix some $J \in \mathbb{N}$ and define $\delta_j = 2^{-j}\delta$, $1 \leqslant j \leqslant J$. For each j let $\widehat{\theta}^{[j]}_m$ stand for a MSSA estimate with parameters $z^{[j]}_k = z_k(\delta_j)$ defined by the formula (15). Finally, denote

$$\widehat{\theta}^{\langle J\rangle}_m = \left(2^{J+1} - 1\right)^{-1} \sum_{j=1}^{J} 2^j \widehat{\theta}^{[j]}_m \tag{21}$$

A rigorous result is formulated in the next theorem.

Theorem 3. *Under assumptions of Theorem 2, the choice $J = \lceil \frac{2r}{d+2} \log_2 K \rceil$ ensures that*

$$\mathbb{E} \max_{1 \leqslant m \leqslant M} \mathcal{K}^r\left(\theta^*_m, \widehat{\theta}^{\langle J\rangle}_m\right) \leqslant C_7^r \left(\frac{2r}{d+2}\right)^{1+2r/(d+2)} \left(\frac{\log 8KM}{n}\right)^{2r/(d+2)}$$

for all $r \geqslant 1$.

The proof of this result is given in Section 4.4. Note that the modified procedure requires running the MSSA algorithm $O(\log \log n)$ times and does not have a significant influence on the computational time.

With guarantees on the performance of estimation, we are ready to provide bounds on the excess risk of misclassification. Now we assume that a test point X is drawn randomly according to distribution $\mathbb{P}_X$ and Y has a conditional distribution (1).

Theorem 4. *Let $S_n = \{(X_i, Y_i)\}_{i=1}^n$ be a training sample with independent entries and (X, Y) is a test point generated from the distribution $\mathcal{D} = \mathbb{P}(Y|X)\mathbb{P}_X$ and $\mathbb{P}(Y|X)$ is given by (1). Let the multiclass low noise assumption (11) be fulfilled and supposethat for each realization of S_n and (X, Y) a collection of localizing schemes $W_1, \dots, W_K$ is chosen in a way to ensure (A1) and (A3) with probability 1. Suppose that (A2) holds. Choose a constant $\alpha \in (0, 1)$ from the condition (14) and set parameters z_k according to (15). Let*

$$|\theta_m^*(x+t) - \theta^*(x)| \leqslant Lh, \quad \forall\, |t| \leqslant h,\ 1 \leqslant m \leqslant M$$

and select

$$h = C_8 \left(\frac{\log n}{L^2 n}\right)^{1/(d+2)} \tag{22}$$

Suppose that for each m from 1 to M there exists a localizing scheme $W_{k^*(m)}$, $1 \leqslant k \leqslant K$ such that

$$w_i^{(k^*(m))} = 0, \quad \forall\, i : |X_i - x| > h$$

and

$$\alpha_1 n h^d \leqslant N_{k^*(m)} \leqslant \alpha_2 n h^d \tag{23}$$

for some positive constants $0 < \alpha_1 \leqslant \alpha_2$. Then for the excess risk $\mathcal{E}(\widehat{f})$ one has

$$\mathcal{E}(\widehat{f}) \leqslant B C_9^{1+\alpha} (13 + 12\alpha) \left(\frac{2+\alpha}{d+2}\right)^{1+(1+\alpha)/(d+2)} \left(\frac{\log 8KM}{n}\right)^{(1+\alpha)/(d+2)}$$

for some positive constant C_9.

4.2. Proof of Theorem 1

We start with an auxiliary result, which were obtained in [21] and [6] respectively.

Lemma 1. *Let (A1) and (A2) be fulfilled. Then for the likelihood estimate $\widetilde{\theta}_m$ of the form (7) and arbitrary $z > 0$ it holds*

$$\mathbb{P}\left(N\mathcal{K}\left(\widetilde{\theta}_m, \overline{\theta}_m\right) > z\right) \leqslant 2e^{-z/\varkappa^2}$$

Lemma 2. *Under assumptions (A1)–(A3), it holds*

$$N_k \mathcal{K}\left(\widehat{\theta}_m^{(k)}, \widehat{\theta}_m^{(k-1)}\right) \leqslant (1+b) z_k$$

Moreover,

$$N_{k'} \mathcal{K}\left(\widehat{\theta}_m^{(k')}, \widehat{\theta}_m^{(k)}\right) \leqslant (1+b)\varkappa^2 c_u^2 \overline{z}_k, \quad \forall\, k' > k$$

where $c_u = (u^{-\frac{1}{2}} - 1)^{-1}$ and $\overline{z}_k = \max_{l \geqslant k} z_l$.

We also use a reparametrization

$$\nu = \log \frac{\theta}{1-\theta} \tag{24}$$

throughout the proof.

Proof of Theorem 1. Let us fix an arbitrary m from 1 to M and $\delta \in (0,1)$.

First, note that (14) and (15) imply

$$N_k \mathcal{K}\left(\overline{\theta}_m^{(k)}, \overline{\theta}_m^{(k-1)}\right) \leqslant (1-\alpha)^2 b z_k, \quad m = \overline{1, M},\, k = \overline{2, K}$$

Now we bound $\mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right)$. Note that

$$\begin{aligned}
&\mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \widetilde{\theta}_m^{(k-1)}\right) \leqslant \sqrt{\frac{\kappa_2}{2}} \left|\widetilde{\nu}_m^{(k)} - \widetilde{\nu}_m^{(k-1)}\right| \\
&\leqslant \sqrt{\frac{\kappa_2}{2}} \left(\left|\widetilde{\nu}_m^{(k)} - \overline{\nu}_m^{(k)}\right| + \left|\widetilde{\nu}_m^{(k-1)} - \overline{\nu}_m^{(k-1)}\right| + \left|\overline{\nu}_m^{(k)} - \overline{\nu}_m^{(k-1)}\right|\right) \\
&\leqslant \varkappa \left(\mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \overline{\theta}_m^{(k)}\right) + \mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k-1)}, \overline{\theta}_m^{(k-1)}\right) + \mathcal{K}^{1/2}\left(\overline{\theta}_m^{(k)}, \overline{\theta}_m^{(k-1)}\right)\right)
\end{aligned}$$

Then it holds

$$\begin{aligned}
&\mathbb{P}\left(N_k \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \widetilde{\theta}_m^{(k-1)}\right) > b z_k\right) \leqslant \mathbb{P}\left(\varkappa \mathcal{K}\mathcal{L}^{\frac{1}{2}}\left(\widetilde{\theta}_m^{(k)}, \overline{\theta}_m^{(k)}\right)\right. \\
&\left. + \varkappa \mathcal{K}\mathcal{L}^{\frac{1}{2}}\left(\widetilde{\theta}_m^{(k-1)}, \overline{\theta}_m^{(k-1)}\right) + \varkappa \mathcal{K}\mathcal{L}^{\frac{1}{2}}\left(\overline{\theta}_m^{(k)}, \overline{\theta}_m^{(k-1)}\right) > \sqrt{\frac{b z_k}{N_k}}\right) \\
&\leqslant \mathbb{P}\left(\varkappa \mathcal{K}\mathcal{L}^{\frac{1}{2}}\left(\widetilde{\theta}_m^{(k)}, \overline{\theta}_m^{(k)}\right) + \varkappa \mathcal{K}\mathcal{L}^{\frac{1}{2}}\left(\widetilde{\theta}_m^{(k-1)}, \overline{\theta}_m^{(k-1)}\right) > \alpha \sqrt{\frac{b z_k}{N_k}}\right) \\
&\leqslant \mathbb{P}\left(2\varkappa^2 \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \overline{\theta}_m^{(k)}\right) + 2\varkappa^2 \mathcal{K}\left(\widetilde{\theta}_m^{(k-1)}, \overline{\theta}_m^{(k-1)}\right) > \frac{\alpha^2 b z_k}{N_k}\right) \\
&\leqslant \mathbb{P}\left(2\varkappa^2 \mathcal{K}\left(\widetilde{\theta}_m^{(k-1)}, \overline{\theta}_m^{(k-1)}\right) > \frac{\alpha^2 b z_k}{2N_k}\right) + \mathbb{P}\left(2\varkappa^2 \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \overline{\theta}_m^{(k)}\right) > \frac{\alpha^2 b z_k}{2N_k}\right) \\
&\leqslant \mathbb{P}\left(N_{k-1} \mathcal{K}\left(\widetilde{\theta}_m^{(k-1)}, \overline{\theta}_m^{(k-1)}\right) > \frac{\alpha^2 u_0 b z_k}{4\varkappa^2}\right) \\
&+ \mathbb{P}\left(N_k \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \overline{\theta}_m^{(k)}\right) > \frac{\alpha^2 b z_k}{4\varkappa^2}\right) \leqslant 4 e^{-\frac{\alpha^2 u_0 b z_k}{4\varkappa^4}}
\end{aligned}$$

Define an event $A_k = \left\{N_k \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \widetilde{\theta}_m^{(k-1)}\right) \leqslant b z_k\right\}$. It is clear that on the event $\bigcap\limits_{i=1}^{k} A_i$ it holds $\widetilde{\theta}_m^{(k)} = \widehat{\theta}_m^{(k)}$. Then, due to the union bound, we have

$$\mathbb{P}\left(\widetilde{\theta}_m^{(k)} \neq \widehat{\theta}_m^{(k)}\right) \leqslant \mathbb{P}\left(\bigcup_{j=1}^{k} \overline{A}_j\right) \leqslant \sum_{j=1}^{k} \mathbb{P}\left(\overline{A}_j\right) \leqslant \sum_{j=1}^{k} 4 e^{-\alpha^2 u_0 b z_j/(4\varkappa^4)}$$

One can easily check that the choice

$$z_k = \frac{4\varkappa^4}{\alpha^2 u_0 b} \log \frac{4KM}{\delta} \tag{25}$$

ensures

$$\mathbb{P}\left(\widetilde{\theta}_m^{(k)} \neq \widehat{\theta}_m^{(k)}\right) \leqslant \sum_{j=1}^{k} \frac{\delta}{KM} \leqslant \frac{\delta}{M}, \quad \forall\, k = \overline{1, K}$$

Further we work on the event $\left\{\widetilde{\theta}_m^{(k)} = \widehat{\theta}_m^{(k)}\right\}$, which has a probability measure at least $1 - \frac{\delta}{M}$. On this event

$$\mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right) = \mathcal{K}^{1/2}\left(\widehat{\theta}_m^{(k)}, \widehat{\theta}_m\right) \leqslant \sqrt{\frac{\kappa_2}{2}} \left|\widehat{\nu}_m^{(k)} - \widehat{\nu}_m\right|$$

$$\leqslant \sqrt{\frac{\kappa_2}{2}} \sum_{j=k}^{K-1} \left|\widehat{\nu}_m^{(j+1)} - \widehat{\nu}_m^{(j)}\right| \leqslant \sum_{j=k}^{K-1} \varkappa \mathcal{K}^{1/2}\left(\widehat{\theta}_m^{(j+1)}, \widehat{\theta}_m^{(j)}\right)$$

This and Lemma 2 imply

$$\mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right) \leqslant \sum_{j=k}^{K-1} \varkappa \sqrt{(1+b)\frac{z_{j+1}}{N_{j+1}}} \leqslant \varkappa \sqrt{(1+b) z_k} \sum_{j=k+1}^{K} \frac{1}{\sqrt{N_j}}$$

$$\leqslant \varkappa \sqrt{(1+b)\frac{z_k}{N_k}} \sum_{j=1}^{K-k+1} \frac{1}{\sqrt{u_0}} \leqslant \frac{\varkappa}{1 - \sqrt{u_0}} \sqrt{(1+b) u_0 \frac{z_k}{N_k}}$$

Using (15), one obtains that with probability at least $1 - \frac{\delta}{M}$

$$\mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right) \leqslant \sqrt{\frac{1+b}{b}} \cdot \frac{2\varkappa^3}{\alpha(1 - \sqrt{u_0})} \sqrt{\frac{\log \frac{4KM}{\delta}}{N_k}} \tag{26}$$

for every fixed m. The union bound implies that (26) holds simultaneously for all m from 1 to M with probability at least $1 - \delta$.

Second, we show that for all k from 1 to K it holds

$$\mathcal{K}^{1/2}\left(\theta_m^*, \widehat{\theta}_m\right) \leqslant \mathcal{K}^{1/2}\left(\theta_m^*, \widetilde{\theta}_m^{(k)}\right) + \varkappa^2 \mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right) \tag{27}$$

Use reparametrization $\theta = \theta(\nu)$ and observe that for all k

$$\begin{aligned}
\mathcal{K}\left(\theta_m^*, \widehat{\theta}_m\right) - \mathcal{K}\left(\theta_m^*, \widetilde{\theta}_m^{(k)}\right) &= \mathcal{K}\left(\nu_m^*, \widehat{\nu}_m\right) - \mathcal{K}\left(\nu_m^*, \widetilde{\nu}_m^{(k)}\right) \\
&= D'(\nu_m^*)\left(\widetilde{\nu}_m^{(k)} - \widehat{\nu}_m\right) - D(\widetilde{\nu}_m^{(k)}) + D(\widehat{\nu}_m) \\
&\leqslant \mathcal{K}\left(\widetilde{\nu}_m^{(k)}, \widehat{\nu}_m\right) + (D'(\nu_m^*) - D'(\widetilde{\nu}_m^{(k)}))(\widetilde{\nu}_m^{(k)} - \widehat{\nu}_m) \\
&\leqslant \mathcal{K}\left(\widetilde{\nu}_m^{(k)}, \widehat{\nu}_m\right) + \kappa_2 |\nu_m^* - \widetilde{\nu}_m^{(k)}||\widetilde{\nu}_m^{(k)} - \widehat{\nu}_m| \\
&\leqslant \mathcal{K}\left(\widetilde{\nu}_m^{(k)}, \widehat{\nu}_m\right) + 2\varkappa^2 \mathcal{K}^{1/2}\left(\nu_m^*, \widetilde{\nu}_m^{(k)}\right) \mathcal{K}^{1/2}\left(\widetilde{\nu}_m^{(k)}, \widehat{\nu}_m\right) \\
&= \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right) + 2\varkappa^2 \mathcal{K}^{1/2}\left(\theta_m^*, \widetilde{\theta}_m^{(k)}\right) \mathcal{K}^{1/2}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right)
\end{aligned}$$

Thus, (27) holds. This and (26) imply (16).

Now fix $\varepsilon > 0$. Using the Cauchy-Schwartz inequality one obtains

$$\mathcal{K}\left(\theta_m^*, \widehat{\theta}_m\right) \leqslant (1+\varepsilon)\mathcal{K}\left(\theta_m^*, \widetilde{\theta}_m^{(k)}\right) + \varkappa^4 \left(1 + \frac{1}{\varepsilon}\right) \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right)$$

Similarly for $(\theta_m^* - \widehat{\theta}_m)^2$ one has

$$|\theta_m^* - \widehat{\theta}_m| - |\theta_m^* - \widetilde{\theta}_m^{(k)}| \leqslant |\widetilde{\theta}_m^{(k)} - \widehat{\theta}_m|$$

and

$$\begin{aligned}(\theta_m^* - \widehat{\theta}_m)^2 \leqslant &(1+\varepsilon)(\theta_m^* - \widetilde{\theta}_m^{(k)})^2 + \left(1 + \frac{1}{\varepsilon}\right)(\widetilde{\theta}_m^{(k)} - \widehat{\theta}_m)^2 \leqslant \\ &(1+\varepsilon)(\theta_m^* - \widetilde{\theta}_m^{(k)})^2 + 2\varkappa^2\kappa_2 \left(1 + \frac{1}{\varepsilon}\right) \mathcal{K}\left(\widetilde{\theta}_m^{(k)}, \widehat{\theta}_m\right)\end{aligned}$$

and (17) follows.

$\square$

4.3. Proof of Theorem 2

From the Theorem 1 with $\varepsilon = 1$ on event with probability at least $1-\delta$ we have for all m

$$\left(\theta_m^* - \widehat{\theta}_m\right)^2 \leqslant 2\left(\theta_m^* - \widetilde{\theta}_m^{(k^*(m))}\right)^2 + \frac{C}{N_{k^*(m)}} \log \frac{4KM}{\delta}$$

with a constant $C > 0$.

Consider $\left(\theta_m^* - \widetilde{\theta}^{(k^*(m))}\right)^2$:

$$|\theta_m^* - \widetilde{\theta}^{(k^*(m))}| \leqslant |\theta_m^* - \overline{\theta}^{(k^*(m))}| + |\overline{\theta}^{(k^*(m))} - \widetilde{\theta}^{(k^*(m))}|$$

It holds

$$\begin{aligned}|\theta_m^* - \overline{\theta}^{(k^*(m))}| &= \left|\sum_{i=1}^n \frac{w_i^{(k^*(m))}}{N_{k^*(m)}} (\theta_{m,i} - \theta_m^*)\right| \\ &= \left|\sum_{i:\|X_i - X\| \leqslant h_{k^*(m)}} \frac{w_i^{(k^*(m))}}{N_{k^*(m)}} (\theta_{m,i} - \theta_m^*)\right| \\ &\leqslant \sum_{i:\|X_i - X\| \leqslant h_{k^*(m)}} \frac{w_i^{(k^*(m))}}{N_{k^*(m)}} \left|\theta_{m,i} - \theta_m^*\right| \leqslant Lh\end{aligned}$$

Since Bernoulli random variables are sub-Gaussian with the variance proxy $\frac{1}{4}$, from the Hoeffding inequality we have

$$\mathbb{P}\left(N_{k^*(m)}|\overline{\theta}^{(k^*(m))} - \widetilde{\theta}^{(k^*(m))}| > t\right) \leqslant 2\exp\left(-\frac{2t^2}{\sum\limits_{i=1}^n \left[w_i^{(k^*(m))}\right]^2}\right) \leqslant 2\exp\left(-\frac{2t^2}{N_{k^*(m)}}\right)$$

and then with probability at least $1 - \frac{\delta}{M}$ it holds

$$\left(\overline{\theta}^{(k^*(m))} - \widetilde{\theta}^{(k^*(m))}\right)^2 \leqslant \frac{1}{2N_{k^*(m)}} \log\frac{2M}{\delta}$$

Using the union bound, one has

$$\left(\overline{\theta}^{(k^*(m))} - \widetilde{\theta}^{(k^*(m))}\right)^2 \leqslant \frac{1}{2N_{k^*(m)}} \log\frac{2M}{\delta}$$

simultaneously for all m from 1 to M on the event with probability at least $1-\delta$. Again, using the union bound, with probability at least $1-2\delta$ it holds

$$\left(\theta_m^* - \widehat{\theta}_m\right)^2 \leqslant 4L^2 h_{k^*(m)}^2 + \frac{2}{N_{k^*(m)}}\log\frac{2M}{\delta} + \frac{C}{N_{k^*(m)}}\log\frac{4KM}{\delta}$$

simultaneously for all m from 1 to M and the claim of Theorem 2 follows from (19) and (20).

4.4. *Proof of Theorem 3*

Denote $g_m(x_m) = \mathcal{K}(\theta_m^*, x_m)$. Note that $g_m(x_m)$, $1 \leqslant m \leqslant M$, are convex functions on $[0,1]$. Since for all $r \geqslant 1$ $h(x) = x^r$ is convex increasing function on $[0,+\infty)$ then $g_m^r(x_m)$, $1 \leqslant m \leqslant M$, are convex too. Finally, the function $\max\limits_{1\leqslant m\leqslant M}\{g_1(x_1),\dots,g_M(x_M)\}$ is convex and this fact implies

$$\begin{aligned}
\mathbb{E}\max_{1\leqslant m\leqslant M}\mathcal{K}^r\left(\theta_m^*, \widehat{\theta}_m^{\langle J\rangle}\right) &\leqslant \sum_{j=1}^J \frac{2^j}{2^{J+1}-2}\mathbb{E}\max_{1\leqslant m\leqslant M}\mathcal{K}^r\left(\theta_m^*, \widehat{\theta}_m^{[j]}\right) \\
&\leqslant \sum_{j=1}^J 2^{j-J}\left[\frac{\varkappa^{2r}}{2^r\kappa_1^r}\delta_j + C_5^r\left(\frac{\log(8KM/\delta_j)}{n}\right)^{2r/(d+2)}\right] \\
&\leqslant \sum_{j=1}^J 2^{-J}\left[\frac{\varkappa^{2r}}{2^r\kappa_1^r}\delta + C_5^r\left(\frac{j\log 2 + \log(8KM/\delta)}{n}\right)^{2r/(d+2)}\right] \\
&\leqslant J2^{-J}\left[\frac{\varkappa^{2r}}{2^r\kappa_1^r}\delta + C_5^r\left(\frac{J\log 2 + \log(8KM/\delta)}{n}\right)^{2r/(d+2)}\right]
\end{aligned}$$

Now choose $J = \lceil \frac{2r}{d+2} \log_2 K \rceil$ and $\delta = \left(\frac{K}{n}\right)^{\frac{2r}{d+2}} \cdot \log_2^{2r/(d+2)-1} K$. Then

$$
\begin{aligned}
&\mathbb{E} \max_{1 \leqslant m \leqslant M} \mathcal{K}^r \left(\theta^*_m, \widehat{\theta}^{\langle J \rangle}_m \right) \\
&\leqslant \frac{1 + \frac{2r}{d+2} \log_2 K}{K^{2r/(d+2)}} \left[\frac{\varkappa^{2r}}{2^r \kappa_1^r} \left(\frac{K}{n} \right)^{2r/(d+2)} \log_2^{2r/(d+2)-1} K \right. \\
&\left. + C_5^r \left(\frac{1 + \frac{2r}{d+2} \log n + \log 8KM + (1 - 2r/(d+2)) \log\log K}{n} \right)^{2r/(d+2)} \right] \\
&\leqslant C_7^r \left(\frac{2r}{d+2} \right)^{1+2r/(d+2)} \left(\frac{\log 8KM}{n} \right)^{2r/(d+2)},
\end{aligned}
$$

and the last inequality holds since $\log n \lesssim K$.

4.5. Proof of Theorem 4

For the excess risk $\mathcal{E}(f)$ the next representation from [14] will be useful.

Lemma 3. *For any classifier $f(X)$ we have*

$$
\mathcal{E}(f) = R(f) - R(f^*) = \mathbb{E}\left[\theta^*_{f^*(X)}(X) - \theta^*_{f(X)}(X) \right]
$$

The next theorem guarantees that under the low noise assumption (11) the MSSA procedure achieves fast rates.

Lemma 4. *Let the low noise condition (11) be fulfilled. Choose $r > 1 + \alpha$ and set z_k as in the formula (15). Suppose that for this r and for all m from 1 to M it holds*

$$
\mathbb{E} \max_{1 \leqslant m \leqslant M} |\widehat{\theta}_m - \theta^*_m|^r \leqslant \mu_r
$$

Then for the excess risk $\mathcal{E}(\hat{f}_n)$ it holds

$$
\mathcal{E}(\hat{f}_n) \leqslant B \left(1 + \frac{6(r + \alpha + 2)}{r - \alpha - 1} \right) \mu_r^{\frac{1+\alpha}{r}}
$$

The proof of this result is given in Appendix A and uses the same idea of slicing as in [5] and [14] with a significant modification that unlike in those papers, we prove a similar result for the procedure, which confidence set levels depend on its parameters z_k.

Note that the Theorem 3 guarantees that

$$
\mu_r \leqslant C_9^r \left(\frac{r}{d+2} \right)^{\frac{r}{d+2}+1} \left(\frac{\log 8KM}{n} \right)^{\frac{r}{d+2}}
$$

and this result for $r = 2 + \alpha$ combined with Lemma 4 finishes the proof of Theorem 4.

References

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

[2] Allwein, E. L., Schapire, R. E. and Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *J. Mach. Learn. Res.* **1** 113–141. MR1884092

[3] Aly, M. (2005). Survey on multiclass classification methods. *Neural networks* 1-9.

[4] Athitsos, V. and Sclaroff, S. (2005). Boosting nearest neighbor classifiers for multiclass recognition. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops* 45-45.

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

[6] Belomestny, D. and Spokoiny, V. (2007). Spatial aggregation of local likelihood estimates with applications to classification. *Ann. Statist.* **35** 2287–2311. MR2363972

[7] Butucea, C., Delmas, J.-F., Dutfoy, A. and Fischer, R. (2017). Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Electron. J. Stat.* **11** 2258–2294. MR3654825

[8] Crammer, K. and Singer, Y. (2002). On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *J. Mach. Learn. Res.* **2** 265–292.

[9] Dai, D., Rigollet, P. and Zhang, T. (2012). Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.* **40** 1878–1905. MR3015047

[10] Daniely, A., Sabato, S. and Shwartz, S. S. (2012). Multiclass Learning Approaches: A Theoretical Comparison with Implications. In *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 485–493. Curran Associates, Inc.

[11] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A probabilistic theory of pattern recognition. Applications of Mathematics (New York)* **31**. Springer-Verlag, New York. MR1383093

[12] Dheeru, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository.

[13] Dietterich, T. G. and Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-correcting Output Codes. *J. Artif. Int. Res.* **2** 263–286.

[14] Dinh, V., Ho, L. S. T., Cuong, N. V., Nguyen, D. and Nguyen, B. T. (2015). Learning from non-iid data: fast rates for the one-vs-all multiclass plug-in classifiers. In *Theory and applications of models of computation. Lecture Notes in Comput. Sci.* **9076** 375–387. Springer, Cham. MR3368631

[15] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712. MR1792783

[16] Juditsky, A., Rigollet, P. and Tsybakov, A. B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. MR2458184

[17] Lecué, G. (2006). Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.* **7** 971–981. MR2274393

[18] Lecué, G. (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli* **13** 1000–1022. MR2364224

[19] Lecué, G. (2013). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli* **19** 2153–2166. MR3160549

[20] Lecué, G. and Rigollet, P. (2014). Optimal learning with Q-aggregation. *Ann. Statist.* **42** 211–224. MR3178462

[21] Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation. *Probab. Theory Related Fields* **135** 335–362. MR2240690

[22] Reitmaier, T., Calma, A. and Sick, B. (2015). Transductive active learning A new semi-supervised learning approach based on iteratively refined generative models to capture structure in data. *Information Sciences* **293** 275 - 298.

[23] Rifkin, R. and Klautau, A. (2003/04). In defense of one-vs-all classification. *J. Mach. Learn. Res.* **5** 101–141. MR2247975

[24] Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639–665. MR2933661

[25] Rigollet, P. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. MR2356821

[26] Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. MR3025134

[27] Rubinstein, B. I., Bartlett, P. L. and Rubinstein, J. H. (2007). Shifting, One-Inclusion Mistake Bounds and Tight Multiclass Expected Risk Bounds. In *Advances in Neural Information Processing Systems 19* (B. Schölkopf, J. C. Platt and T. Hoffman, eds.) 1193–1200. MIT Press.

[28] Tsybakov, A. B. (2003). *Optimal Rates of Aggregation* In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings* 303–313. Springer Berlin Heidelberg, Berlin, Heidelberg.

[29] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[30] Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inform. Theory* **45** 2271–2284. MR1725115

[31] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47. MR2044592

[32] Yuditskiĭ, A. B., Nazin, A. V., Tsybakov, A. B. and Vayatis, N. (2005). Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii* **41** 78–96. MR2198228

Appendix A: Proof of Lemma 4

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

$$A_i = \left\{ \beta^{i-1} t < \theta^*_{f^*(X)}(X) - \theta^*_{\widehat{f}(X)}(X) \leqslant \beta^i t \right\}, \quad i \geqslant 0$$

Then

$$\begin{aligned}
R(\widehat{f}) - R(f^*) &= \mathbb{E}\left[\theta^*_{f^*(X)}(X) - \theta^*_{\widehat{f}(X)}(X) \right] \\
&= \mathbb{E}\left[\theta^*_{f^*(X)}(X) - \theta^*_{\widehat{f}(X)}(X) \right] \mathbb{1}\left(f^*(X) \neq \widehat{f}(X) \right) \\
&= \sum_{i=0}^{\infty} \mathbb{E}\left[\theta^*_{f^*(X)}(X) - \theta^*_{\widehat{f}(X)}(X) \right] \mathbb{1}(f^*(X) \neq \widehat{f}(X)) \mathbb{1}(A_i) \\
&\leqslant t \mathbb{P}\left\{ 0 < \theta^*_{f^*(X)}(X) - \theta^*_{\widehat{f}(X)}(X) \leqslant t \right\} \\
&+ \sum_{i=1}^{\infty} \beta^i t \mathbb{E}\left[\mathbb{1}(f^*(X) \neq \widehat{f}(X)) \mathbb{1}(A_i) \right]
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\mathbb{1}\Big(f^*(X) \neq \widehat{f}(X) \Big) \mathbb{1}(A_i) &\leqslant \mathbb{E}\mathbb{1}\Big(\theta^*_{f^*(X)}(X) \leqslant \theta^*_{\widehat{f}(X)}(X) \\
&+ |\widehat{\theta}_{f^*(X)}(X) - \theta^*_{f^*(X)}(X)| + |\widehat{\theta}_{\widehat{f}(X)}(X) - \theta^*_{\widehat{f}(X)}(X)| \Big) \mathbb{1}(A_i) \\
&\leqslant \mathbb{E}\mathbb{1}\Big(|\widehat{\theta}_{f^*(X)}(X) - \theta^*_{f^*(X)}(X)| \\
&+ |\widehat{\theta}_{\widehat{f}(X)}(X) - \theta^*_{\widehat{f}(X)}(X)| \geqslant \beta^{i-1} t \Big) \mathbb{1}(A_i) \\
&\leqslant \mathbb{E}\mathbb{1}\Big(|\widehat{\theta}_{f^*(X)}(X) - \theta^*_{f^*(X)}(X)| \geqslant \beta^{i-2} t) \\
&+ \mathbb{1}(|\widehat{\theta}_{\widehat{f}(X)}(X) - \theta^*_{\widehat{f}(X)}(X)| \geqslant \beta^{i-2} t \Big) \mathbb{1}(A_i) \\
&\leqslant 2\mathbb{E}\mathbb{1}\left(\max_{1 \leqslant m \leqslant M} |\widehat{\theta}_m(X) - \theta^*_m(X)| \geqslant \beta^{i-2} t \right) \mathbb{1}(A_i) \\
&\leqslant 2\mathbb{P}\left(\max_{1 \leqslant m \leqslant M} |\widehat{\theta}_m(X) - \theta^*_m(X)| \geqslant \beta^{i-2} t \right) \mathbb{P}(A_i) \\
&\leqslant 2\mathbb{P}(A_i) \frac{\mu_r}{\beta^{r(i-2)} t^r} \leqslant \frac{2\mu_r}{\beta^{r(i-2)} t^r} \cdot B \beta^{\alpha i} t^\alpha = 2B\beta^{2r} \cdot \frac{\mu_r}{(\beta^i t)^{r-\alpha}}
\end{aligned}$$

Then

$$
\begin{aligned}
R(\widehat{f}) - R(f^*) &\leqslant Bt^{1+\alpha} + \sum_{i=1}^{\infty} 2B\beta^{2r} \cdot \frac{\mu_r}{(\beta^i t)^{r-\alpha-1}} \\
&= Bt^{1+\alpha} \left(1 + \sum_{i=1}^{\infty} 2\beta^{2r} \cdot \frac{\mu_r}{\beta^{i(r-\alpha-1)} t^r} \right) \\
&= Bt^{1+\alpha} \left(1 + \sum_{i=1}^{\infty} 2\beta^{2r} \cdot \frac{\mu_r}{\beta^{i(r-\alpha-1)} t^r} \right) \\
&= Bt^{1+\alpha} \left(1 + \frac{2\beta^{2r}\mu_r}{(\beta^{r-\alpha-1} - 1)t^r} \right) \\
&\leqslant Bt^{1+\alpha} \left(1 + \frac{2(r+\alpha+2)\beta^{r+\alpha+1}\mu_r}{(r-\alpha-1)t^r} \right)
\end{aligned}
$$

Note that

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

Now, the choice

$$
t = \mu_r^{\frac{1}{r}},
$$

implies the assertion of theorem.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.
002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.
003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.
004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.
005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.
006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.
007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.
008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.
009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.
010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.
011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.
012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.
013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.
014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.
015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.
016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.
017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.
018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.
019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.
021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.
022 " Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.
023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.
024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.
025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.
026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.
027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.
028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.
029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.