National Research University Higher School of Economics

Faculty of Mathematics

As a manuscript

Nikita Puchkin

Adaptive weights methods for minimax optimal estimation in problems of supervised and unsupervised learning

Summary of the PhD Thesis
for the purpose of obtaining academic degree
Doctor of Philosophy in Mathematics

Academic Supervisor
PhD
Vladimir Spokoiny

Moscow – 2023

Table of Contents

Introduction

The problem of pattern recognition in large datasets have recently attracted huge attention. The problem of statistical learning finds more and more applications in the era of big data. The diversity of areas of utilization is also impressive: from bioinformatics (e.g., protein fold prediction [1]) and finance (credit rating [2] and bankruptcy [3] prediction) to speech recognition [4] and image analysis [5, 6] and restoration [7, 8]. Such an interest caused a substantial progress in theoretical investigation of the existing methods of machine learning and motivated researches develop new, more efficient ones. Analysis of learning algorithms usually relies on the tools from probability theory and mathematical statistics. That is, a learner assumes that the data he possesses is generated from a statistical model, and he must perform an inference based on this model.

Speaking of statistical learning problems, one can distinguish between three major groups: supervised, semi-supervised, and unsupervised learning. In supervised learning problems, each instance of a given sample has a label, also referred to as a response or a target variable (usually, it is a real number). The goal of the learner is to predict a label of a newly arrived instance, which is not presented in the training sample. The most popular examples of supervised learning problems are classification and regression. In opposite, in an unsupervised learning problem, the learner has to recognize patterns in unlabelled data. The most common examples of unsupervised learning problems include clustering and manifold learning. Finally, in the semi-supervised setup the learner has to perform an inference based on a small portion of labelled instances and a large number of unlabelled ones.

The present thesis pursues methological and theoretical purposes. It aims at developing new adaptive algorithms for supervised and unsupervised learning problems with strong theoretical guarantees on their performance. Chapter 1 is devoted multiclass classification. We suggest an adaptive multiclass nearest neighbor classifier, Algorithm 1. Though the statisticians are familiar with k-nearest neighbor rules for a long time, their nonasymptotic analysis was performed quite recently. In [9], the author proved a minimax optimal upper bound on the excess risk of a weighted nearest neighbor classifier. Unfortunately, the approach of [9] requires quite restrictive assumptions. For instance, the author assumed that the distribution of feature vectors satisfies the so-called strong density assumption. This issue was partially addressed in [10] and [11], where the authors introduced a novel variant of smoothness of the target function, connecting the distribution of feature vectors and labels. Major steps towards understanding the limitations of k-nearest neighbor rules were made in [12] and [13]. In particular, in [12] the authors showed that a universal choice of k in the nearest neighbor classifier leads to strictly suboptimal performance under quite realistic assumptions. In [13], the authors suggested to use auxiliary unlabelled data for the point-dependent choice of k. However, this approach is not always applicable, since the unlabelled data may be unavailable to the statistician. In the present thesis, we develop the ideas of spatial stagewise aggregation [3] and suggest an adaptive weights method based on combination of nearest neighbor estimates. Though the extension of the algorithm in [3] to the multiclass case is rather straightforward, its theoretical analysis is much more involved. In contrast to [3], we impose much weaker assumptions and show that, under the same conditions as in [12], our procedure attains minimax optimal rates of convergence up to a logarithmic factor (see Theorem 1). Unlike [13], our procedure does not involve any auxiliary data. To our knowledge, this is the first such estimator in the literature.

Chapter 2 is devoted to a manifold learning problem, that is, estimation of a smooth low-dimensional submanifold in $\mathbb{R}^D$ from noisy observations. This problem was extensively studied in the literature. Unfortunately, the existing methods of manifold estimation either require the noise magnitude be extremely small (e.g., [8, 14–16]) or assume the noise distribution is known (e.g. [17, 18]). In the present thesis, we focus on a setup, which was not previously considered in the literature. Namely, in contrast to [17, 18], we impose mild assumptions on the noise distribution and we allow the noise magnitude be much larger than in [8, 14–16, 19]. Though our assumptions are quite realistic, they significantly differ from the usual ones. Hence, it is not surprising that the existing algorithms either have no theoretical guarantees in our setup or show suboptimal performance. We faced a challenging problem of suggesting an optimal manifold estimate under mild conditions on the noise distribution. In the present thesis, we extend the idea of structural adaptation [20, 21] and suggest a novel algorithm (Algorithm 2) for manifold denoising. The algorithm allows us to construct a manifold estimate with strong theoretical guarantees (see Theorem 4). We also prove a new minimax lower bound on the accuracy of manifold estimation (Theorem 5), which yields the optimality of our method.

The problems considered in Chapters 1 and 2 are very different, but the methods we proposed for these problems, Algorithm 1 and Algorithm 2, share similar ideas. The core of the algorithms is the so called adaptive weights approach, properly tailored to the problems of multiclass classification and manifold denoising. The fact that the suggested procedures are adaptive (that is, they implicitly perform a partial parameter tuning, simplifying the model selection) increase their practical value. We also would like to note that one may combine Algorithm 1 and Algorithm 2 and apply them to semi-supervised learning problems.

The contribution of the present thesis is as follows.

1. We propose an algorithm for multiclass classification, Algorithm 1, 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. Besides, it adapts to the smoothness of the target function.

2. We prove a large deviation bound on the excess risk of the estimate, returned by Algorithm 1, under mild assumptions. The obtained theoretical results are new in the literature and claim optimal accuracy of classification with only a logarithmic payment for adaptation.

3. We suggest a novel algorithm for manifold denoising, Algorithm 2, based on the idea of structural adaptation.

4. Based on Algorithm 2, we construct a new manifold estimate and prove a new upper bound on the accuracy of manifold estimation.

5. We provide a new minimax lower bound on the accuracy of manifold estimation, claiming the optimality of our procedure.

Main results of the thesis were presented at the following conferences, workshops, schools, and seminars.

1. Workshop “New frontiers in high-dimensional probability and statistics”, Moscow, February 23–24, 2018. Talk: “Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification”.

2. 12th International Vilnius Conference on Probability Theory and Mathematical Statistics and 2018 IMS Annual Meeting on Probability and Statistics, Vilnius, Lithuania, July 2–6, 2018. Poster: “Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification”

3. Research seminar “Structural Learning”, Moscow, October 11, 2018. Talk: “Manifold Learning”.

4. Winter school “New frontiers in high-dimensional probability and statistics 2”, Moscow, February 22–23, 2019. Talk: “Manifold estimation from noisy observations”.

5. 49th Saint-Flour Summer School, Saint-Flour, France, July 7–19, 2019. Talk: “Manifold estimation from noisy observations”.

6. Conference “Structural Inference in High-Dimensional Models 2”, Pushkin, Saint-Petersburg, August 26–30, 2019. Poster: “Structure-adaptive manifold estimation”.

7. HSE-Yandex Autumn School on Generative Models, Moscow, November 26–29, 2019. Talk: “Structure-adaptive manifold estimation”.

8. Research seminar “Structural Learning”, Moscow, December 3, 2019. Talk: “Sample complexity of learning a manifold with an unknown dimension”.

9. Conference “Mathematical Methods of Statistics”, Luminy, France, December 16–20, 2019. Talk: “Structure-adaptive manifold estimation”.

10. HSE Faculty of Computer Science conference on Machine Learning, Fundamental Research track, Moscow, November 18–20, 2020. Talk: “Structure-adaptive manifold estimation”.

Main results of the thesis were published in two papers in peer-reviewed journals [22, 23].

The thesis contents and the presented main results reflect the author’s personal contribution. The author prepared the results for publication in collaboration with the scientific advisor. The author’s contribution is primary. All the presented results were obtained personally by the author.

The thesis consists of introduction, 2 chapters, conclusion, and bibliography. Each chapter starts with the literature review on the relevant topic. The thesis is 100 pages long, including 95 pages of the text, 5 tables, and 5 figures. The bibliography is 5 pages long and it includes 83 items.

Chapter 1

Multiclass classification

1.1. Problem statement

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 $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $Y_i \in \mathcal{Y} = \{1, \ldots, M\}$, $1 \leqslant i \leqslant n$, $M > 2$. 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 $f : \mathcal{X} \to \{1, \ldots, M\}$ 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 [1] or when one wants to classify DNA microarrays [24], finance when one predicts a corporate credit rating [2], image analysis [5] when one tries to classify an object on an image, speech recognition [4], and others.

1.2. Nonparametric multiclass classification and literature review

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 $\mathbb{P}\left(Y = m | X\right)$, $1 \leqslant m \leqslant M$ by $\eta_m(X)$. For this model, the optimal classifier f^* can be found analytically

$$f^*(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \eta_m(X). \tag{1.1}$$

Unfortunately, true values $\eta_1(X), \ldots, \eta_M(X)$ 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 $\widehat{\eta}_1^{(NN)}(X), \ldots, \widehat{\eta}_M^{(NN)}(X)$ of $\eta_1(X), \ldots, \eta_M(X)$ 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 [9], 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 [9] 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 [10] and [11], authors address this issue. In [10], the authors introduce a novel Hölder-like smoothness condition on $\eta_1(x)$ tailored to nearest neighbor. This trick allows to avoid the strong density assumption and boundedness of features. The disadvantage of the modified smoothness condition in [10] is that it is implicit. Instead of this condition, in [12], 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 [13] and [12], 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 it complicates the tuning procedure. 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 [3], where an aggregation of binary classifiers was studied. Unlike many other aggregation procedures, such as exponential weighting [25–27], mirror averaging [28, 29], empirical risk minimization [30], and Q-aggregation [31, 32], 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 [3] 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, [33, 34]) 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 [35] and then in [36] 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.

1.3. Contribution

Main contributions of the present chapter 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 $\eta_1(\cdot), \dots, \eta_M(\cdot)$;

- 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.

1.4. Notation and model assumptions

We start with a simple observation. Introduce a function

$$\varphi(t) = \left(\frac{1}{2M} \vee t\right) \wedge \left(1 - \frac{1}{2M}\right). \tag{1.2}$$

It is easy to show that the composition

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

satisfies the equality

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

where, as before, f^* stands for the Bayes classifier. Hence, instead of $\eta_m(x)$, one can estimate $\theta_m(x)$ at x and then use a plug-in classifier

$$\widehat{f}(X) = \operatorname*{argmax}_{1 \leqslant m \leqslant M} \widehat{\theta}_m(X), \tag{1.3}$$

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

The problem is how to construct the estimates of $\theta_m(x)$, $1 \leqslant m \leqslant M$. Fix some m and transform the labels into binarized ones: $\mathbb{1}\left(Y_i = m\right)$. It is clear that

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

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 $\widetilde{\theta}^w_m(x) = \varphi(\widetilde{\eta}^w_m(x))$, where

$$\widetilde{\eta}^w_m(x) = \frac{\sum\limits_{i=1}^n w_i(X_i, x)\mathbb{1}(Y_i = m)}{\sum\limits_{i=1}^n w_i(X_i, x)} \equiv \frac{S^w_m(x)}{N_w(x)} \tag{1.4}$$

is a weighted nearest neighbor estimate of $\eta_m(x)$. Here we introduced the notations $S^w_m(x) = \sum\limits_{i=1}^n w_i(X_i, x)\mathbb{1}(Y_i = m)$, $N_w(x) = \sum\limits_{i=1}^n w_i(X_i, x)$. 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 chapter, we consider the weights of the following form:

$$w_i = w_i(X_i, x) = \mathcal{K}\left(\frac{\|X_i - x\|}{h}\right), \tag{1.5}$$

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

- $\mathcal{K}(t)$ is a non-increasing funciton,
- $\mathcal{K}(0) = 1,$ (A1)
- $\mathcal{K}(1) \geqslant \frac{1}{2},$
- $\mathcal{K}(t) = 0, \quad \forall\, t > 1.$

This assumption can be easily satisfied. First, note that the rectangular kernel $\mathcal{K}(t) = \mathbb{1}\left(0 \leqslant t \leqslant 1\right)$ meets these requirements and, therefore, (A1) holds for the case of ordinary nearest neighbor estimates. There are other examples of such kernels $\mathcal{K}$. For instance, one can easily check that Epanechnikov-like and Gaussian-like kernels, $\mathcal{K}(t) = (1 - t^2/2)\mathbb{1}(0 \leqslant t \leqslant 1)$ and $\mathcal{K}(t) = e^{-t^2/2}\mathbb{1}(0 \leqslant t \leqslant 1)$ respectively, fulfill (A1). It is also important to mention that here and further in this chapter, 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 [10].

The nearest neighbor estimate (1.4) 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 $a, b > 0$, and $0 < u_0 < u < 1$ such that

$$n_1 \leqslant a, \quad n_K \geqslant b n^{2/(d+2)}, \quad \text{and} \quad 2u_0 \leqslant \frac{n_{k-1}}{n_k} \leqslant \frac{u}{2}, \quad \text{for all } 1 \leqslant k \leqslant K. \tag{A2}$$

Each n_k induces a set of weights $w_1^{(k)}, \dots, w_n^{(k)}$ with

$$w_i^{(k)} = w_i^{(k)}(X_i, x) = \mathcal{K}\left(\frac{\|X_i - x\|}{h_k}\right), \tag{1.6}$$

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

$$\widetilde{\theta}_m^{(k)}(x) = \varphi\left(\widetilde{\eta}_m^{(k)}(x)\right) \equiv \left(\frac{1}{2M} \vee \widetilde{\eta}_m^{(k)}(x)\right) \wedge \left(1 - \frac{1}{2M}\right), \tag{1.7}$$

$$\widetilde{\eta}_m^{(k)}(x) = \frac{S_m^{(k)}(x)}{N_k(x)}, \tag{1.8}$$

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 [3] 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 (1.3). We will refer to the procedure as multiclass spatial stagewise aggregation (MSSA for short).

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)-Hölder continuous, that is, there exist $L > 0$ and $\alpha > 0$ such that for all $x, x' \in \mathcal{X}$ and $1 \leqslant m \leqslant M$ it holds that

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

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

$$\begin{aligned} R(f) &= 1 - \mathbb{E}_{(X,Y)\sim\mathcal{D}}\mathbb{1}(Y = f(X)) \\ &= 1 - \mathbb{E}_X\mathbb{P}(Y = f(X)|X) = 1 - \mathbb{E}_X\eta_{f(X)}(X). \end{aligned} \tag{1.9}$$

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 high probability. The well-known Mammen-Tsybakov noise condition [37] 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 that

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

This assumption can be extended to the multiclass case. For any x, 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 (1.4) for the multiclass classification can be formulated as follows (see [38, 39]): there exist $B > 0$ and $\beta \geqslant 0$ such that the following holds for all $t > 0$:

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

We will use this assumption to establish fast rates for the plug-in classifier $\widehat{f}(X)$ in Section 1.6.

There are two more requirements we need: the minimal mass assumption and the tail assumption introduced in [12]. The first one assumes that there exist $\varkappa > 0$ and $r_0 > 0$, such that

$$\mathbb{P}(X \in B(x, r)) \geqslant \varkappa p(x) r^d \quad \text{for all } r \in (0, r_0] \text{ and } x \in \operatorname{supp}(\mathbb{P}_X), \tag{A5}$$

where $B(x, r)$ stands for the Euclidean ball of radius r centered at x and $p(x)$ is the density of the marginal distribution $\mathbb{P}_X$ of X with respect to the Lebesgue 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 that

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

It was discussed in [12] (Theorem 4.1) that the conditions (A5) and (A6) are necessary for quantitative analysis of classifiers and cannot be removed.

One can highlight a simple case of a bounded away from zero density when for any $x \in \operatorname{supp}(\mathbb{P}_X)$ it holds that $p(x) \geqslant p_0 > 0$ 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 [10, 11]. In those papers, the authors assume that there are constants $L > 0$ and $\alpha \in (0, 1]$, such that for all $x, x' \in \mathcal{X}$ it holds that

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

This assumption ensures that in the regions with a small density $p(x)$ the function $\eta(x)$ is (L', α)-Hölder continuous with a small constant L'. An approach, considered in [12], relies on the 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, the univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$, exponential distribution $\operatorname{Exp}(\lambda)$, gamma-distribution $\operatorname{Gamma}(k, \lambda)$, 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 [12, Example 4.1] for the details). A special case, in which one may be interested in, is the case when $\operatorname{supp}(\mathbb{P}_X)$ is compact. In this case,

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

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 provide a simple sufficient condition to check (A6).

Proposition 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:

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

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

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

□

In what is going further, we require p in (A6) to be larger than $\alpha/(2\alpha + d)$. By Proposition 1, any $\mathbb{P}_X$, such that $\mathbb{E}\|X\|^r < \infty$ for some $r > \alpha d/(\alpha + d)$, satisfies (A6) with $p > \alpha/(2\alpha + d)$.

1.5. An adaptive weights method for multiclass classification

In this section, we present the multiclass spatial stagewise aggregation (MSSA) procedure, which is 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 $z_1, \dots, z_K$ will be referred to as critical values.

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 $O\left(Mn \log n\right)$ 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 $O\left(Mn \log n\right)$. If there are several test points, then the computations can be done in parallel.

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 (1.6) with a kernel $\mathcal{K}$ satisfying (A1) and calculate $\widetilde{\theta}_m^{(k)}(x)$ according to (1.7) and (1.8).
    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 $\widehat{\theta}_m^{(k)}(x) = \gamma_k \widetilde{\theta}_m^{(k)}(x) + (1 - \gamma_k)\widehat{\theta}_m^{(k-1)}(x)$.
    Put the final estimate $\widehat{\theta}_m(x) = \widehat{\theta}_m^{(K)}(x)$.
return the predicted label $\widehat{f}(x) = \operatorname*{argmax}_{1 \leqslant m \leqslant m} \left\{\widehat{\theta}_m(x)\right\}$.
```

1.6. Theoretical properties of the MSSA procedure

1.6.1. Main result

Theorem 1. *Grant the assumptions* (A1) – (A5) *and let* (A6) *hold with* $p > \alpha/(2\alpha + d)$. *Choose the parameters* $z_1, \dots, z_K$ *according to the formula*

$$z_k = \frac{8M^2}{u_0} \log \frac{12KM}{\delta_*}, \quad 1 \leqslant k \leqslant K, \tag{1.10}$$

where

$$\delta_* = \begin{cases} \left(\frac{M^3 \log n}{np_0}\right)^{\frac{\alpha(2+\beta)}{2\alpha+d}}, & \text{if } \exists\, p_0 : p(x) \geqslant p_0 \,\forall\, x \in supp(\mathbb{P}_X), \\ \psi_*^{r_*}, & \text{otherwise}, \end{cases} \tag{1.11}$$

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

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

Let $\widehat{\theta}_1(\cdot), \dots, \widehat{\theta}_M(\cdot)$ *be the corresponding MSSA estimates. Then, if the sample size* n *is sufficiently large, 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*

$$\mathbb{E}_{S_n} \mathcal{E}(\widehat{f}) \lesssim \begin{cases} \left(\frac{M^3 \log n}{np_0}\right)^{\frac{\alpha(1+\beta)}{2\alpha+d}}, & \text{if } \exists\, p_0 : p(x) \geqslant p_0 \,\forall\, x \in supp(\mathbb{P}_X), \\ \left(\frac{M^3 \log^2 n}{n}\right)^{\frac{\alpha(1+\beta)}{\alpha\beta/p + 2\alpha + d}}, & \text{otherwise}. \end{cases} \tag{1.12}$$

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

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

then, on with probability at least $(1-\delta)$ over $S_n \sim \mathcal{D}^{\otimes n}$, it holds that

$$\mathcal{E}(\widehat{f}) \leqslant \mathbb{P}(\widehat{f}(X) \neq f^*(X)) \lesssim \delta + \left(\frac{M^3 \log(12KM/\delta)}{n}\right)^{\frac{\alpha\beta}{\alpha\beta/p+(2\alpha+d)}}. \tag{1.13}$$

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

There are some comments we have. First, the rates (1.12) are optimal up to a logarithmic factor (see [35, Theorem 3.2] for the case of bounded away from zero density, [35, Theorem 4.1] for the case of bounded support (i.e. $p = 1$ in (A6)), and [12, Theorem 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 (1.13) transforms into

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

recovering the result of Theorem 7 in [10].

1.6.2. Comparison with the nearest neighbor rule

Theorem 2. *Assume (A1), (A3), and (A5). Fix any m from $\{1, \dots, M\}$, and a test point $x \in \mathcal{X}$. Then, for the weighted nearest neighbor estimate $\widetilde{\eta}^w_m(x)$ defined by (1.4) and (1.5), with probability at least $(1-\delta)$ over $S_n \sim \mathcal{D}^{\otimes n}$, it holds that*

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

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 bound in Theorem 2 improves the result for the nearest neighbor regression obtained in [11] since it controls large deviations of $|\eta_m(x) - \widetilde{\eta}^w_m(x)|$ rather than its mean value. For the case of a bounded away from zero density, Theorem 2 and the union bound immediately yield

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

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

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

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 of convergence $O(n^{-(1+\beta)/(2\alpha+d)})$, 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 [12, Theorems 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 $O(n^{-\frac{\alpha(1+\beta)}{\alpha(1+\beta)/p+2\alpha+d}})$, while Theorem 1 guarantees that the MSSA classifier has a minimax rate of convergence up to a logarithmic factor. It was also shown in [12, Theorems 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 $O\left(\left[(\log n)/n\right]^{\frac{\alpha(1+\beta)}{\alpha\beta/p+2\alpha+d}} \right)$ 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. On the other hand, there is a simple way to tune the 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.

Chapter 2
Manifold learning

2.1. Literature review

We consider a problem of manifold learning, that is, to recover a smooth low dimensional manifold from a cloud of points in a high dimensional space. This problem is of great theoretical and practical interest. For instance, if one deals with a problem of supervised or semi-supervised regression, the feature vectors, though lying in a very high-dimensional space, may occupy only a low-dimensional subset. In this case, one can hope to obtain a rate of prediction which depends on the intrinsic dimension of the data rather than on the ambient one and escape the curse of dimensionality. At the beginning of the century, the popularity of manifold learning gave rise to several novel nonlinear dimension reduction procedures, such as Isomap [40], locally linear embedding [41, LLE] and its modification [42], Laplacian eigenmaps [43], and t-SNE [44]. More recent works include interpolation on manifolds via geometric multi-resolution analysis [15], local polynomial estimators [16] and numerical solution of PDE [45]. It is worth mentioning that all these works assume that the data points either lie exactly on the manifold or in its very small vicinity (which shrinks as the sample size n tends to infinity), so the noise ε is so negligible that it may be ignored and put into a remainder term in Taylor's expansion. However, in practice, this assumption can be too resrictive. and the observed data do not exactly lie on a manifold. One may think of this situation as there are unobserved "true" features that lie exactly on the manifold and the learner observes its corrupted versions. Such noise corruption leads to a dramatic decrease in the quality of manifold reconstruction for those algorithms which misspecify the model and assume that the data lies exactly on the manifold. Therefore, one has to do a preliminary step, which is called manifold denoising (see e.g. [14, 46, 47]), to first project the data onto the manifold. Such methods usually act locally, i.e. consider a set of small neighborhoods, determined by a smoothing parameter (e.g. a number of neighbors or a radius h), and construct local approximations based on these neighborhoods. The problem of this approach is that the size of the neighborhood must be large compared to the noise magnitude M, which may lead to a non-optimal choice of the smoothing parameter. The exclusion is the class of procedures, based on an optimization problem, such as mean-shift [48, 49] and its variants [46, 50, 51]. The mean-shift algorithm may be viewed as a generalized EM algorithm applied to the kernel density estimate (see [52]). However, since the mean shift algorithm and its variants approximate the true density of $Y_1, \dots, Y_n$ by the kernel density estimate, they may suffer from the curse of dimensionality and the rates of convergence we found in the literature depend on the ambient dimension rather than on the intrinsic one in the noisy case. To our best knowledge, only papers [17, 18] consider the case, when the noise magnitude does not tend to zero as n grows. However, the approach in [17, 18] assumes that the noise distribution is known and has a very special structure. For instance, considered in [17], the noise has a uniform distribution in the direction orthogonal to the manifold tangent space. Without belittling a significant impact of this paper, the assumption about the uniform distribution is unlikely to hold in practice. Thus, there are two well studied extremal situations in manifold learning. The first one corresponds to the case of totally unknown noise distribution but extremely small noise magnitude, and the other one corresponds to the case of large noise, which distribution is completely known. This thesis aims at studying the problem of manifold recovery under weak and realistic assumptions on the noise.

2.2. Manifold learning and structural adaptation

Below we focus on a model with additive noise. Suppose we are given an i.i.d. sample $\mathbb{Y}_n = (Y_1, \dots, Y_n)$, where Y_i are independent copies of a random vector Y in $\mathbb{R}^D$, generated from the model

$$Y = X + \varepsilon. \tag{2.1}$$

Here X is a random element whose distribution is supported on a low-dimensional manifold $\mathcal{M}^* \subset \mathbb{R}^D$, $\dim(\mathcal{M}^*) = d < D$, and ε is a full dimensional noise. The goal of a statistician is to recover the corresponding unobserved variables $\mathbb{X}_n = \{X_1, \dots, X_n\}$, which lie on the manifold $\mathcal{M}^*$, and estimate $\mathcal{M}^*$ itself. Assumptions on the noise are crucial for the quality of estimation. One usually assumes that the noise is not too large, that is, $\|\varepsilon\| \leqslant M$ almost surely for some relatively small noise magnitude M. If the value M is smaller than the reach[1] of the manifold then the noise can be naturally decomposed in a component aligned with the manifold tangent space and another component describing the departure from the manifold. It is clear that the impact of these two components is different, and it is natural to consider an anisotropic noise. For this purpose, we introduce a free parameter b which controls the norm of the tangent component of the noise; see (A3) for the precise definition. The pair of parameters (M, b) characterizes the noise structure more precisely than just the noise magnitude M and allows us to understand the influence of the noise anisotropy on the rates of convergence.

As already mentioned, most of the existing manifold denoising procedures involve some nonparametric local smoothing methods with a corresponding bandwidth. The use of isotropic smoothing leads to the constraint that the noise magnitude is significantly smaller than the width of local neighborhoods; see e.g. [8, 14–16]. Similar problem arises even the case of effective dimension reduction in regression corresponding to the case of linear manifolds. The use of anisotropic smoothing helps to overcome this difficulty and to build efficient and asymptotically optimal estimation procedures; see e.g. [53] or [21]. The thesis extends the idea of *structural adaptation* proposed in [20, 21]. In our method, we construct cylindric neighborhoods, which are stretched in a normal direction to the manifold. However, our result is not a formal generalization of [20] and [21]. Those papers considered a regression setup, while our study focuses on a special unsupervised learning problem. This requires to develop essentially different technique and use different mathematical tools for theoretical study and substantially modify of the procedure. Also to mention that a general manifold learning is much more involved than just linear dimension reduction, and a straightforward extension from the linear case is not possible.

2.3. Contribution

Let us briefly describe our procedure and the main contributions of the present chapter. Many manifold denoising procedures (see, for instance, [8, 14, 19, 51]) act in an iterative manner and our procedure is not an exception. We start with some guesses $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ of the projectors onto the tangent spaces of $\mathcal{M}^*$ at the points $X_1, \dots, X_n$, respectively. These guesses may be very poor, in fact. Nevertheless, they give a bit of information, which can be used to construct initial estimates $\widehat{X}_1^{(0)}, \dots, \widehat{X}_n^{(0)}$. On the other hand, the estimates $\widehat{X}_1^{(0)}, \dots, \widehat{X}_n^{(0)}$ help to construct the estimates $\widehat{\mathbf{\Pi}}_1^{(1)}, \dots, \widehat{\mathbf{\Pi}}_n^{(1)}$ of the projectors onto the tangent spaces of $\mathcal{M}^*$ at the points $X_1, \dots, X_n$, respectively, which are better than $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$. One can repeat these two steps to iteratively refine the estimates of $X_1, \dots, X_n$ and of the manifold $\mathcal{M}^*$ itself. We call this approach a *structure-adaptive manifold estimation* (SAME). We show that SAME constructs such

[1] A reader is referred to Section 2.4 for the definition.

estimates $\widehat{X}_1, \dots, \widehat{X}_n$ of $X_1, \dots, X_n$ and a manifold estimate $\widehat{\mathcal{M}}$ of $\mathcal{M}^*$, such that

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb \vee Mh \vee h^2}{\varkappa} + \sqrt{\frac{D(h^2 \vee M^2) \log n}{nh^d}}, \qquad \text{(Theorem 3)}$$

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \left(\frac{M^2 b^2}{\varkappa^3} \vee \frac{h^2}{\varkappa} \right) + \sqrt{\frac{D(h^4/\varkappa^2 \vee M^2) \log n}{nh^d}}, \qquad \text{(Theorem 4)}$$

provided that $h \gtrsim \left((D \log n/n)^{1/d} \vee (DM^2 \varkappa^2 \log n/n)^{1/(d+4)} \right)$ and M and, possibly, b degrade to zero fast enough, and both inequalities hold with an overwhelming probability. Here h is the width of a cylindrical neighborhood, which we are able to control, $\varkappa$ is a lower bound for the reach of $\mathcal{M}^*$ (see Section 2.4 for the definition of reach). Moreover, our algorithm estimates projectors $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ onto tangent spaces at $X_1, \dots, X_n$. It produces estimates $\widehat{\mathbf{\Pi}}_1, \dots, \widehat{\mathbf{\Pi}}_n$, such that

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i - \mathbf{\Pi}(X_i)\| \lesssim \frac{h}{\varkappa} + h^{-1} \sqrt{\frac{D(h^4/\varkappa^2 \vee M^2) \log n}{nh^d}} \qquad \text{(Theorem 3)}$$

with high probability. Here, for any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ denotes its spectral norm. The notation $f(n) \lesssim g(n)$ means that there exists a constant $c > 0$, which does not depend on n, such that $f(n) \leqslant cg(n)$. $d_H(\cdot, \cdot)$ denotes the Hausdorff distance and it is defined as follows:

$$d_H(\mathcal{M}_1, \mathcal{M}_2) = \inf \left\{ \varepsilon > 0 : \mathcal{M}_1 \subseteq \mathcal{M}_2 \oplus \mathcal{B}(0, \varepsilon), \, \mathcal{M}_2 \subseteq \mathcal{M}_1 \oplus \mathcal{B}(0, \varepsilon) \right\},$$

where $\oplus$ stands for the Minkowski sum and $\mathcal{B}(0, r)$ is a Euclidean ball in $\mathbb{R}^D$ of radius r.

The optimal choice of h yields

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb}{\varkappa} \vee \frac{1}{\varkappa} \left(\frac{D\varkappa^2 \log n}{n} \right)^{\frac{2}{d+2}} \vee \frac{M}{\varkappa} \left(\frac{DM^2 \varkappa^2 \log n}{n} \right)^{\frac{1}{d+4}},$$

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \frac{M^2 b^2}{\varkappa^3} \vee \frac{1}{\varkappa} \left(\frac{D \log n}{n} \right)^{\frac{2}{d}} \vee \frac{1}{\varkappa} \left(\frac{DM^2 \varkappa^2 \log n}{n} \right)^{\frac{2}{d+4}}$$

and

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i - \mathbf{\Pi}(X_i)\| \lesssim \frac{1}{\varkappa} \left(\frac{D \log n}{n} \right)^{\frac{1}{d}} \vee \frac{1}{\varkappa} \left(\frac{DM^2 \varkappa^2 \log n}{n} \right)^{\frac{1}{d+4}}.$$

Note that the optimal choice of h is much smaller than a possible value $n^{-2/(3d+8)}$ of the noise magnitude M. Besides, we prove a lower bound

$$\inf_{\widehat{\mathcal{M}}} \sup_{\mathcal{M}^*} \mathbb{E} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \frac{M^2 b^2}{\varkappa^3} \vee \varkappa^{-1} \left(\frac{M^2 \varkappa^2 \log n}{n} \right)^{\frac{2}{d+4}} \qquad \text{(Theorem 5)}$$

which has never appeared in the manifold learning literature. Here $\widehat{\mathcal{M}}$ is an arbitrary estimate of $\mathcal{M}^*$ and $\mathcal{M}^*$ fulfills some regularity conditions, which are precisely specified in Theorem 5. Theorem 5, together with Theorem 1 from [54], where the authors managed to obtain the lower bound $\inf_{\widehat{\mathcal{M}}} \sup_{\mathcal{M}^*} \mathbb{E} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim (\log n/n)^{2/d}$, claims optimality of our method.

2.4. Model assumptions

Let us remind that we consider the model (2.1), where X belongs to the manifold $\mathcal{M}^*$ and the distribution of the error vector ε will be described a bit later in this section. First, we require regularity of the underlying manifold $\mathcal{M}^*$. We assume that it belongs to a class $\mathfrak{M}^d_\varkappa$ of twice differentiable, compact, connected manifolds without a boundary, contained in a ball $\mathcal{B}(0,R)$, with a reach, bounded below by $\varkappa$, and dimension d:

$$\begin{aligned}\mathcal{M}^* \in \mathbb{M}^d_{\varkappa} = \big\{\mathcal{M} \subset \mathbb{R}^D : \mathcal{M} \text{ is a compact, connected manifold}\\ \text{without a boundary}, \mathcal{M} \in \mathcal{C}^2, \mathcal{M} \subseteq \mathcal{B}(0,R),\\ \text{reach}\,(\mathcal{M}) \geqslant \varkappa, \dim(\mathcal{M}) = d < D\big\}\,.\end{aligned} \tag{A1}$$

The reach of a manifold $\mathcal{M}$ is defined as a supremum of such r that any point in $\mathcal{M} \oplus \mathcal{B}(0,r)$ has a unique (Euclidean) projection onto $\mathcal{M}$. Here $\oplus$ stands for the Minkowski sum and $\mathcal{B}(0,r)$ is a Euclidean ball in $\mathbb{R}^D$ of radius r. The requirement that the reach is bounded away from zero prevents $\mathcal{M}^*$ from having a large curvature. In fact, if the reach of $\mathcal{M}^*$ is at least $\varkappa$, then the curvature of any geodesic on $\mathcal{M}^*$ is bounded by $1/\varkappa$ (see [17, Lemma 3]).

Second, the density $p(x)$ of X (with respect to the d-dimensional Hausdorff measure on $\mathcal{M}^*$) meets the following condition:

$$\begin{aligned}&\exists\, p_1 \geqslant p_0 > 0 : \forall x \in \mathcal{M}^* \quad p_0 \leqslant p(x) \leqslant p_1,\\ &\exists L \geqslant 0 : \forall\, x, x' \in \mathcal{M}^* \quad |p(x) - p(x')| \leqslant \frac{L\|x - x'\|}{\varkappa}.\end{aligned} \tag{A2}$$

Besides the aforementioned conditions on $\mathcal{M}^*$ and X, we require some properties of the noise ε. We suppose that, given $X \in \mathcal{M}^*$, the conditional distribution $(\varepsilon \,|\, X)$ fulfils the following assumption: there exist $0 \leqslant M < \varkappa$ and $0 \leqslant b \leqslant \varkappa$, such that

$$\begin{aligned}&\mathbb{E}(\varepsilon \,|\, X) = 0,\ \|\varepsilon\| \leqslant M < \varkappa,\\ &\|\mathbf{\Pi}(X)\varepsilon\| \leqslant \frac{Mb}{\varkappa} \qquad \mathbb{P}(\cdot \,|\, X)\text{-almost surely,}\end{aligned} \tag{A3}$$

where $\mathbf{\Pi}(X)$ is the projector onto the tangent space $\mathcal{T}_X\mathcal{M}^*$ of $\mathcal{M}^*$ at X. The model with manifold $\mathcal{M}^* \in \mathbb{M}^d_{\varkappa}$ and the bounded noise has been extensively studied in literature (see [15–17, 19, 55]). In [56], the authors consider the Gaussian noise, which is unbounded, but they restrict themselves on the event $\max\limits_{1 \leqslant i \leqslant n} \|\varepsilon_i\| \leqslant \varkappa$, which is essentially similar to the case of bounded noise. In our work, we introduce an additional parameter $b \in [0, \varkappa]$, which characterises maximal deviation in tangent direction.

The pair of parameters (M, b) determines the noise structure more precisely than just the noise magnitude M. If $b = 0$, we deal with perpendicular noise, which was studied in [16, 17]. The case $b = \varkappa$ corresponds to the bounded noise, which is not constrained to be orthogonal. Such model was considered, for instance, in [19]. In our work, we provide upper bounds on accuracy of manifold estimation for all pairs (M, b) satisfying the following conditions:

$$\begin{cases} M \leqslant An^{-\frac{2}{3d+8}},\\ M^3 b^2 \leqslant \alpha\varkappa \left[\left(\frac{D\log n}{n}\right)^{\frac{4}{d}} \vee \left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{4}{d+4}}\right], \end{cases} \tag{A4}$$

where A and α are some positive constants. Among all the pairs (M, b), satisfying (A4), we can highlight two cases. The first one is the case of maximal admissible magnitude:

$$\begin{aligned}&M = M(n) \leqslant An^{-\frac{2}{3d+8}},\\ &b = b(n) \leqslant \frac{\sqrt{\alpha\varkappa}}{A^{3/2}} \left[\left(\frac{D\log n}{n}\right)^{\frac{1}{d}} \vee \left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{1}{d+4}}\right].\end{aligned} \tag{A4.1}$$

The second one is the case of maximal admissible angle:

$$b = \varkappa, \quad M = M(n) \leqslant \left(\frac{D^4 \alpha^{d+4}}{\varkappa^{d-4}} \right)^{\frac{1}{3d+4}} n^{-\frac{4}{3d+4}}. \tag{A4.2}$$

If (A4.1) holds, we deal with *almost* perpendicular noise. Note that in this case the condition (A3) ensures that X is very close to the projection $\pi_{\mathcal{M}^*}(Y)$ of Y onto $\mathcal{M}^*$. Here and further in this thesis, for a closed set $\mathcal{M}$ and a point x, $\pi_{\mathcal{M}}(x)$ stands for a Euclidean projection of x onto $\mathcal{M}$. Thus, estimating $X_1, \dots, X_n$, we also estimate the projections of $Y_1, \dots, Y_n$ onto $\mathcal{M}^*$. Also, we admit that the noise magnitude M may decrease as slow as $n^{-2/(3d+8)}$. We discuss this condition in details in Section 2.6 after Theorem 3 and compare it with other papers to convince the reader that the assumption $M \leqslant A n^{-2/(3d+8)}$ is mild. In fact, to the best of our knowledge, only in [17, 18] the authors impose weaker assumptions on the noise magnitude. At the first glance, the condition (A4.1) looks very similar to the case of orthogonal noise $b = 0$. However, our theoretical study reveals a surprising effect: the existing lower bounds for manifold estimation in the case of perpendicular noise are different from the rates we prove for the case of almost perpendicular noise satisfying (A4.1). We provide the detailed discussion in Section 2.6 below.

Finally, if (A4.2) holds, the noise is not constrained to be orthogonal. However, in this case, we must impose more restrictive condition on the noise magnitude than in (A4.1). Nevertheless, under the condition (A4.2), we show that the result of [19], Theorem 2.7, where the authors also consider bounded noise, can be improved if one additionally assumes that the log-density $\log p(x)$ is Lipschitz. A more detailed discussion is provided in Section 2.6.

2.5. An adaptive weights method for manifold denoising

In this section we propose a novel manifold estimation procedure based on a nonparametric smoothing technique and structural adaptation idea. One of the most popular methods in nonparametric estimation is weighted averaging:

$$\widehat{X}_i^{(loc)} = \frac{\sum\limits_{j=1}^{n} w_{ij}^{(loc)} Y_j}{\sum\limits_{j=1}^{n} w_{ij}^{(loc)}}, \quad 1 \leqslant i \leqslant n, \tag{2.2}$$

and $w_{ij}^{(loc)}$ are the localizing weights defined by

$$w_{ij}^{(loc)} = \mathcal{K}\left(\frac{\|Y_i - Y_j\|^2}{h^2} \right), \quad 1 \leqslant i, j \leqslant n,$$

where $\mathcal{K}(\cdot)$ is a smoothing kernel and the bandwidth $h = h(n)$ is a tuning parameter. In this paper, we consider the kernel $\mathcal{K}(t) = e^{-t}$.

Remark 1. *Instead of $\mathcal{K}(t) = e^{-t}$, one can take any two times differentiable, monotonously decreasing on $\mathbb{R}_+$ function such that it and its first and second derivatives have either exponential decay or finite support. We use $\mathcal{K}(t) = e^{-t}$ to avoid further complications of the proofs.*

The estimate (2.2) has an obvious limitation. Consider a pair on indices (i, j) such that $\|X_i - X_j\| < h$ and $h = h(n)$ is of order $(\log n/n)^{1/d}$, which is known to be the optimal choice in the presence of small noise (see [19, Proposition 5.1] and [16, Theorem 6]). If the noise magnitude M is much larger than $(\log n/n)^{1/d}$ (which is the case we also consider), then $M > h$ and the weights $w_{ij}^{(loc)}$ carry wrong information about the neighborhood of X_i, i.e. $w_{ij}^{(loc)}$ can be very small even if the distance $\|X_i - X_j\|$ is smaller than h. This leads to a large variance of the estimate (2.2) when h is of order $(\log n/n)^{1/d}$, and one has to increase the bandwidth h, inevitably making the bias of the estimate larger.

The argument in the previous paragraph leads to the conclusion that the weights $w_{ij}^{(loc)}$ must be adjusted. Let us fix any i from 1 to n. "Ideal" localizing weights w_{ij} are such that they take into account only those indices j, for which the norm $\|X_i - X_j\|$ does not exceed the bandwidth h too much. Of course, we do not have access to compute the norms $\|X_i - X_j\|$ for all pairs but assume for a second that the projector $\mathbf{\Pi}(X_i)$ onto the tangent space $\mathcal{T}_{X_i}\mathcal{M}^*$ was known. Then, instead of the weights $w_{ij}^{(loc)}$, one would rather use the ones of the form

$$w_{ij}(\mathbf{\Pi}(X_i)) = \mathcal{K}\left(\frac{\|\mathbf{\Pi}(X_i)(Y_i - Y_j)\|^2}{h^2}\right), \quad 1 \leqslant j \leqslant n,$$

to remove a large orthogonal component of the noise. The norm $\|\mathbf{\Pi}(X_i)(Y_i - Y_j)\|$ turns out to be closer to $\|X_i - X_j\|$ than $\|Y_i - Y_j\|$, especially if the ambient dimension is large. Thus, instead of the ball $\{Y \colon \|Y - Y_i\| \leqslant h\}$ around Y_i, we consider a cylinder $\{Y \colon \|\mathbf{\Pi}_i(Y_i - Y)\| \leqslant h\}$, where $\mathbf{\Pi}_i$ is a projector, which is assumed to be close to $\mathbf{\Pi}(X_i)$. One just has to ensure that the cylinder does not intersect $\mathcal{M}^*$ several times. For this purpose, we introduce the weights

$$w_{ij}(\mathbf{\Pi}_i) = \mathcal{K}\left(\frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2}{h^2}\right)\mathbb{1}\left(\|Y_i - Y_j\| \leqslant \tau\right), \quad 1 \leqslant j \leqslant n, \tag{2.3}$$

with a constant $\tau < \varkappa$.

The adjusted weights (2.3) require a "good" guess $\mathbf{\Pi}_i$ of the projector $\mathbf{\Pi}(X_i)$. The question is how to find this guess. We use the following strategy. We start with poor estimates $\widehat{\mathbf{\Pi}}_1{}^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ of $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ and take a large bandwidth h_0. Then we compute the weighted average estimates $\widehat{X}_1^{(1)}, \dots, \widehat{X}_n^{(1)}$ with the adjusted weights (2.3) and the bandwidth h_0. These estimates can be then used to construct estimates $\widehat{\mathbf{\Pi}}_1^{(1)}, \dots, \widehat{\mathbf{\Pi}}_n^{(1)}$ of $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$, which are better than $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$. After that, we repeat the described steps with a bandwidth $h_1 < h_0$. This leads us to an iterative procedure, which is given by Algorithm 2.

The computational complexity of Algorithm 2 is $O(n^2D^2K + nD^3K)$. This includes $O(n^2D^2)$ operations to update the weights $w_{ij}^{(k)}$, $1 \leqslant i, j \leqslant n$, and the estimates $\widehat{X}_i^{(k)}$ and $\widehat{\mathbf{\Sigma}}_i^{(k)}$, $1 \leqslant i \leqslant n$, on each iteration and $O(nD^3)$ operations to update the projectors $\widehat{\mathbf{\Pi}}_i^{(k)}$, $1 \leqslant i \leqslant n$, on each iteration.

2.6. Theoretical properties of SAME

This section states the main results. Here and everywhere in this thesis, for any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ denotes its spectral norm. The notation $f(n) \asymp g(n)$ means $f(n) \lesssim g(n) \lesssim f(n)$.

Theorem 3. *Assume (A1), (A2), (A3), and (A4) . Let the initial guesses $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ of $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ be such that on an event with probability at least $1 - n^{-1}$ it holds*

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(0)} - \mathbf{\Pi}(X_i)\| \leqslant \frac{\Delta h_0}{\varkappa}$$

with a constant Δ, such that $\Delta h_0 \leqslant \varkappa/4$, and $h_0 = C_0/\log n$, where $C_0 > 0$ is an absolute constant. Choose $\tau = 2C_0/\sqrt{\log n}$ and set any $a \in (1, 2]$. If n is larger than a constant N_Δ, depending on Δ, and $h_K \gtrsim \left((D \log n/n)^{1/d} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$ (with a sufficiently large hidden constant, which is greater than 1) then there exists a choice of γ, such that after K iterations Algorithm 2

Algorithm 2 Structure-adaptive manifold estimator (SAME)

```
The sample of noisy observations $\mathbb{Y}_n = (Y_1, \dots, Y_n)$, the initial guesses $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ of $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$, the number of iterations $K + 1$, an initial bandwidth $h_0$, the threshold $\tau$ and constants $a > 1$ and $\gamma > 0$ are given.
for $k$ from 0 to $K$ do
    Compute the weights $w_{ij}^{(k)}$ according to the formula
        $w_{ij}^{(k)} = \mathcal{K}\left(\frac{\|\widehat{\mathbf{\Pi}}_i^{(k)}(Y_i - Y_j)\|^2}{h_k^2}\right) \mathbb{1}\left(\|Y_i - Y_j\| \leqslant \tau\right), \quad 1 \leqslant i, j \leqslant n.$
    Compute the estimates
        $\widehat{X}_i^{(k)} = \sum_{j=1}^{n} w_{ij}^{(k)} Y_j \Big/ \left(\sum_{j=1}^{n} w_{ij}^{(k)}\right), \quad 1 \leqslant i \leqslant n.$    (2.4)
    If $k < K$, for each $i$ from 1 to n, define a set $\mathcal{J}_i^{(k)} = \{j : \|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\| \leqslant \gamma h_k\}$ and compute the matrices
        $\widehat{\mathbf{\Sigma}}_i^{(k)} = \sum_{j \in \mathcal{J}_i^{(k)}} (\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})(\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})^T, \quad 1 \leqslant i \leqslant n.$
    If $k < K$, for each $i$ from 1 to n, define $\widehat{\mathbf{\Pi}}_i^{(k+1)}$ as a projector onto a linear span of eigenvectors of $\widehat{\mathbf{\Sigma}}_i^{(k)}$, corresponding to the largest $d$ eigenvalues.
    If $k < K$, set $h_{k+1} = a^{-1} h_k$.
return the estimates $\widehat{X}_1 = \widehat{X}_1^{(K)}, \dots, \widehat{X}_n = \widehat{X}_n^{(K)}$.
```

produces estimates $\widehat{X}_1, \dots, \widehat{X}_n$*, such that, with probability at least* $1 - (5K + 4)/n$*, it holds*

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb \vee Mh_K \vee (1 + \Phi_{M,b,h_K,\varkappa})h_K^2}{\varkappa} + \sqrt{\frac{D(h_K^2 \vee M^2)\log n}{nh_K^d}},$$

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(K)} - \mathbf{\Pi}(X_i)\| \lesssim \Psi_{M,b,h_K,\varkappa}\left(\frac{h_K}{\varkappa} + h_K^{-1}\sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2)\log n}{nh_K^d}}\right),$$

where

$$\begin{aligned}
\Phi_{M,b,h_K,\varkappa} &= \frac{M^3(1 + b/h_K)^2}{h_K^2 \varkappa} + \frac{M^2(1 + b/h_K + \sqrt{\log h_K^{-1}})}{\varkappa h_K} + \frac{Mh_K^2}{\varkappa^3} \\
&\lesssim \alpha + o(1), \quad n \to \infty, \\
\Psi_{M,b,h_K,\varkappa} &= \left(1 + \frac{M(1 + b/h_K) \vee (1 + \Phi_{M,b,h_K,\varkappa})h_K}{\varkappa}\right)^{d+1}(1 + \Phi_{M,b,h_K,\varkappa}) \qquad (2.5) \\
&\leqslant (1 + \alpha)\left(4^{d+1} + (2\sqrt{\alpha})^{d+1}\right).
\end{aligned}$$

In particular, if one chooses the parameter a *and the number of iterations* K *in such a way that* $h_K \asymp \left((D\varkappa^2 \log n/n)^{1/(d+2)} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$ *then*

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i - X_i\| \lesssim \frac{Mb}{\varkappa} + \frac{1}{\varkappa}\left(\frac{D\varkappa^2 \log n}{n}\right)^{\frac{2}{d+2}} \vee \frac{M}{\varkappa}\left(\frac{DM^2\varkappa^2 \log n}{n}\right)^{\frac{1}{d+4}}.$$

If $h_K \asymp \left((D\log n/n)^{1/d} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$ *then*

$$\max_{1\leqslant i\leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(K)} - \mathbf{\Pi}(X_i)\| \lesssim \frac{1}{\varkappa}\left(\frac{D\log n}{n}\right)^{\frac{1}{d}} \vee \frac{1}{\varkappa}\left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{1}{d+4}}.$$

Note that one has to take the number of iterations K of order $\log n$ since the sequence of bandwidths $h_1, \dots, h_K$ decreases exponentially.

In Theorem 3, we assume that $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ may depend on $Y_1, \dots, Y_n$. The natural question is how to construct the initial guesses $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$ of the projectors $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$. We propose a strategy for initialization of our procedure. One can use [19, Proposition 5.1] to get the estimates $\widehat{\mathbf{\Pi}}_1^{(0)}, \dots, \widehat{\mathbf{\Pi}}_n^{(0)}$. For each i from 1 to n introduce

$$\widehat{\mathbf{\Sigma}}_i^{(0)} = \frac{1}{n-1}\sum_{j\neq i}(Y_j - \overline{Y}_i)(Y_j - \overline{Y}_i)^T \mathbb{1}(Y_j \in \mathcal{B}(Y_i, h_0)),$$

where $\overline{Y}_i = \frac{1}{N_i}\sum_{j\neq i} Y_j \mathbb{1}(Y_j \in \mathcal{B}(Y_i, h_0))$, $N_i = |\{j : Y_j \in \mathcal{B}(Y_i, h)\}|$. Let $\widehat{\mathbf{\Pi}}_i^{(0)}$ be the projector onto the linear span of the d largest eigenvectors of $\widehat{\mathbf{\Sigma}}_i^{(0)}$. Following the lines of the proof of [19, Proposition 5.1], one can show that, with probability larger than $1 - n^{-1}$, it holds

$$\max_{1\leqslant i\leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(0)} - \mathbf{\Pi}(X_i)\| \lesssim \frac{h_0}{\varkappa} + \frac{M}{h_0},$$

provided that $h_0 \gtrsim (\log n/n)^{1/d}$, $h_0 = h_0(n) = o(1)$ as $n \to \infty$, and n is sufficiently large.

Condition (A4) and the choice of h_K in Theorem 3 yield that $M = M(n)$ can decrease almost as slow as $h_K^{2/3} = h_K^{2/3}(n)$. Thus, we admit the situation when the noise magnitude M is much larger than the smoothing parameter h_K. For instance, in [16], the authors use local polynomial estimates and require $M = O(h^2)$ and $h = h(n) \asymp n^{-1/d}$. In [19], the authors assume $M \lesssim \lambda(\log n/n)^{1/d}$. In [56], the authors deal with Gaussian noise $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_D)$ and get the accuracy of manifold estimation $O(\sigma\sqrt{D})$ using $O(\sigma^{-d})$ samples. This means that $\sigma = O(n^{-1/d})$, which yields that

$$\max_{1\leqslant i\leqslant n} \|\varepsilon_i\| \lesssim n^{-1/d}\sqrt{D\log n}$$

with overwhelming probability. A similar situation is observed in [51], where the authors also consider the Gaussian noise $\mathbb{N}(0, \sigma^2 \boldsymbol{I}_D)$ and, using the kernel density estimate with bandwidth h, obtain the upper bound

$$O\left(\sigma^2 \log\sigma^{-1} + h^2 + \sqrt{\frac{\log n}{nh^D}}\right)$$

on the Hausdorff distance between $\mathcal{M}^*$ and their estimate. In order to balance the first and the second terms, one must take $\sigma = O(h/\sqrt{\log h^{-1}})$, which means that

$$\max_{1\leqslant i\leqslant n} \|\varepsilon_i\| \lesssim h\sqrt{\frac{D\log n}{\log h^{-1}}},$$

while we allow $\max_{1\leqslant i\leqslant n} \|\varepsilon_i\|$ be as large as $h_K^{2/3}$. Finally, in [14] the authors require $M = O(h)$. So, we see that the condition (A4) is quite mild.

Theorem 3 claims that, despite the relatively large noise, our procedure constructs consistent estimates of the projections of the sample points onto the manifold $\mathcal{M}^*$. The accuracy of the projection estimation is a bit worse than the accuracy of manifold estimation, which we provide in Theorem 4 below. The reason for that is the fact that the estimate $\widehat{X}_i$ is significantly shifted with respect to X_i in a tangent direction, while the orthogonal component of $(\widehat{X}_i - X_i)$ is small.

A similar phenomenon was already known in the problem of efficient dimension reduction. For instance, in [20, 21] the authors managed to obtain the rate $n^{-2/3}$ for the bias of the component, which is orthogonal to the efficient dimension reduction space, while the rate of the bias in the index estimation was only $n^{-1/2}$. Moreover, the term Mh_K in Theorem 3 appears because of the correlation between the weights $w_{ij}^{(k)}$ and the sample points Y_j.

We proceed with upper bounds on the estimation of the manifold $\mathcal{M}^*$.

Theorem 4. *Assume conditions of Theorem 3. Consider the piecewise linear manifold estimate*

$$\widehat{\mathcal{M}} = \left\{ \widehat{X}_i + h_K \widehat{\mathbf{\Pi}}_i^{(K)} u : 1 \leqslant i \leqslant n,\ u \in \mathcal{B}(0,1) \subset \mathbb{R}^D \right\},$$

where $\widehat{\mathbf{\Pi}}_i^{(K)}$ is a projector onto d-dimensional space obtained on the K-th iteration of Algorithm 2. Then, as long as $h_K \gtrsim \left((D \log n/n)^{1/d} \vee (DM^2 \varkappa^2 \log n/n)^{1/(d+4)}\right)$ (with a sufficiently large hidden constant, which is greater than 1), on an event with probability at least $1-(5K+5)/n$, it holds

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \left(\frac{(1 + \Phi_{M,b,h_K,\varkappa} + \Psi_{M,b,h_K,\varkappa}) h_K^2}{\varkappa} \vee \frac{M^2 b^2}{\varkappa^3} \right) + \sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2) \log h_K^{-1}}{n h_K^d}},$$

where $\Phi_{M,b,h_K,\varkappa}$ and $\Psi_{M,b,h_K,\varkappa}$ are defined in (2.5). In particular, if a and K are chosen such that $h_K \asymp \left((D \log n/n)^{1/d} \vee (DM^2\varkappa^2/n \log n)^{1/(d+4)}\right)$, then

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \lesssim \frac{M^2 b^2}{\varkappa^3} \vee \varkappa^{-1} \left(\frac{D \log n}{n} \right)^{\frac{2}{d}} \vee \varkappa^{-1} \left(\frac{DM^2 \varkappa^2 \log n}{n} \right)^{\frac{2}{d+4}}.$$

Let us elaborate on the result of Theorem 4. First, let us discuss the case of bounded non-orthogonal noise, that is, the situation when (A4.2) holds. The model with bounded noise was considered in [19], where the authors assumed that $\mathcal{M}^*$ satisfies (A1) and the density of X fulfils

$$0 < p_0 \leqslant p(x) \leqslant p_1, \quad \forall x \in \mathcal{M}^*,$$

for some constants p_0, p_1. Note that this is a slightly more general setup, since we additionally assume that the log-density is Lipschitz. Under these assumptions, [19] proved (Theorem 2.7) the following upper bound on the Hausdorff distance using the tangential Delaunay complex (TDC):

$$d_H(\widehat{\mathcal{M}}_{TDC}, \mathcal{M}^*) \lesssim \left(\frac{\log n}{n} \right)^{2/d} + M^2 \left(\frac{\log n}{n} \right)^{-2/d},$$

provided that $M \lesssim (\log n/n)^{1/d}$. To the best of our knowledge, the situation, when (A1), (A2), (A3), and (A4.2) hold, was not studied in the manifold learning literature. One can observe that both TDC and SAME achieve the rate $O\left(\log n/n\right)^{2/d}$ in the case of extremely small noise $M \lesssim (\log n/n)^{2/d}$. However, if $(\log n/n)^{2/d} \lesssim M \lesssim n^{-4/(3d+4)}$ then the rate of convergence of SAME in the case of the density $p(x)$ satisfying (A2) improves over the known rates of TDC in the case of bounded away from 0 and ∞ density $p(x)$.

Now, let us discuss the case of almost orthogonal noise, i.e. when (A4.1) holds. This model is completely new in the manifold learning literature. The most similar one considered in the prior work is the model with perpendicular noise studied in [16, 17], so we find it useful to compare this more restrictive model with our upper bounds for the case of almost orthogonal noise. In [17], the authors obtain the rates $O(\log n/n)^{2/(d+2)}$ assuming that, given X, the noise ε has a uniform distribution on $\mathcal{B}(X, M) \cap (\mathcal{T}_X \mathcal{M}^*)^\perp$. In their work, the authors do not assume that M tends to zero as n tends to infinity, however, they put a far more restrictive assumption on the noise distribution than we do. In [16, Theorem 6], the authors use local polynomial estimate $\widehat{\mathcal{M}}_{LP}$ to prove the upper bound

$$d_H(\widehat{\mathcal{M}}_{LP}, \mathcal{M}^*) \lesssim \left(\frac{\log n}{n}\right)^{k/d} \vee M$$

for the case when $\mathcal{M}^*$ is a $\mathcal{C}^k$-manifold with dimension d and reach at least $\varkappa$ without a boundary. If $\mathcal{M}^*$ is a $\mathcal{C}^2$-manifold, this rate is minimax optimal for the case of extremely small noise $M \lesssim (\log n/n)^{2/d}$ but it can be improved when the noise magnitude exceeds $(\log n/n)^{2/d}$.

The result of Theorem 4 cannot be improved for the case of general additive noise, which fulfils the assumption (A3) with $b \gtrsim \left((\log n/n)^{1/d} \vee \ (M^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$. We justify this discussion by the following theorem.

Theorem 5. *Suppose that the sample $\mathbb{Y}_n = \{Y_1, \dots, Y_n\}$ is generated according to the model (2.1), where $\mathcal{M}^* \in \mathfrak{M}^d_\varkappa$, the density $p(x)$ of X fulfils (A2) (with sufficiently large $\mathfrak{p}_1, L$ and sufficiently small $\mathfrak{p}_0$) and the noise ε satisfies (A3). Then, for any estimate $\widehat{\mathcal{M}}$, it holds that*

$$\sup_{\mathcal{M}^* \in \mathfrak{M}^d_\varkappa} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \frac{M^2 b^2}{\varkappa^3}. \tag{2.6}$$

Moreover, if, in addition, n is sufficiently large, $M\varkappa \gtrsim (\log n/n)^{2/d}$, and the parameter b in (A3) is such that

$$b \gtrsim \left((\log n/n)^{1/d} \vee (M^2\varkappa^2 \log n/n)^{1/(d+4)}\right),$$

with a large enough hidden constant, then, for any estimate $\widehat{\mathcal{M}}$, it holds that

$$\sup_{\mathcal{M}^* \in \mathfrak{M}^d_\varkappa} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \varkappa^{-1} \left(\frac{M^2\varkappa^2 \log n}{n}\right)^{\frac{2}{d+4}}. \tag{2.7}$$

Theorem 5 studies the case $M \gtrsim (\log n/n)^{2/d}$. In [54], the authors proved the minimax lower bound

$$\inf_{\widehat{\mathcal{M}}} \sup_{\mathcal{M}^* \in \mathfrak{M}^d_\varkappa} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \left(\frac{\log n}{n}\right)^{2/d}$$

for the noiseless case, which is also tight for $M \lesssim (\log n/n)^{2/d}$. Theorem 5, together with [54, Theorem 1] yields that SAME is minimax optimal in the model with almost orthogonal noise. The lower bounds (2.6) and (2.7) are completely new and are different from the currently known results on manifold estimation from [17] and [16], where the authors studied a perpendicular noise fulfilling (A3) with $b = 0$.

Conclusion

1. We proposed an adaptive algorithm for 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 and adapts to the smoothness of the underlying target function.

2. We proved an upper bound on the excess risk of the classifier, returned by the proposed algorithm, under mild assumptions. This is the first theoretical result matching the minimax lower bound up to a logarithmic factor under these assumptions.

3. We developed a new structure-adaptive manifold estimation procedure for manifold denoising. The procedure turns out to be more robust to orthogonal noise, than the existing methods.

4. We carried out theoretical analysis of the proposed procedure. We proved new upper and lower bounds on the accuracy of manifold estimation. The bounds coincide up to a mutiplicative constant, claiming optimality of the proposed algorithm in the minimax sense.

References

1. C. H.Q. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–58, 2001.
2. H. Ahn and K.-J. Kim. Corporate credit rating using multiclass classification models with order information. *World Academy of Science, Engineering and Technology*, 60:95–100, 12 2011.
3. D. Belomestny and V. Spokoiny. Spatial aggregation of local likelihood estimates with applications to classification. *The Annals of Statistics*, 35(5):2287–2311, 2007.
4. A. Ganapathiraju, J.E. Hamaker, and J. Picone. Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing*, 52(8):2348–2355, 2004.
5. J. Kittler, R. Ghaderi, T. Windeatt, and J. Matas. Face verification via error correcting output codes. *Image and Vision Computing*, 21:1163–1169, 12 2003.
6. D. Li and D. B. Dunson. Classification via local manifold approximation. *Biometrika*, 107(4):1013–1020, 2020.
7. G. Peyré. Manifold models for signals and images. *Computer Vision and Image Understanding*, 113(2):249–260, 2009.
8. S. Osher, Z. Shi, and W. Zhu. Low dimensional manifold model for image processing. *SIAM Journal on Imaging Sciences*, 10(4):1669–1690, 2017.
9. R. J. Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.
10. K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3437–3445. MIT Press, 2014.
11. M. Döring, L. Györfi, and H. Walk. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18(227):1–16, 2018.
12. S. Gadat, T. Klein, and C. Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3):982–1009, 2016.
13. T. I. Cannings, T. B. Berrett, and R. J. Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *The Annals of Statistics*, 48(3):1789–1814, 2020.
14. M. Hein and M. Maier. Manifold denoising. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.
15. M. Maggioni, S. Minsker, and N. Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *Journal of Machine Learning Research*, 17(2):1–51, 2016.
16. E. Aamari and C. Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.
17. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13(43):1263–1291, 2012.
18. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012.
19. E. Aamari and C. Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete & Computational Geometry*, 59(4):923–971, 2018.
20. M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, 29(3):595–623, 2001.
21. M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001.
22. N. Puchkin and V. Spokoiny. An adaptive multiclass nearest neighbor classifier. *ESAIM: Probability and Statistics*, 24:69–99, 2020.
23. N. Puchkin and V. Spokoiny. Structure-adaptive manifold estimation. *Journal of Machine Learning Research*, 23(40):1–62, 2022.

24. R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18(1):104–117, 2003.
25. A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer Berlin Heidelberg, 2003.
26. G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.
27. P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 27(4):558–575, 2012.
28. 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(4):78–96, 2005.
29. A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.
30. G. Lecué. Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*, 19(5B):2153–2166, 2013.
31. D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905, 2012.
32. G. Lecué and P. Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.
33. C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. *Electronic Journal of Statistics*, 11(1):2258–2294, 2017.
34. P. Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2):639–665, 2012.
35. J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.
36. V. Dinh, L. S. T. Ho, N. V. Cuong, D. Nguyen, and B. T. Nguyen. Learning from non-iid data: fast rates for the one-vs-all multiclass plug-in classifiers. In *Theory and applications of models of computation*, volume 9076 of *Lecture Notes in Comput. Sci.*, pages 375–387. Springer, Cham, 2015.
37. E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.
38. A. Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 (3), pages 1220–1228. PMLR, 2013.
39. V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.
40. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
41. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
42. Z. Zhang and J. Wang. MLLE: Modified locally linear embedding using multiple weights. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.
43. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
44. L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
45. Z. Shi and J. Sun. Convergence of the point integral method for Laplace-Beltrami equation on point cloud. *Research in the Mathematical Sciences*, 4(1):22, 2017.
46. W. Wang and M. Á. Carreira-Perpiñán. Manifold blurring mean shift algorithms for manifold

denoising. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1759–1766, 2010.

47. D. Gong, F. Sha, and G. Medioni. Locally linear denoising on image manifolds. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 265–272. PMLR, 2010.
48. K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.
49. Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.
50. U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12(34):1249–1286, 2011.
51. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.
52. M. Á. Carreira-Perpiñán. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.
53. Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):363–410, 2002.
54. A. K. H. Kim and H. H. Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. *Electronic Journal of Statistics*, 9(1):1562–1582, 2015.
55. N. G. Trillos, D. Sanz-Alonso, and R. Yang. Local regularization of noisy point clouds: Improved global geometric estimates and data analysis. *Journal of Machine Learning Research*, 20(136):1–37, 2019.
56. C. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan. Fitting a putative manifold to noisy data. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 688–720. PMLR, 2018.