Journal of Machine Learning Research 23 (2022) 1-62 Submitted 3/21; Revised 12/21; Published 2/22

Structure-Adaptive Manifold Estimation

Nikita Puchkin NPUCHKIN@HSE.RU
National Research University Higher School of Economics,
Pokrovsky boulevard 11, 109028 Moscow, Russian Federation
and
Institute for Information Transmission Problems RAS,
Bolshoy Karetny per. 19, build.1, 127051 Moscow, Russian Federation

Vladimir Spokoiny SPOKOINY@WIAS-BERLIN.DE
Weierstrass Institute and Humboldt University,
Mohrenstrasse 39, 10117 Berlin, Germany
and
National Research University Higher School of Economics,
Pokrovsky boulevard 11, 109028 Moscow, Russian Federation
and
Institute for Information Transmission Problems RAS,
Bolshoy Karetny per. 19, build.1, 127051 Moscow, Russian Federation

Editor: Miguel Carreira-Perpinan

Abstract

We consider a problem of manifold estimation from noisy observations. Many manifold learning procedures locally approximate a manifold by a weighted average over a small neighborhood. However, in the presence of large noise, the assigned weights become so corrupted that the averaged estimate shows very poor performance. We suggest a structure-adaptive procedure, which simultaneously reconstructs a smooth manifold and estimates projections of the point cloud onto this manifold. The proposed approach iteratively refines the weights on each step, using the structural information obtained at previous steps. After several iterations, we obtain nearly "oracle" weights, so that the final estimates are nearly efficient even in the presence of relatively large noise. In our theoretical study, we establish tight lower and upper bounds proving asymptotic optimality of the method for manifold estimation under the Hausdorff loss, provided that the noise degrades to zero fast enough.

Keywords: manifold learning, manifold denoising, structural adaptation, adaptive procedures, minimax

1. Introduction

We consider a problem of manifold learning, that is, to recover a 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

©2022 Nikita Puchkin and Vladimir Spokoiny.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-0338.html.

learning gave rise to several novel nonlinear dimension reduction procedures, such as Isomap (Tenenbaum et al., 2000), locally linear embedding (Roweis and Saul, 2000, LLE) and its modification (Zhang and Wang, 2006), Laplacian eigenmaps (Belkin and Niyogi, 2003), and t-SNE (van der Maaten and Hinton, 2008). More recent works include interpolation on manifolds via geometric multi-resolution analysis (Maggioni et al., 2016), local polynomial estimators (Aamari and Levrard, 2019) and numerical solution of PDE (Shi and Sun, 2017). 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. (Hein and Maier, 2006; Wang and Carreira-Perpinan, 2010; Gong et al., 2010)), 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 (Fukunaga and Hostetler, 1975; Cheng, 1995) and its variants (Wang and Carreira-Perpinan, 2010; Ozertem and Erdogmus, 2011; Genovese et al., 2014). The mean-shift algorithm may be viewed as a generalized EM algorithm applied to the kernel density estimate (see (Carreira-Perpinan, 2007)). This algorithm and its modifications were extensively studied in the literature (Comaniciu and Meer, 2002; Hein and Maier, 2006; Li et al., 2007; Genovese et al., 2014; Arias-Castro et al., 2016). For a comprehensive review on mean-shift algorithms, a reader is referred to (Carreira-Perpinan, 2015). Though mean-shift algorithm was initially proposed for mode seeking and clustering, it found its applications in manifold denoising (see e.g. (Hein and Maier, 2006; Wang and Carreira-Perpinan, 2010; Ozertem and Erdogmus, 2011; Genovese et al., 2014; Carreira-Perpinan, 2015)). If the observations lie around a smooth manifold, then few iterations of the mean-shift algorithm move the data towards the manifold. However, since the mean shift algorithm and its variants (for example, subspace-constrained mean-shift (Ozertem and Erdogmus, 2011; Genovese et al., 2014) which is based on density ridges (Eberly et al., 1994)) 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 (Genovese et al., 2012a,b) consider the case, when the noise magnitude does not tend to zero as n grows. However, the approach in (Genovese et al., 2012a,b) assumes that the noise distribution is known and has a very special structure. For instance, considered in (Genovese et al., 2012a), 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. Moreover, the authors point out that their goal was to establish minimax rates rather than propose a practical estimator. 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 paper aims at studying the problem of manifold recovery under weak and realistic assumptions on the noise.

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

$$Y = X + \varepsilon. \tag{1}$$

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. In our work we are particularly interested in situations when b is of order 1 (non-orthogonal noise) and when b is small (*nearly* orthogonal noise) but our theoretical study is also valid for intermediate values of b (see Equation A4 below). We still have to assume that the noise magnitude $M = M(n)$ tends to zero as n tends to infinity but aim at describing the best possible rate of convergence still ensuring a consistent estimation. More precisely, if b is sufficiently small we allow M to be of order $n^{-2/(3d+8)}$, which is much slower than, for instance, $(\log n/n)^{2/d}$ and $(\log n/n)^{1/d}$, considered in (Aamari and Levrard, 2019) and (Aamari and Levrard, 2018), respectively (see the assumption (A4.1) for the precise statement). To the best of our knowledge, this is the first paper which provides a rigorous theoretical study in this setup as well as the setup for intermediate b.

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. (Hein and Maier, 2006; Maggioni et al., 2016; Osher et al., 2017; Aamari and Levrard, 2019). 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. (Xia et al., 2002) or (Hristache et al., 2001a). This paper extends the idea of *structural adaptation* proposed in (Hristache et al., 2001b,a). In these papers, the authors suggested to use anisotropic elliptic neighborhoods

1. A reader is referred to Section 2 for the definition.

with axes shinking in the direction of the estimated effective dimension reduction (e.d.r.) subspace and stretching in the orthogonal directions to estimate the e.d.r. subspace. As the shape of the local neighborhoods depends on the unknown e.d.r. structure, the procedure learns this structure from the data using iterations. This explains the name "structural adaptation". The use of anisotropic smoothing allows to obtain semiparametrically efficient root-n consistent estimates of the e.d.r. space (Xia et al., 2002; Hristache et al., 2001a). In our method, we construct cylindric neighborhoods, which are stretched in a normal direction to the manifold. However, our paper is not a formal generalization of (Hristache et al., 2001b) and (Hristache et al., 2001a). 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.

Now we briefly describe our procedure. Many manifold denoising procedures (see, for instance, (Hein and Maier, 2006; Genovese et al., 2014; Osher et al., 2017; Aamari and Levrard, 2018)) act in an iterative manner and our procedure is not an exception. We start with some guesses $\widehat{\boldsymbol{\Pi}}_1^{(0)}, \dots, \widehat{\boldsymbol{\Pi}}_n^{(0)}$ 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 $\widehat{\boldsymbol{\Pi}}_1^{(1)}, \dots, \widehat{\boldsymbol{\Pi}}_n^{(1)}$ of the projectors onto the tangent spaces of $\mathcal{M}^*$ at the points $X_1, \dots, X_n$, respectively, which are better than $\widehat{\boldsymbol{\Pi}}_1^{(0)}, \dots, \widehat{\boldsymbol{\Pi}}_n^{(0)}$. 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 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

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

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

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 for the definition of reach). Moreover, our algorithm estimates projectors $\boldsymbol{\Pi}(X_1), \dots, \boldsymbol{\Pi}(X_n)$ onto tangent spaces at $X_1, \dots, X_n$. It produces estimates $\widehat{\boldsymbol{\Pi}}_1, \dots, \widehat{\boldsymbol{\Pi}}_n$, such that

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

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

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

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

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. As pointed out in (Genovese et al., 2012b), the manifold estimation can be considered as a particular case of the error-in-variables regression problem. Then the rate $(M^2/n\log n)^{2/(d+4)}$ makes sense since it corresponds to an optimal accuracy of locally linear estimation with respect to $\|\cdot\|_\infty$-norm in a nonparametric regression problem (which is also $(M^2/n\log n)^{2/(d+4)}$). Besides, we prove a lower bound

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

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 3. Theorem 3, together with Theorem 1 from (Kim and Zhou, 2015), 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.

The rest of this paper is organized as follows. In Section 2, we formulate model assumptions and introduce notations. In Section 3, we provide our algorithm for manifold denoising and then illustrate its performance in Section 4. Finally, in Section 5, we give a theoretical justification of the algorithm and discuss its optimality. The proofs of the main results are collected in Section 6. Many technical details are contained in Appendix.

2. Model Assumptions

Let us remind that we consider the model (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 $\mathscr{M}^d_\varkappa$ 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:

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

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. One can also use the following equivalent definition of the reach (see (Genovese et al., 2012a, Section 2.1)). For a point $x \in \mathcal{M}$, let $\mathcal{T}_x\mathcal{M}$ stand for a tangent space of $\mathcal{M}$ at x, i. e. a linear space spanned by the derivative vectors of smooth curves on the manifold passing through x, and define a fiber

$$F_r(x) = \left(\{x\} \oplus \left(\mathcal{T}_x\mathcal{M}\right)^{\perp}\right) \cap \mathcal{B}(x, r),$$

where $(\mathcal{T}_x\mathcal{M})^{\perp}$ is an orthogonal complement of $\mathcal{T}_x\mathcal{M}$. Then $\operatorname{reach}(\mathcal{M})$ is a supremum of such $r > 0$ that for any $x, x' \in \mathcal{M}$, $x \neq x'$, the sets $F_r(x)$ and $F_r(x')$ do not intersect:

$$\operatorname{reach}(\mathcal{M}) = \sup\left\{r > 0 : \forall\, x, x' \in \mathcal{M}, x \neq x',\ F_r(x) \cap F_r(x') = \emptyset\right\}.$$

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 (Genovese et al., 2012a, 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

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

where $\boldsymbol{\Pi}(X)$ is the projector onto the tangent space $\mathcal{T}_X\mathcal{M}^*$. The model with manifold $\mathcal{M}^* \in \mathscr{M}^d_\varkappa$ and the bounded noise has been extensively studied in literature (see (Genovese et al., 2012a; Maggioni et al., 2016; Aamari and Levrard, 2018, 2019; Trillos et al., 2019)). In (Fefferman et al., 2018), 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 (Genovese et al., 2012a; Aamari and Levrard, 2019). The case $b = \varkappa$ corresponds to the bounded noise, which is not constrained to be orthogonal. Such model was considered, for instance, in (Aamari and Levrard, 2018). 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:

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

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 paper, 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 5 after Theorem 1 and compare it with other papers to convince the reader that the assumption $M \leqslant An^{-2/(3d+8)}$ is mild. In fact, to the best of our knowledge, only in (Genovese et al., 2012a,b) 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 5 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 Aamari and Levrard (2018), 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 5.

3. A Structure-adaptive Manifold Estimator (SAME)

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:

$$\widehat{X}_i^{(loc)} = \frac{\sum\limits_{j=1}^{n} w_{ij}^{(loc)} Y_j}{\sum\limits_{j=1}^{n} w_{ij}^{(loc)}}, \quad 1 \leqslant i \leqslant n, \tag{2}$$

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) 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 (Aamari and Levrard, 2018, Proposition 5.1) and (Aamari and Levrard, 2019, 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) 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 $\{Y : \|Y - Y_i\| \leqslant h\}$ around Y_i, we consider a cylinder $\{Y : \|\mathbf{\Pi}_i(Y_i - Y)\| \leqslant h\}$, 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

$$w_{ij}(\mathbf{\Pi}_i) = \mathcal{K}\left(\frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2}{h^2}\right) \mathbb{1}\left(\|Y_i - Y_j\| \leqslant \tau\right), \quad 1 \leqslant j \leqslant n, \tag{3}$$

with a constant $\tau < \varkappa$.

The adjusted weights (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 $\widehat{\mathbf{\Pi}}_1{}^{(0)}, \ldots, \widehat{\mathbf{\Pi}}_n^{(0)}$ of $\mathbf{\Pi}(X_1), \ldots, \mathbf{\Pi}(X_n)$ and take a large bandwidth h_0. Then we compute the weighted average estimates $\widehat{X}_1^{(1)}, \ldots, \widehat{X}_n^{(1)}$ with the adjusted weights (3) and the bandwidth h_0. These estimates can be then used to construct estimates $\widehat{\mathbf{\Pi}}_1^{(1)}, \ldots, \widehat{\mathbf{\Pi}}_n^{(1)}$ of $\mathbf{\Pi}(X_1), \ldots, \mathbf{\Pi}(X_n)$, which are better than $\widehat{\mathbf{\Pi}}_1^{(0)}, \ldots, \widehat{\mathbf{\Pi}}_n^{(0)}$. 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 1.

Let us discuss the role of the parameter γ in Algorithm 1. After the computation of the estimates $\widehat{X}_1^{(k)}, \ldots, \widehat{X}_n^{(k)}$, our goal is to use them to update the projectors $\widehat{\mathbf{\Pi}}_1^{(k)}, \ldots, \widehat{\mathbf{\Pi}}_n^{(k)}$.

Algorithm 1 Structure-adaptive manifold estimator (SAME)

1: 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 τ and constants $a > 1$ and $\gamma > 0$ are given.

2: **for** k from 0 to K **do**

3: Compute the weights $w_{ij}^{(k)}$ according to the formula

$$w_{ij}^{(k)} = \mathcal{K}\left(\frac{\|\widehat{\mathbf{\Pi}}_i^{(k)}(Y_i - Y_j)\|^2}{h_k^2}\right) \mathbb{1}\left(\|Y_i - Y_j\| \leqslant \tau\right), \quad 1 \leqslant i, j \leqslant n\,.$$

4: Compute the estimates

$$\widehat{X}_i^{(k)} = \sum_{j=1}^n w_{ij}^{(k)} Y_j \Big/ \left(\sum_{j=1}^n w_{ij}^{(k)}\right), \quad 1 \leqslant i \leqslant n\,. \tag{4}$$

5: 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

$$\widehat{\mathbf{\Sigma}}_i^{(k)} = \sum_{j \in \mathcal{J}_i^{(k)}} (\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})(\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})^T, \quad 1 \leqslant i \leqslant n\,.$$

6: 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.

7: If $k < K$, set $h_{k+1} = a^{-1} h_k$.

return the estimates $\widehat{X}_1 = \widehat{X}_1^{(K)}, \dots, \widehat{X}_n = \widehat{X}_n^{(K)}$.

However, our theoretical analysis (see Lemma 3 in the proof of Theorem 1 below) reveals that the weighted average (4) removes well an orthogonal component of the noise but causes a shift of $\widehat{X}_1^{(k)}, \dots, \widehat{X}_n^{(k)}$ in tangent direction. An illustration is given in Figure 1. Even if the noise is nearly orthogonal, the tangent component of $\widehat{X}_i^{(k)} - X_i$ is $O(h_k)$ while the orthogonal component is only $O(h_k^2/\varkappa)$. This means that if we take such $\widehat{X}_j^{(k)}$'s that X_j is close to X_i we can get a good estimate of the projector $\widehat{\mathbf{\Pi}}_i^{(k)}$. However, even if X_i and X_j are close, the distance between $\widehat{X}_i^{(k)}$ and $\widehat{X}_j^{(k)}$ can be as large as $O(h_K)$, because of the shift in the tangent direction. We take it into account and introduce an auxiliary parameter γ to construct a set of indices $\mathcal{J}_i^{(k)}$ which includes only those j's that X_j is close to X_i.

The computational complexity of Algorithm 1 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. SAME requires slightly more time than, for instance, the manifold blurring mean shift algorithm ((Wang and Carreira-Perpinan, 2010, MBMS), see the pseudocode in Appendix H below). The complexity of MBMS is $O(n^2D + n(D+\mathsf{k})(D \wedge \mathsf{k})^2)$ per iteration. Here k is the number of neighbors used by MBMS to perform local

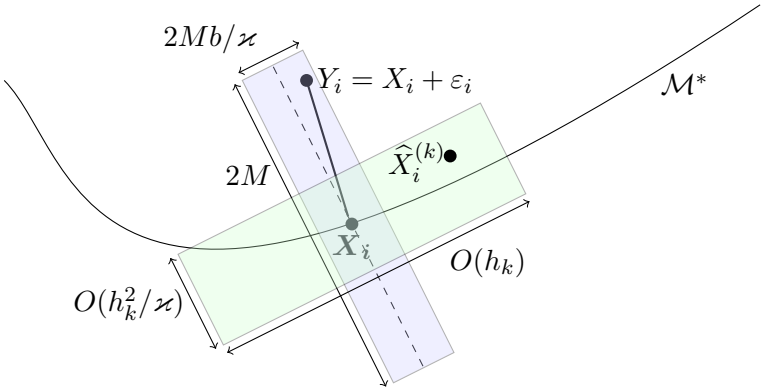

Figure 1: An illustration of how the weighted average estimate (4) induces a shift in a tangent direction. Even if the noise is nearly orthogonal and the observation lies in a thin blue cylinder around the point of interest, the weighted average estimate falls into the green rectangle stretched in the tangent directions with high probability. Nevertheless, the averaging successfully removes the orthogonal component of the noise.

PCA. SAME needs more operations to update the weights $w_{ij}^{(k)}$, $1 \leqslant i, j \leqslant n$, because of multiplication of the projectors $\widehat{\mathbf{\Pi}}_i^{(k)}$, $1 \leqslant i \leqslant n$, by the vectors $(Y_j - Y_i)$, $1 \leqslant i, j \leqslant n$. If the parameter k in MBMS is greater than D, then SAME and MBMS require the same time to perform PCA-type procedures.

4. Numerical Experiments

In this section, we carry out simulations to illustrate the performance of SAME. For convenience, theoretical results were obtained for manifolds without a boundary (which is a common assumption in the manifold learning literature) but we use some well-known surfaces with boundary in the experiments. The source code of all the numerical experiments described in this section is available on GitHub (link).

4.1 Manifold Denoising and Dimension Reduction

In this section, we present the performance of SAME on two widely known artificial data sets: Swiss Roll and S-shape. First, we show how our estimator denoises the manifold. We start with the description of the experiment with the Swiss Roll. We sampled $n = 2500$ points on a two-dimensional manifold in $\mathbb{R}^3$ and then embedded the surface into $\mathbb{R}^{20}$ adding 17 dummy coordinates. After that, we added a uniform noise with a magnitude 0.75 to each coordinate (thus, the noise magnitude M was equal to $0.75 \cdot \sqrt{20}$). In our algorithm, we initialized $\widehat{\mathbf{\Pi}}_i^{(0)} = \boldsymbol{I}_{20}$ for all i from 1 to n and made 6 iterations with $h_k^2 = h_0^2 \cdot 1.25^{-k}$, $0 \leqslant k \leqslant 5$, $\tau = h_0$, and $\gamma = 4$. To choose the initial bandwidth h_0, we took $\alpha = 0.015$ and put h_0 equal to the distance to the $\lfloor \alpha n \rfloor$-th nearest neighbor of the first sample point.

Data set	MSE of SAME, $\times 10^2$	MSE of MBMS, $\times 10^2$
Swiss Roll	**67.4**	69.3
S-shape	**3.9**	4.7

Table 1: Mean squared errors (MSE, (5)) of SAME and MBMS algorithms. Best results are boldfaced.

The parameter γ had minor influence on the behaviour of the algorithm and it was set to 4 in all the experiments. To quantify the performance of the algorithm, we used the mean squared error

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{X}_i - X_i\|^2. \tag{5}$$

The results are shown in Figure 2 (top) and in Table 1.

We compare SAME with the manifold blurring mean shift algorithm (Wang and Carreira-Perpinan, 2010, MBMS). We provide a pseudocode of MBMS in Appendix H below to make the paper self-contained. The parameters σ and k of MBMS as well as the number of iterations were chosen such that they minimized the mean squared error (5) over a range of parameters. The dimension d was set to 2. The smallest MSE was achieved with $\sigma = 2.6/\sqrt{2}$, $k = 150$ and only 1 iteration. The results are given in Figure 2 (top, right column). We observed that the mean squared error of MBMS grew after 1 or 2 iterations. In contrast to SAME, MBMS required much less iterations. However, SAME recovered the surface better than MBMS. The reason for that is hidden in the localizing weights $w_{ij}^{(k)}$, $1 \leqslant i, j \leqslant n$, $1 \leqslant k \leqslant K$. Projection onto a tangent hyperplane removes a large orthogonal component of the noise and, hence, $\|\widehat{\boldsymbol{\Pi}}_i^{(k)}(Y_j - Y_i)\|$ better approximates the distance between X_j and X_i than $\|Y_j - Y_i\|$. Besides, a weighted average estimate induces a shift in tangent direction (see the discussion after Algorithm 1 in Section 3) which may be harmful on early iterations. As a consequence, $\|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\|$ may also be a bad estimate for $\|X_j - X_i\|$.

The experiment with the S-shape manifold was carried in a similar way. We took $n = 1500$ points on the manifold in $\mathbb{R}^3$, embedded the surface into $\mathbb{R}^{30}$, and added a uniform noise with a magnitude 0.2 to each coordinate (thus, $M = 0.2 \cdot \sqrt{30}$). Again, we initialized $\widehat{\boldsymbol{\Pi}}_i^{(0)} = \boldsymbol{I}_{30}$ for all i from 1 to n. Then we put $h_k^2 = h_0^2 \cdot 1.25^{-k}$, $0 \leqslant k \leqslant 13$, i. e. the algorithm ran 14 iterations. The initial bandwidth h_0 was equal to the distance to the $\lfloor 0.1n \rfloor$-th nearest neighbor of the first sample point. The parameters τ and γ were equal to h_0 and 4, respectively. For MBMS, we took $\sigma = 0.45/\sqrt{2}$, $k = 300$, $d = 2$, and made 2 iterations. The tuning procedure of the parameters of MBMS was the same as in the example with the S-shape data set. The result of this experiment is displayed in Figure 2 (bottom) and in Table 1.

Next, we show how the preliminary denoising step may improve a dimension reduction. We consider the modified locally linear procedure (Zhang and Wang, 2006, MLLE for short), which is often used in applications due to its quality and computational efficiency. MLLE takes high-dimensional vectors as an input and returns their low-dimensional representation. In the case of S-shape and Swiss Roll data sets, one can easily find this map by straightening the curved surfaces into a plane. In the noiseless case, MLLE solves this

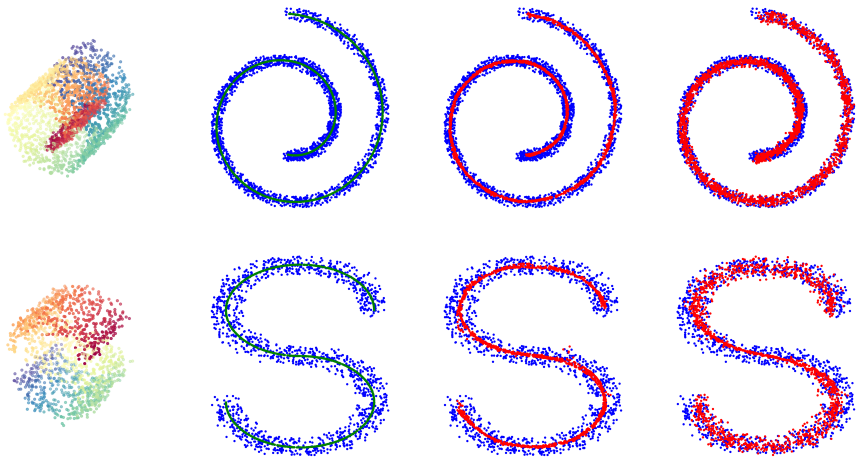

Figure 2: Perfomance of SAME and MBMS on the Swiss Roll data set (top) and on the S-shape data set (bottom). Column 1: noisy observations lying near a two-dimensional manifold. Column 2: noisy observations (blue) and the true manifold (green). Column 3: noisy observations (blue) and the projections onto the manifold (red), recovered by SAME. Column 4: noisy observations (blue) and the projections onto the manifold (red), recovered by MBMS.

task. However, as the other non-linear dimension reduction procedures based on Taylor's expansion, this algorithm deteriorates its performance in the presence of significant noise. In Figure 3 (center images) one can clearly observe that the MLLE procedure is not able to recognize a two-dimensional structure in the noisy data set. Instead of a rectangular-like shape, which would be a natural choice to represent the two-dimensional structure of the S-shape and Swiss Roll data sets, we have a curve. However, if one first uses SAME for manifold denoising and only after that applies MLLE for dimension reduction, then one obtains the desired result: both surfaces are straightened into planes. Of course, popular dimension reduction methods (e.g. Isomap (Tenenbaum et al., 2000), LLE (Roweis and Saul, 2000), MLLE (Zhang and Wang, 2006), Laplacian eigenmaps (Belkin and Niyogi, 2003), t-SNE (van der Maaten and Hinton, 2008)) still perform well in the presence of small noise. However, a researcher should consider an option of using preliminary manifold denoising before dimension reduction in the case of larger noise.

4.2 Manifold Denoising and Semi-supervised Learning

Manifold denoising can be a preprocessing step in semi-supervised learning. In the problem of semi-supervised learning, a statistician is usually given small amount of labelled data and a lot of unlabelled data. The goal is to recover the labels of the unlabelled points (transductive semi-supervised learning) or to propose a rule for prediction of label of a test point x ("true" semi-supervised learning). In semi-supervised learning, it is usually assumed that the unlabelled data carries useful information, which may be useful for prediction. The

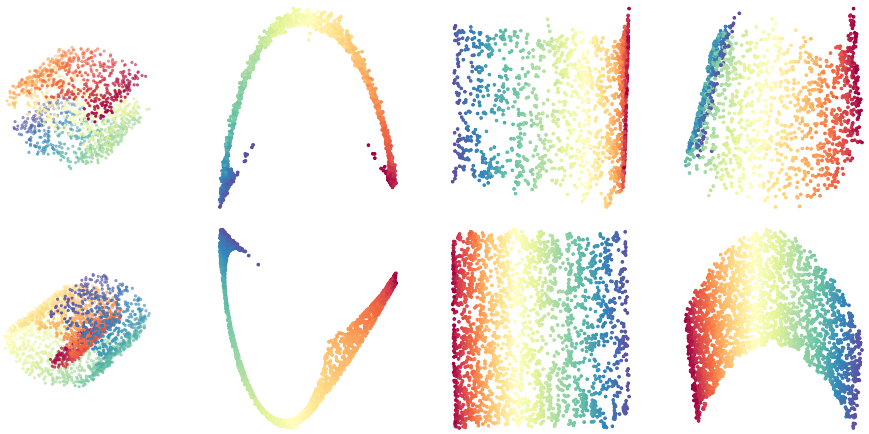

Figure 3: The role of manifold denoising in a successful dimension reduction for the S-shape data set (top) and Swiss Roll data set (Bottom). Column 1: noisy observations. Column 2: application of MLLE to the data set without denoising. Column 3: application of MLLE to the data set with a preliminary denoising via SAME. Column 4: application of MLLE to the data set with a preliminary denoising via MBMS.

most popular assumptions is that the data has a cluster structure or data points lie in a vicinity of a low-dimensional manifold.

In this section, we pursue the goal of recovering labels of unlabelled points. We take two artificial data sets g241c and g241n, which are described in (Chapelle et al., 2010). The data sets g241c and g241n have $n = 1500$ pairs (Y_i, Z_i), $1 \leqslant i \leqslant n$, where $Z_i \in \{-1, 1\}$ is a binary label and $Y \in \mathbb{R}^D$, $D = 241$ is a high-dimensional feature vector. According to (Chapelle et al., 2010), the data sets g241c and g241n were generated in such a way that they have a cluster structure and do not have hidden manifold structure. However, in (Hein and Maier, 2006), the authors report that preliminary manifold denoising step, applied to these data sets, improves the quality of classification. In this section, we illustrate that preliminary denoising with SAME also improves classification error.

We split the data sets into 100 train points and 1400 test points. We use k-nearest neighbors classifier as a baseline for two reasons. First, k-NN method is popular and often used in practice. Second, k-NN classifier is based on pairwise distances between feature vectors and should gain from the manifold denoising. For each of the data sets we perform the following procedure. First, we apply k-NN method without denoising. Then we make manifold denoising using SAME and apply k-NN to the denoised data set. In the case of g241c data set, we took $d = 10$, $\tau = 22$, $\gamma = 4$ and $h_k = 20 \cdot 1.2^{-k}$, $0 \leqslant k \leqslant 1$. In the case of g241n data set, we took $d = 6$, $\tau = 21$, $\gamma = 4$ and $h_k = 20 \cdot 1.2^{-k}$, $0 \leqslant k \leqslant 2$. The results are summarized in Table 2. We observe that preliminary denoising improves quality of prediction.

Data set	Best number of neighbors k without denoising	k-NN error without denoising (%)	Best number of neighbors k after denoising	denoised k-NN error (%)
g241c	21	31.3	15	27.8
g241n	18	27.1	12	25.3

Table 2: Error rates with and without manifold denoising via SAME for k-NN method, applied to artificial data sets g241c and g241n.

5. Theoretical Properties of SAME

This section states the main results. Here and everywhere in this paper, 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 1 *Assume* (A1), (A2), (A3), *and* (A4) . *Let the initial guesses* $\widehat{\boldsymbol{\Pi}}_1^{(0)}, \dots, \widehat{\boldsymbol{\Pi}}_n^{(0)}$ *of* $\boldsymbol{\Pi}(X_1), \dots, \boldsymbol{\Pi}(X_n)$ *be such that on an event with probability at least* $1 - n^{-1}$ *it holds*

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

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

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

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

where

$$\Phi_{M,b,h_K,\varkappa} = \frac{M^3(1 + b/h_K)^2}{h_K^2\varkappa} + \frac{M^2(1 + b/h_K + \sqrt{\log h_K^{-1}})}{\varkappa h_K} + \frac{Mh_K^2}{\varkappa^3} \lesssim \alpha + o(1), \quad n \to \infty,$$

$$\Psi_{M,b,h_K,\varkappa} = \left(1 + \frac{M(1 + b/h_K) \vee (1 + \Phi_{M,b,h_K,\varkappa})h_K}{\varkappa}\right)^{d+1} (1 + \Phi_{M,b,h_K,\varkappa}) \tag{6}$$

$$\leqslant (1 + \alpha)\left(4^{d+1} + (2\sqrt{\alpha})^{d+1}\right).$$

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 1, 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 (Aamari and Levrard, 2018, 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 eigenvalues of $\widehat{\mathbf{\Sigma}}_i^{(0)}$. Then the following result holds.

Proposition 1 (Aamari and Levrard (2018), Proposition 5.1) *Assume* (A1), (A2), (A3). *Set* $h_0 \gtrsim (\log n/n)^{1/d}$ *for large enough hidden constant. Let* $M/h_0 \leqslant 1/4$ *and let* $h_0 = h_0(n) = o(1)$, $n \to \infty$. *Then for* n *large enough, with probability larger than* $1 - n^{-1}$, *it holds*

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

Remark 2 *In (Aamari and Levrard, 2018), the authors take* $h_0 \asymp (\log n/n)^{1/d}$. *Nevertheless, a careful reading of the proofs reveals that one can also take larger values of* h_0.

Remark 3 *One can also use local PCA procedure from (Cheng and Wu, 2013) with a more sophisticated choice of the neighborhood for initialization.*

The condition (A4) and the choice of h_K in Theorem 1 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 (Aamari and Levrard, 2019), the authors use local polynomial estimates and require $M = O(h^2)$ and $h = h(n) \asymp n^{-1/d}$. In (Aamari and Levrard, 2018), the authors assume $M \leqslant \lambda(\log n/n)^{1/d}$, and λ does not exceed a constant λ_{d,p_0,p_1}, depending on d, p_0 and p_1. In (Fefferman et al., 2018), 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 (Genovese et al., 2014), where the authors also consider the Gaussian noise $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_D)$ 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 (Hein and Maier, 2006) the authors require $M = O(h)$. So, we see that the condition (A4) is quite mild.

Theorem 1 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 2 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 (Hristache et al., 2001b,a) 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 1 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 2 *Assume conditions of Theorem 1. Consider the piecewise linear manifold estimate*

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

where $\widehat{\boldsymbol{\Pi}}_i^{(K)}$ is a projector onto d-dimensional space obtained on the K-th iteration of Algorithm 1. 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

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

where $\Phi_{M,b,h_K,\varkappa}$ and $\Psi_{M,b,h_K,\varkappa}$ are defined in (6). 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

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

Remark 4 *For manifold reconstruction, one can also use a different technique, based on (Aamari and Levrard, 2018, Theorem 4.1) and tangential Delaunay complexes. We emphasize that the manifold estimate $\widehat{\mathcal{M}}$, used in Theorem 2, is not a manifold itself but a union of d-dimensional discs.*

Let us elaborate on the result of Theorem 2. 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 (Aamari and Levrard, 2018), 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, Aamari and Levrard (2018) 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 $M \leqslant \lambda_{d,p_0,p_1} (\log n/n)^{1/d}$, where the constant λ_{d,p_0,p_1} depends on d, p_0 and p_1. 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) 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 (Genovese et al., 2012a; Aamari and Levrard, 2019), so we find it useful to compare this more restrictive model with our upper bounds for the case of almost orthogonal noise. In (Genovese et al., 2012a), 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 (Aamari and Levrard, 2019, 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 $(\log n/n)^{2/d}$.

Theorem 2 shows that our procedure achieves the classical nonparametric rate, where the bias and the variance terms correspond to the best one can hope for when deals with the

locally linear estimator. In the case of small noise ($M \lesssim (\log n/n)^{2/d}$), the result of Theorem 2 matches the lower bound obtained in (Kim and Zhou, 2015) and the upper bound from (Aamari and Levrard, 2019). It is not surprising, because when $\mathcal{M}^*$ is a $\mathcal{C}^2$-manifold, the local polynomial estimate considered in (Aamari and Levrard, 2019) becomes a piecewise linear estimate, based on local PCA, and achieves the optimal rate in the case of small noise. Our algorithm acts in a similar manner and the only significant difference is hidden in the weights. However, if the noise is very small, there is no need to adjust the weights, so local PCA and SAME behave comparably in this regime.

The same concerns the projector estimates $\widehat{\mathbf{\Pi}}_1^{(K)}, \dots, \widehat{\mathbf{\Pi}}_n^{(K)}$ from Theorem 1. In the case of small noise, we recover the minimax rate $(\log n/n)^{1/d}$ obtained in (Aamari and Levrard, 2018, 2019). However, as the magnitude of the noise grows, our procedure shows superior performance, compared to the estimates in (Aamari and Levrard, 2018, 2019).

The result of Theorem 2 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 3 *Suppose that the sample $\mathbb{Y}_n = \{Y_1, \dots, Y_n\}$ is generated according to the model* (1)*, where $\mathcal{M}^* \in \mathscr{M}_\varkappa^d$, the density $p(x)$ of X fulfils* (A2) *(with sufficiently large p_1, L and sufficiently small p_0) and the noise ε satisfies* (A3)*. Then, for any estimate $\widehat{\mathcal{M}}$, it holds that*

$$\sup_{\mathcal{M}^* \in \mathscr{M}_\varkappa^d} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \frac{M^2 b^2}{\varkappa^3}. \tag{7}$$

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

$$\sup_{\mathcal{M}^* \in \mathscr{M}_\varkappa^d} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim \varkappa^{-1} \left(\frac{M^2\varkappa^2 \log n}{n}\right)^{\frac{2}{d+4}}. \tag{8}$$

Theorem 3 studies the case $M \gtrsim (\log n/n)^{2/d}$. In (Kim and Zhou, 2015), the authors proved the minimax lower bound

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

for the noiseless case, which is also tight for $M \lesssim (\log n/n)^{2/d}$. Theorem 3, together with (Kim and Zhou, 2015, Theorem 1) yields that SAME is minimax optimal in the model with almost orthogonal noise. The lower bounds (7) and (8) are completely new and are different from the currently known results on manifold estimation from (Genovese et al., 2012a) and (Aamari and Levrard, 2019), where the authors studied a perpendicular noise fulfilling (A3) with $b = 0$. In (Genovese et al., 2012a), the authors focused on the case of uniform noise

and proved the lower bound $(M/n)^{2/(d+2)}$ when $\mathcal{M}^*$ fulfils (A1) and $p_0 \leqslant p(x) \leqslant p_1$ for all $x \in \mathcal{M}^*$. In (Aamari and Levrard, 2019), the authors went further and proved that

$$\inf_{\widehat{\mathcal{M}}} \sup_{\mathcal{M}^*} \mathbb{E} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \gtrsim (M/n)^{k/(d+k)} \vee n^{-k/d},$$

where $\mathcal{M}^*$ runs over a class of compact, connected $\mathcal{C}^k$-manifolds of dimension d without a boundary and $\operatorname{reach}(\mathcal{M}^*) \geqslant \varkappa$. Theorem 3 reveals a surprising effect: if one allows small deviations of ε_i's in tangent directions then the problem of manifold estimation becomes harder and this fact is reflected in the minimax rates of convergence. Namely, if $b \gtrsim (\log n/n)^{1/d} \vee (M^2\varkappa^2 \log n/n)^{1/(d+4)}$, the minimax rate is $M^2b^2/\varkappa^3 \vee \left(M^2 \log n/n\right)^{2/(d+4)} \vee (\log n/n)^{2/d}$ which is different from the best known lower bound $(M/n)^{2/(d+2)} \vee (\log n/n)^{2/d}$ for the case of perpendicular noise.

6. Proofs

This section collects the proof of the main results.

6.1 Proof of Theorem 1

The proof of Theorem 1 is given in several steps. First, we show that the adjusted weights $w_{ij}(\mathbf{\Pi}_i)$ are informative, i.e. significant weights correspond only to points X_j, which are close to X_i.

Lemma 1 *Assume* (A1), (A3). *Let* $\mathbf{\Pi}_i$ *be any projector, such that* $\|\mathbf{\Pi}_i - \mathbf{\Pi}(X_i)\| \leqslant \frac{\Delta h}{\varkappa}$. *Assume that* $M \leqslant \varkappa/16$, $\Delta h \leqslant \varkappa/4$, *and* $M(\Delta + b/h) \leqslant \varkappa/4$. *Then for any* i *and* j, *such that* $\|Y_i - Y_j\| \leqslant 0.5\varkappa$, *it holds*

$$\frac{1}{2}\|\mathbf{\Pi}_i(Y_i - Y_j)\| - \frac{2M(\Delta h + b)}{\varkappa} \leqslant \|X_i - X_j\| \leqslant 2\|\mathbf{\Pi}_i(Y_i - Y_j)\| + \frac{4M(\Delta h + b)}{\varkappa}.$$

Lemma 1 quantifies the informal statement $\mathbf{\Pi}_i(Y_j - Y_i) \approx X_j - X_i$, giving explicit error bound depending on the error in the guess of the projector. Note that if (A4) holds, $h \in [h_K, h_0]$ and h_0, h_K satisfy the conditions of Theorem 1 then $M \leqslant \varkappa/16$, $M(\Delta + b/h) \leqslant \varkappa/4$, $\Delta h \leqslant \varkappa/4$ if n is large enough. The next step is to show that the cylinder $\{y : \|\mathbf{\Pi}_i(y - Y_i)\| \leqslant h, \|y - Y_i\| \leqslant \tau\}$ contains enough sample points. For this purpose, we prove the regularity of the design points in the following sense.

Lemma 2 *Assume* (A1)—(A3). *Fix any* i *from* 1 *to* n *and let* $\mathbf{\Pi}_i$ *be any projector, such that* $\|\mathbf{\Pi}_i - \mathbf{\Pi}(X_i)\| \leqslant \frac{\Delta h}{\varkappa}$, *where* $(\log n/n)^{1/d} \lesssim h < h_0$, $\Delta h \leqslant \varkappa/4$, $M \leqslant \varkappa/16$, *and* $M(\Delta + b/h) \leqslant \varkappa/4$. *Suppose that* h_0 *is chosen in a such way that* $h_0 \leqslant 0.5\tau$, *and* n *is sufficiently large. Then, on an event with probability at least* $1 - n^{-2}$, *it holds*

$$\sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) \geqslant C' n h^d \tag{9}$$

with an absolute constant $C' > 0$.

Roughly speaking, the cylinder $\{y : \|\mathbf{\Pi}_i(y - Y_i)\| \leqslant h, \|y - Y_i\| \leqslant \tau\}$ contains $\sim nh^d$ sample points with high probability. It is important, because the ball $\mathcal{B}(Y_i, h)$ would contain $\sim nh^D$ sample points if $h < M$. The sum of weights controls the variance of our estimates. From this point of view, the choice of cylindric neighborhoods instead of the balls yields much better rates.

Now, we are ready to make the main step in the proof of Theorem 1.

Lemma 3 *Assume conditions of Theorem 1. Let $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_n$ be any (possibly random) projectors, such that $\|\mathbf{\Pi}_i - \mathbf{\Pi}(X_i)\| \leqslant \frac{\Delta h}{\varkappa}$ almost surely, $(\log n/n)^{1/d} \lesssim h \leqslant h_0$, $\Delta h \leqslant \varkappa/4$, $M \leqslant \varkappa/16$, and $M(\Delta + b/h) \leqslant \varkappa/4$. Let $w_{ij}(\mathbf{\Pi}_i)$, $1 \leqslant i, j \leqslant n$, be the localizing weights, computed according to*

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

with a constant $\tau < 0.5\varkappa$. Then, conditionally on $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_n$, on an event with probability at least $1 - 2n^{-1}$, it simultaneously holds

$$\begin{aligned}
&\max_{1 \leqslant i \leqslant n} \left\| \sum_{j=1}^n w_{ij}(\mathbf{\Pi}_i)(Y_j - X_i) \right\| \lesssim \left(M\left(\Delta + \frac{b}{h}\right) \vee h \vee \frac{\Delta^2 h^2}{\varkappa} \right) \frac{h^{d+1}}{\varkappa} \\
&+ \Phi_{M,b,h,\varkappa,\Delta} \frac{nh^{d+2}}{\varkappa} + \sqrt{D(h^2 \vee M^2) nh^d \log n}, \\
&\max_{1 \leqslant i \leqslant n} \left\| \sum_{j=1}^n w_{ij}(\mathbf{\Pi}_i)\big((\boldsymbol{I} - \mathbf{\Pi}(X_i))(Y_j - X_i)\big) \right\| \\
&\lesssim \left(1 + \Phi_{M,b,h,\varkappa,\Delta}\right) \frac{nh^{d+2}}{\varkappa} + \sqrt{D(h^4/\varkappa^2 \vee M^2) nh^d \log n},
\end{aligned}$$

where

$$\Phi_{M,b,h,\varkappa,\Delta} = \frac{M^3(1 + \Delta + b/h)^2}{h^2 \varkappa} + \frac{M^2(\Delta + b/h + \sqrt{\log h^{-1}})}{\varkappa h} + \frac{(1 + \Delta^4) M h^2}{\varkappa^3},$$

and the hidden constants do not depend on Δ.

The proof of Lemma 3 is moved to Appendix C. In Lemma 3, the assumption (A4) comes into play. The condition (A4) implies that $\Phi_{M,b,h,\varkappa,\Delta} \leqslant \alpha + o(1)$ as $n \to \infty$. This follows from the fact that, under (A4), for any $h = h(n) \geqslant h_K$, it holds $M^3 = o(h^2)$, $n \to \infty$. Indeed, we have

$$\begin{aligned}
\frac{M^3}{h^2} &\leqslant \frac{M^3}{h_K^2} = \frac{M^3}{(M^2/n)^{2/(d+4)}} \cdot \frac{(M^2/n)^{2/(d+4)}}{h_K^2} \\
&= M^{(3d+8)/(d+4)} \cdot n^{2/(d+4)} \cdot \frac{(M^2/n)^{2/(d+4)}}{h_K^2} \\
&\leqslant A \cdot \frac{(M^2/n)^{2/(d+4)}}{h_K^2} \leqslant \frac{A}{(\log n)^{2/(d+4)}} \to 0, \quad n \to \infty.
\end{aligned}$$

Moreover, under (A4), it holds

$$\frac{M^3b^2}{\varkappa h^4} \leqslant \frac{M^3b^2}{\varkappa h_K^4} \leqslant \frac{M^3b^2}{\varkappa\left(\frac{D\log n}{n}\right)^{\frac{4}{d}} \vee \varkappa\left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{4}{d+4}}} \leqslant \alpha.$$

Therefore, $\Phi_{M,b,h,\varkappa,\Delta} \leqslant \alpha + o(1)$ as $n \to \infty$. Similarly, if (A4.2) holds, we have $M^3b^2 = M^3\varkappa^2 = o(h_K^4)$, which also yields $\Phi_{M,b,h,\varkappa,\Delta} \to 0$.

We need one more auxiliary result. Lemma 1, 2 and 3 imply that if we have good guesses of the projectors $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ then we can get good estimates of $X_1, \dots, X_n$ even if the noise magnitude M is quite large. Now we have to show that these estimates of $X_1, \dots, X_n$ can be used to construct good estimates of $\mathbf{\Pi}(X_1), \dots, \mathbf{\Pi}(X_n)$ and then we can carry out the proof by induction. Our next lemma is devoted to this problem.

Let us work on the event, on which (9) holds with $h = h_k$. Note that

$$\|\widehat{X}_i^{(k)} - X_i\| = \left\| \frac{\sum\limits_{j=1}^n w_{ij}^{(k)} Y_j}{\sum\limits_{j=1}^n w_{ij}^{(k)}} - X_i \right\| = \frac{\left\| \sum\limits_{j=1}^n w_{ij}^{(k)} (Y_j - X_i) \right\|}{\sum\limits_{j=1}^n w_{ij}^{(k)}}$$

and

$$\begin{aligned} d(\widehat{X}_i^{(k)}, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) &= \left\| \widehat{X}_i^{(k)} - X_i - \mathbf{\Pi}(X_i)(\widehat{X}_i^{(k)} - X_i) \right\| \\ &= \frac{\left\| \sum\limits_{j=1}^n w_{ij}^{(k)} \big(Y_j - X_i - \mathbf{\Pi}(X_i)(Y_j - X_i)\big) \right\|}{\sum\limits_{j=1}^n w_{ij}^{(k)}}. \end{aligned}$$

Here we used the fact that the projection of a point x onto the tangent plane $\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*$ is given by

$$\pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}(x) = X_i + \mathbf{\Pi}(X_i)(x - X_i).$$

Then Lemma 3 and Lemma 2 immediately yield that, if we have

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(k)} - \mathbf{\Pi}(X_i)\| \leqslant \Delta h_k/\varkappa$$

on the k-th iteration with probability at least $1 - (5k+1)/n$, then

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i^{(k)} - X_i\| \lesssim \left(\frac{M(\Delta h_k + b) \vee (1 + \Phi_{M,b,h_k,\varkappa,\Delta}) h_k^2}{\varkappa} \vee \frac{\Delta^2 h_k^3}{\varkappa^2} \right) + \sqrt{\frac{D(h_k^2 \vee M^2)\log h_k^{-1}}{nh_k^d}},$$

$$\max_{1 \leqslant i \leqslant n} d(\widehat{X}_i^{(k)}, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) \lesssim \frac{(1 + \Phi_{M,b,h_k,\varkappa,\Delta}) h_k^2}{\varkappa} + \sqrt{\frac{D(h_k^4/\varkappa^2 \vee M^2)\log h_k^{-1}}{nh_k^d}}$$

with probability at least $1 - (5k+1)/n - 3/n = 1 - (5k+4)/n$. It only remains to check that the projector estimates $\widehat{\mathbf{\Pi}}_1^{(k+1)}, \dots, \widehat{\mathbf{\Pi}}_n^{(k+1)}$ also satisfy

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(k+1)} - \mathbf{\Pi}(X_i)\| \lesssim \frac{h_{k+1}}{\varkappa}$$

with high probability. The precise statement is given in the following lemma.

Lemma 4 *Assume conditions of Theorem 1. Let Ω_k be an event, such that on this event it holds*

$$\max_{1 \leqslant i \leqslant n} \|\widehat{X}_i^{(k)} - X_i\| \leqslant \beta_1 \left(h_k + \sqrt{\frac{D(h_k^2 \vee M^2) \log h_k^{-1}}{n h_k^d}} \right) \leqslant 2\beta_1 h_k,$$

$$\max_{1 \leqslant i \leqslant n} d(\widehat{X}_i^{(k)}, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) \leqslant \beta_2 \left(\frac{h_k^2}{\varkappa} + \sqrt{\frac{D(h_k^4/\varkappa^2 \vee M^2) \log h_k^{-1}}{n h_k^d}} \right) \leqslant \frac{2\beta_2 h_k^2}{\varkappa}. \quad (10)$$

Then there exists $\gamma \asymp 1 + \beta_1$ such that, with probability at least $\mathbb{P}(\Omega_k) - 2n^{-1}$, it holds

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(k+1)} - \mathbf{\Pi}(X_i)\| \lesssim \frac{\gamma(\gamma + 4\beta_1)^d \beta_2 h_k}{\varkappa} \lesssim \frac{(1+\beta_1)^{d+1} \beta_2 h_k}{\varkappa}.$$

The proof of Lemma 4 can be found in Appendix D. From the derivations before Lemma 4, the event Ω_k from Lemma 4 has probability at least $1 - (5k+4)/n$. Note that, if $h_k \gtrsim \big((D \log n/n)^{1/d} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\big)$ with a hidden constant greater than 1, as given in the conditions of Theorem 1, then the bias terms in (10) are dominating, i.e.

$$h_k + \sqrt{\frac{D(h_k^2 \vee M^2) \log h_k^{-1}}{n h_k^d}} \leqslant 2h_k = 2a h_{k+1}$$

and

$$\frac{h_k^2}{\varkappa} + \sqrt{\frac{D(h_k^4/\varkappa^2 \vee M^2) \log h_k^{-1}}{n h_k^d}} \leqslant \frac{2h_k^2}{\varkappa}.$$

Due to the discussion before Lemma 4, we can take

$$\beta_1 = \left(\frac{M(\Delta + b/h_k) \vee (1 + \Phi_{M,b,h_k,\varkappa,\Delta}) h_k}{\varkappa} \vee \frac{\Delta^2 h_k^2}{\varkappa^2} \right),$$
$$\beta_2 = 1 + \Phi_{M,b,h_k,\varkappa,\Delta}.$$

Then Lemma 4 yields

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(k+1)} - \mathbf{\Pi}(X_i)\| \lesssim a(1+\beta_1)^{d+1} \beta_2 \frac{h_{k+1}}{\varkappa}. \quad (11)$$

The proof of Theorem 1 goes by induction. Let C be the hidden constant in (11). Assume that on the k-th iteration

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(k)} - \mathbf{\Pi}(X_i)\| \leqslant \frac{\Delta \vee (1+\alpha) C a (4^{d+1} + (2\sqrt{\alpha})^{d+1})}{\varkappa} h_k$$

with probability at least $1 - (5k+1)/n$. Here α is the constant from (A4). Lemma 2, Lemma 3 and Lemma 4 imply that, with probability at least $1 - (5(k+1)+1)/n$, it holds

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(k+1)} - \mathbf{\Pi}(X_i)\| \leqslant aC(1+\beta_1)^{d+1} \beta_2 \frac{h_{k+1}}{\varkappa}.$$

We have to check that $(1+\beta_1)^{d+1}\beta_2 \leqslant (1+\alpha)\left(4^{d+1}+(2\sqrt{\alpha})^{d+1}\right)$. From the definition of β_1 and β_2, we have

$$\begin{aligned}
&\beta_2 = 1+\Phi_{M,b,h,\varkappa,\Delta} \leqslant 1+\alpha+o(1), \quad n\to\infty,\\
&\beta_1 \leqslant \frac{M((Ca(1+\alpha)(4^{d+1}+(2\sqrt{\alpha})^{d+1})\vee\Delta)+b/h_K)}{\varkappa}\\
&\quad \vee \frac{(1+\Phi_{M,h,\varkappa,\Delta})h_0}{\varkappa} \vee \frac{(Ca(1+\alpha)(4^{d+1}+(2\sqrt{\alpha})^{d+1})\vee\Delta)^2 h_0^2}{\varkappa^2}\\
&\quad = \frac{Mb}{\varkappa h_K}+o(1), \quad n\to\infty
\end{aligned}$$

Show that (A4) yields $Mb/(\varkappa h_K)\leqslant\sqrt{\alpha}+o(1)$, as $n\to\infty$. Then, if n is sufficiently large, i.e. $n\geqslant N_\Delta$, this will imply $\beta_1\leqslant\sqrt{\alpha}+1$, $\beta_2\leqslant 2(1+\alpha)$. Due to (A4),

$$\frac{M^2b^3}{\varkappa h_K^4} \leqslant \frac{M^2b^3}{\varkappa\left(\frac{D\log n}{n}\right)^{\frac{4}{d}}\vee\varkappa\left(\frac{DM^2\varkappa^2\log n}{n}\right)^{\frac{4}{d+4}}} \leqslant \alpha.$$

If $M\leqslant h_K^2/\varkappa$ then

$$\frac{Mb}{\varkappa h_K} \leqslant \frac{h_K b}{\varkappa^2} \leqslant \frac{h_K}{\varkappa} = o(1), \quad n\to\infty.$$

Otherwise, we have

$$\left(\frac{Mb}{\varkappa h_K}\right)^2 \leqslant \frac{M\varkappa}{h_K^2}\cdot\frac{M^2b^2}{\varkappa^2h_K^2} = \frac{M^3b^2}{\varkappa h_K^4} \leqslant \alpha.$$

Thus, $Mb/(\varkappa h_K)\leqslant\sqrt{\alpha}+o(1)$, $n\to\infty$, and we have $\beta_1\leqslant\sqrt{\alpha}+1$, $\beta_2\leqslant 2(1+\alpha)$ for $n\geqslant N_\Delta$. This yields

$$\begin{aligned}
&(1+\beta_1)^{d+1}\beta_2 \leqslant 2(1+\alpha)\left(2+\sqrt{\alpha}\right)^{d+1}\\
&\leqslant 2(1+\alpha)\cdot 2^d\left(2^{d+1}+\alpha^{(d+1)/2}\right) = (1+\alpha)\left(4^{d+1}+(2\sqrt{\alpha})^{d+1}\right).
\end{aligned}$$

Thus, on the next iteration we have

$$\max_{1\leqslant i\leqslant n}\|\widehat{\mathbf{\Pi}}_i^{(k+1)}-\mathbf{\Pi}(X_i)\| \leqslant \frac{(1+\alpha)Ca\left(4^{d+1}+(2\sqrt{\alpha})^{d+1}\right)}{\varkappa}h_{k+1}$$

with probability at least $1-(5(k+1)+1)/n$. The confidence level $1-(5K+4)/n$ in the claim of Theorem 1 follows from the fact that we do not recompute the projectors on the final step.

6.2 Proof of Theorem 2

Fix any $x\in\widehat{\mathcal{M}}$ By definition of $\widehat{\mathcal{M}}$, there exist i and $u\in\mathcal{B}(0,1)$, such that

$$x = \widehat{X}_i + h_K\widehat{\mathbf{\Pi}}_i^{(K)}u.$$

Lemma 2, Lemma 3, Lemma 4, and the union bound imply that, with probability at least $1 - (5K+4)/n$,

$$\max_{1 \leqslant i \leqslant n} d(\widehat{X}_i, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) \lesssim \frac{(1 + \Phi_{M,b,h_K,\varkappa})h_K^2}{\varkappa} + \sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2)\log h_K^{-1}}{nh_K^d}}$$

and

$$\max_{1 \leqslant i \leqslant n} \|\widehat{\mathbf{\Pi}}_i^{(K)} - \mathbf{\Pi}(X_i)\| \lesssim \Psi_{M,b,h_K,\varkappa}\frac{h_K}{\varkappa},$$

where $\Phi_{M,b,h_K,\varkappa}$ and $\Psi_{M,b,h_K,\varkappa}$ are defined in (6). Recall that $\pi_{\mathcal{M}}(x)$ denotes the projection of x onto a closed set $\mathcal{M}$. Using the result Lemma 4, we immediately obtain

$$\begin{aligned} d(x, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) &= \inf_{v \in \mathbb{R}^D} \|\widehat{X}_i + h_K\widehat{\mathbf{\Pi}}_i^{(K)}u - X_i - \mathbf{\Pi}(X_i)v\| \\ &\leqslant d(\widehat{X}_i, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) \\ &+ \inf_{v \in \mathbb{R}^D} \left\|\pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}\left(\widehat{X}_i\right) + h_K\widehat{\mathbf{\Pi}}_i^{(K)}u - X_i - \mathbf{\Pi}(X_i)v\right\| \end{aligned} \tag{12}$$

Since the vector $\pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}(\widehat{X}_i) - X_i$ belongs to $\mathcal{T}_{X_i}\mathcal{M}^*$, we have

$$\pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}(\widehat{X}_i) - X_i = \mathbf{\Pi}(X_i)\left(\pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}(\widehat{X}_i) - X_i\right).$$

Then, substituting $v + X_i - \pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}(\widehat{X}_i)$ by $\widetilde{v}$, we obtain that the last expression in (12) is equal to

$$\begin{aligned} &d(\widehat{X}_i, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) \\ &+ \inf_{v \in \mathbb{R}^D} \left\|h_K\widehat{\mathbf{\Pi}}_i^{(K)}u - \mathbf{\Pi}(X_i)\left(v + X_i - \pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}\left(\widehat{X}_i\right)\right)\right\| \\ &= d(\widehat{X}_i, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) + \inf_{\widetilde{v} \in \mathbb{R}^D} \|h_K\widehat{\mathbf{\Pi}}_i^{(K)}u - \mathbf{\Pi}(X_i)\widetilde{v}\| \\ &\leqslant d(\widehat{X}_i, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) + \|h_K\widehat{\mathbf{\Pi}}_i^{(K)}u - h_K\mathbf{\Pi}(X_i)u\| \\ &\leqslant d(\widehat{X}_i, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) + h_K\|\widehat{\mathbf{\Pi}}_i^{(K)} - \mathbf{\Pi}(X_i)\| \\ &\lesssim \frac{(1 + \Phi_{M,b,h_K,\varkappa} + \Psi_{M,b,h_K,\varkappa})h_K^2}{\varkappa} + \sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2)\log h_K^{-1}}{nh_K^d}}. \end{aligned}$$

Next, note that, $\|x - \widehat{X}_i\| \leqslant h_K$ and, due to Theorem 1, we have

$$\begin{aligned} \|\pi_{\{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*}(x) - X_i\| &\leqslant \|x - X_i\| \leqslant \|x - \widehat{X}_i\| + \|\widehat{X}_i - X_i\| \\ &\lesssim \frac{Mb \vee Mh_K \vee (1 + \Phi_{M,b,h_K,\varkappa})h_K^2}{\varkappa} + \sqrt{\frac{D(h_K^2 \vee M^2)\log h_K^{-1}}{nh_K^d}} \\ &\lesssim \left(\frac{Mb}{\varkappa} \vee h_K\right) + \sqrt{\frac{D(h_K^2 \vee M^2)\log h_K^{-1}}{nh_K^d}} \lesssim \frac{Mb}{\varkappa} \vee h_K. \end{aligned}$$

The inequalities in the last line follow from the fact that $M < \varkappa$, $\Phi_{M,b,h_K,\varkappa} \lesssim \alpha + o(1)$, $n \to \infty$, and $h_K \geqslant \left((D \log n/n)^{1/d} \vee (DM^2\varkappa^2 \log n/n)^{1/(d+4)}\right)$ due to the conditions of Theorem 1. Since $\mathcal{M}^*$ is a $\mathcal{C}^2$-manifold with a reach at least $\varkappa$, it holds that

$$d(\pi_{\{X_i\}\oplus\mathcal{T}_{X_i}\mathcal{M}^*}(x), \mathcal{M}^*) \lesssim \frac{\|\pi_{\{X_i\}\oplus\mathcal{T}_{X_i}\mathcal{M}^*}(x) - X_i\|^2}{\varkappa} \lesssim \frac{h_K^2}{\varkappa} \vee \frac{M^2b^2}{\varkappa^3}.$$

Finally, we obtain

$$\begin{aligned} d(x, \mathcal{M}^*) &\leqslant d(x, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) + d(\pi_{\{X_i\}\oplus\mathcal{T}_{X_i}\mathcal{M}^*}(x), \mathcal{M}^*) \\ &\lesssim \left(\frac{(1 + \Phi_{M,b,h_K,\varkappa} + \Psi_{M,b,h_K,\varkappa})h_K^2}{\varkappa} \vee \frac{M^2b^2}{\varkappa^3}\right) + \sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2)\log h_K^{-1}}{nh_K^d}}. \end{aligned}$$

Thus, $\widehat{\mathcal{M}} \subseteq \mathcal{M}^* \oplus \mathcal{B}(0, r)$ with

$$r \lesssim \left(\frac{(1 + \Phi_{M,b,h_K,\varkappa} + \Psi_{M,b,h_K,\varkappa})h_K^2}{\varkappa} \vee \frac{M^2b^2}{\varkappa^3}\right) + \sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2)\log h_K^{-1}}{nh_K^d}}.$$

It remains to prove that $\mathcal{M}^* \subseteq \widehat{\mathcal{M}} \oplus \mathcal{B}(0, r)$ with the same r. Fix $x \in \mathcal{M}^*$. Note that there exist constants c_1 and r_0, such that

$$\begin{aligned} &\mathbb{P}_X(X \in \mathcal{B}(x, r)) \geqslant p_0 \mathrm{Vol}(\mathcal{B}(x, r) \cap \mathcal{M}^*) \geqslant c_1 p_0 r^d, \quad \forall r < r_0 \\ &\mathbb{P}_X(X \notin \mathcal{B}(x, r)) \leqslant 1 - c_1 p_0 r^d \leqslant e^{-c_1 p_0 r^d}, \quad \forall r < r_0 \end{aligned}$$

Let $\mathcal{N}_\varepsilon(\mathcal{M}^*)$ stand for an ε-net of $\mathcal{M}^*$. It is known (see, for example, Genovese et al. (2012a), Lemma 3) that $|\mathcal{N}_\varepsilon(\mathcal{M}^*)| \lesssim \varepsilon^{-d}$. Then

$$\begin{aligned} &\mathbb{P}(\exists\, x \in \mathcal{M}^* : \forall\, i \quad X_i \notin \mathcal{B}(x, 2\varepsilon)) \\ &\leqslant \mathbb{P}(\exists\, x \in \mathcal{N}_\varepsilon(\mathcal{M}^*) : \forall\, i \quad X_i \notin \mathcal{B}(x, \varepsilon)) \\ &\leqslant \sum_{x \in \mathcal{N}_\varepsilon(\mathcal{M}^*)} \mathbb{P}(\forall\, i \quad X_i \notin \mathcal{B}(x, \varepsilon)) \lesssim \varepsilon^{-d} e^{-c_1 p_0 n \varepsilon^d}. \end{aligned}$$

This implies that with probability at least $1 - 1/n$

$$\sup_{x \in \mathcal{M}^*} \min_{1 \leqslant i \leqslant n} \|x - X_i\| \lesssim \left(\frac{\log n}{n}\right)^{\frac{1}{d}}.$$

According to (Federer, 1959, Theorem 4.18), on the same event we have

$$\sup_{x \in \mathcal{M}^*} \min_{1 \leqslant i \leqslant n} d(x, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*) \lesssim \varkappa^{-1} \left(\frac{\log n}{n}\right)^{\frac{2}{d}}.$$

Now, fix any $x \in \mathcal{M}^*$. Without loss of generality, assume that $\min_{1 \leqslant i \leqslant n} d(x, \{X_i\} \oplus \mathcal{T}_{X_i}\mathcal{M}^*)$ is attained with $i = 1$. Let $\pi_{\{X_1\}\oplus\mathcal{T}_{X_1}\mathcal{M}^*}(x)$ be the projection of x onto the tangent plane $\{X_1\} \oplus \mathcal{T}_{X_1}\mathcal{M}^*$. It is clear that

$$\left\|\pi_{\{X_1\}\oplus\mathcal{T}_{X_1}\mathcal{M}^*}(x) - X_1\right\| \lesssim \left(\frac{\log n}{n}\right)^{\frac{1}{d}} \leqslant \frac{h_K}{2}.$$

Then there exists $u \in \mathcal{B}(0,1)$, such that

$$\left\| \widehat{X}_1 + h_K \widehat{\mathbf{\Pi}}_1^{(K)} u - \pi_{\{X_1\} \oplus \mathcal{T}_{X_1}\mathcal{M}^*}(x) \right\| \lesssim \left(\frac{(1 + \Phi_{M,b,h_K,\varkappa} + \Psi_{M,b,h_K,\varkappa}) h_K^2}{\varkappa} \vee \frac{M^2 b^2}{\varkappa^3} \right) + \sqrt{\frac{D(h_K^4/\varkappa^2 \vee M^2) \log h_K^{-1}}{n h_K^d}}.$$

Thus,

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

6.3 Proof of Theorem 3

For the sake of convenience, the proof of Theorem 3 is divided into two steps. First, we use the following lemma to obtain the lower bound (7).

Lemma 5 *Suppose that the sample $\mathbb{Y}_n = \{Y_1, \dots, Y_n\}$ is generated according to the model (1), where $\mathcal{M}^* \in \mathscr{M}_\varkappa^d$, $\mathcal{M}^* \subseteq \mathcal{B}(0,R)$ with $R = \sqrt{\varkappa^2 + M^2 b^2/\varkappa^4}$, ε satisfies (A3), and the density $p(x)$ of X fulfils (A2) with $L > 0$, $p_0 \leqslant ((d+1)\omega_{d+1} R^d)^{-1}$, $p_1 \geqslant ((d+1)\omega_{d+1} R^d)^{-1}$, where ω_{d+1} is the volume of the unit Euclidean ball in $\mathbb{R}^{d+1}$. Then, for any estimate $\widehat{\mathcal{M}}$, it holds that*

$$\sup_{\mathcal{M}^* \in \mathscr{M}_\varkappa^d} \mathbb{E}_{\mathcal{M}^*} d_H(\widehat{\mathcal{M}}, \mathcal{M}^*) \geqslant \frac{M^2 b^2}{6\varkappa^3}.$$

The proof of Lemma 5 is moved to Appendix F.1. The construction used in the proof of Lemma 5 is extremely simple. We show that if $(\varepsilon \,|\, X)$ is supported on a tangent space $\mathcal{T}_X \mathcal{M}_0$, where $\mathcal{M}_0$ is a d-dimensional sphere of radius $\varkappa$, then a statistician cannot distinguish between $\mathcal{M}_0$ and a sphere $\mathcal{M}_1$ with greater radius.

Second, we prove the lower bound (8) using Lemma 6 below. The proof of Lemma 6 is based on application of (Tsybakov, 2009, Theorem 2.5) to a family of smooth manifolds with small bumps at different points.

Lemma 6 *Suppose that the sample $\mathbb{Y}_n = \{Y_1, \dots, Y_n\}$ is generated according to the model (1), where $\mathcal{M}^* \in \mathscr{M}_\varkappa^d$, the density $p(x)$ of X fulfils (A2) with sufficiently large p_1, sufficiently small $p_0 > 0$, and $L \geqslant 4p_1/3$. Let the noise ε satisfy (A3) with*

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

where the hidden constant is large enough. Then, for any estimate $\widehat{\mathcal{M}}$, if n is sufficiently large and $M\varkappa \gtrsim (\log n/n)^{2/d}$, it holds that

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

The proof of Lemma 6 is moved to Appendix F.2. The claim of Theorem 3 follows from Lemma 5 and Lemma 6.

Acknowledgments

The authors are grateful to the action editor and three anonymous reviewers for valuable suggestions and remarks. This work was partly supported by the German Ministry for Education and Research as BIFOLD. It was partly carried out within the framework of the HSE University Basic Research Program. Results of Section 5 have been obtained under support of the RSF grant No. 19-71-30020. Nikita Puchkin is a Young Russian Mathematics award winner and would like to thank its sponsors and jury.

Appendix A. Proof of Lemma 1

We have

$$\begin{aligned}
&\|X_j - X_i - \boldsymbol{\Pi}_i(Y_j - Y_i)\| \\
&\leqslant \|X_j - X_i - \boldsymbol{\Pi}_i(X_j - X_i)\| + \|\boldsymbol{\Pi}_i(\varepsilon_j - \varepsilon_i)\| \\
&\leqslant \|X_j - X_i - \boldsymbol{\Pi}(X_i)(X_j - X_i)\| \\
&+ \|\boldsymbol{\Pi}_i - \boldsymbol{\Pi}(X_i)\|\|X_j - X_i\| + \|\boldsymbol{\Pi}_i(\varepsilon_j - \varepsilon_i)\|\,.
\end{aligned}$$

According to (Federer, 1959, Theorem 4.18),

$$\|X_j - X_i - \boldsymbol{\Pi}(X_i)(X_j - X_i)\| \leqslant \frac{\|X_j - X_i\|^2}{2\varkappa}\,.$$

Consider the term $\|\boldsymbol{\Pi}_i(\varepsilon_j - \varepsilon_i)\|$. It holds

$$\begin{aligned}
\|\boldsymbol{\Pi}_i(\varepsilon_j - \varepsilon_i)\| &\leqslant \|(\boldsymbol{\Pi}_i - \boldsymbol{\Pi}(X_i)(\varepsilon_j - \varepsilon_i)\| + \|\boldsymbol{\Pi}(X_i)(\varepsilon_j - \varepsilon_i)\| \\
&\leqslant \frac{2M\Delta h}{\varkappa} + \|\boldsymbol{\Pi}(X_i)\varepsilon_i\| + \|\boldsymbol{\Pi}(X_j)\varepsilon_j\| + \|(\boldsymbol{\Pi}(X_i) - \boldsymbol{\Pi}(X_j))\varepsilon_j\| \qquad (13)\\
&\leqslant \frac{2M\Delta h}{\varkappa} + \frac{2Mb}{\varkappa} + \frac{3M\|X_j - X_i\|}{\varkappa},
\end{aligned}$$

where in the last inequality we used the condition (A3) and applied Proposition 3. Taking into account that $\|\boldsymbol{\Pi}_i - \boldsymbol{\Pi}(X_i)\| \leqslant \frac{\Delta h}{\varkappa}$, we conclude

$$\begin{aligned}
\|X_j - X_i - \boldsymbol{\Pi}_i(Y_j - Y_i)\| &\leqslant \frac{\|X_j - X_i\|^2}{2\varkappa} + \frac{\Delta h\|X_j - X_i\|}{\varkappa} \qquad (14)\\
&+ \frac{2M(\Delta h + b)}{\varkappa} + \frac{3M\|X_i - X_j\|}{\varkappa}.
\end{aligned}$$

Using the triangle inequality

$$\|X_i - X_j\| - \|\boldsymbol{\Pi}_i(Y_i - Y_j)\| \leqslant \|X_j - X_i - \boldsymbol{\Pi}_i(Y_j - Y_i)\|$$

and solving the quadratic inequality

$$\|X_i - X_j\| - \|\mathbf{\Pi}_i(Y_i - Y_j)\| \leqslant \frac{\|X_j - X_i\|^2}{2\varkappa} + \frac{(3M + \Delta h)\|X_j - X_i\|}{\varkappa} + \frac{2M(\Delta h + b)}{\varkappa}$$

with respect to $\|X_i - X_j\|$, we obtain

$$\|X_i - X_j\| \leqslant \frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\| + 2M(\Delta h + b)/\varkappa}{1 - (\Delta h + 3M)/\varkappa} \leqslant 2\|\mathbf{\Pi}_i(Y_i - Y_j)\| + \frac{4M(\Delta h + b)}{\varkappa}.$$

Here we used the fact that, due to condition of the lemma,

$$\Delta h + 3M \leqslant \frac{\varkappa}{4} + \frac{3\varkappa}{16} < \frac{\varkappa}{2}.$$

On the other hand, from (14) we have

$$\|X_i - X_j\| - \|\mathbf{\Pi}_i(Y_i - Y_j)\| \geqslant -\frac{\|X_j - X_i\|^2}{2\varkappa} - \frac{2M(\Delta h + b)}{\varkappa} - \frac{(3M + \Delta h)\|X_i - X_j\|}{\varkappa}.$$

If

$$\frac{\|X_i - X_j\|^2}{2\varkappa} + \frac{3M + \Delta h}{\varkappa}\|X_i - X_j\| + \frac{2M(\Delta h + b)}{\varkappa} \leqslant \frac{\|\mathbf{\Pi}_i(Y_j - Y_i)\|}{2},$$

then $\|X_i - X_j\| \geqslant 0.5\|\mathbf{\Pi}_i(Y_j - Y_i)\|$. Otherwise, it holds

$$\|X_i - X_j\| \geqslant -(3M + \Delta h) + \frac{1}{\varkappa}\sqrt{\left(\frac{\Delta h + 3M}{\varkappa}\right)^2 + \frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\| - 4M(\Delta h + b)/\varkappa}{\varkappa}}.$$

Introduce a function $g(t) = \sqrt{a^2 + t} - a$, $a > 0$, $t \geqslant -a^2$. The function $g(t)$ is concave, increasing and $g(0) = 0$. Therefore, for any t_0 and any $t \in [0, t_0]$ it holds

$$g(t) \geqslant g(t_0)\frac{t}{t_0}.$$

Taking $a = (\Delta h + 3M)/\varkappa$ and $t_0 = 1 - 4M(\Delta h + b)/\varkappa^2$, we immediately obtain

$$\|X_i - X_j\| \geqslant \left(-\frac{3M + \Delta h}{\varkappa} + \sqrt{\left(\frac{3M + \Delta h}{\varkappa}\right)^2 + \frac{\varkappa - 4M\Delta h/\varkappa}{\varkappa}}\right) \cdot \left(\|\mathbf{\Pi}_i(Y_i - Y_j)\| - \frac{4M(\Delta h + b)}{\varkappa}\right).$$

Now it is easy to see that, if $M \leqslant \varkappa/16$, $\Delta h \leqslant \varkappa/4$, and $M(\Delta + b/h) \leqslant \varkappa/4$ then $3M + \Delta h < \varkappa/2 < \varkappa$ and

$$1 - \frac{4M(\Delta h + b)}{\varkappa^2} \geqslant \frac{3}{4}\frac{(3M + \Delta h)^2}{\varkappa^2}.$$

The last inequality yields

$$-\frac{3M+\Delta h}{\varkappa}+\sqrt{\left(\frac{3M+\Delta h}{\varkappa}\right)^2+\frac{\varkappa-4M\Delta h/\varkappa}{\varkappa}}\geqslant\frac{1}{2},$$

which, in its turn, implies

$$\|X_i-X_j\|\geqslant\frac{\|\mathbf{\Pi}_i(Y_i-Y_j)\|}{2}-2M(\Delta h+b).$$

Appendix B. Proof of Lemma 2

Show that for any $\mathbf{\Pi}_i$, such that $\|\mathbf{\Pi}_i-\mathbf{\Pi}(X_i)\|\leqslant\Delta h/\varkappa$, it holds

$$\mathbb{E}^{(-i)}w_{i1}(\mathbf{\Pi}_i)\geqslant C_1h^d.$$

Here and further in this paper, $\mathbb{E}^{(-i)}(\cdot)\equiv\mathbb{E}(\cdot\,|\,(X_i,Y_i))$. Due to Lemma 1, we have

$$\|\mathbf{\Pi}_i(Y_i-Y_1)\|\leqslant2\|X_i-X_1\|+\frac{4M(\Delta h+b)}{\varkappa},$$

which yields

$$\|\mathbf{\Pi}_i(Y_i-Y_1)\|^2\leqslant8\|X_i-X_1\|^2+\frac{32M^2(\Delta h+b)^2}{\varkappa},$$

$$\begin{aligned}
&\mathbb{E}^{(-i)}w_{i1}(\mathbf{\Pi}_i)=\mathbb{E}^{(-i)}e^{-\frac{\|\mathbf{\Pi}_i(Y_i-Y_1)\|^2}{h^2}}\mathbb{1}\left(\|Y_i-Y_1\|\leqslant\tau\right)\\
&\geqslant\mathbb{E}^{(-i)}e^{-32M^2(\Delta+\alpha)^2/\varkappa^2}e^{-\frac{8\|X_i-X_1\|^2}{h^2}}\mathbb{1}\left(\|X_i-X_1\|\leqslant\tau-2M\right)\\
&\geqslant e^{-2}\int\limits_{\mathcal{M}^*\cap\mathcal{B}(X_i,\tau-2M)}e^{-\frac{8\|X_i-x\|^2}{h^2}}p(x)dW(x)\\
&\geqslant e^{-2}p_0\int\limits_{\mathcal{B}(X_i,h)}e^{-\frac{8\|X_i-x\|^2}{h^2}}dW(x)\\
&=\frac{p_0}{e^2}\int\limits_{\mathcal{E}^{-1}(\mathcal{B}(X_i,h))}e^{-\frac{8\|\mathcal{E}_{X_i}(p)-\mathcal{E}_{X_i}(0)\|^2}{h^2}}\sqrt{\det g(v)}dv.
\end{aligned}$$

Here we used the fact that, due to conditions of the lemma, $M(\Delta+b/h)\leqslant\varkappa/4$ and, also, $\tau-2M\geqslant0.5\tau\geqslant h_0\geqslant h$, if h_0 is chosen sufficiently small. Next, due to (Aamari and Levrard, 2019, Lemma 1),

$$\|\mathcal{E}_{X_i}(v)-\mathcal{E}_{X_i}(0)-v\|\leqslant C_\perp\|v\|^2\leqslant C_\perp\varkappa\|v\|.$$

It also holds $\det g(v)\geqslant\frac{1}{2}$ for any $v\in\mathcal{E}^{-1}(\mathcal{B}(X_i,h))$. Then there exists a constant C', depending on d, p_0 and $\varkappa$, such that

$$\mathbb{E}^{(-i)}w_{i1}(\mathbf{\Pi}_i)\geqslant2C'h^d.$$

Now, consider the sum

$$\sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i)$$

Given Y_i, the weights $w_{ij}(\mathbf{\Pi}_i)$ are conditionally independent and identically distributed. The Bernstein's inequality implies that

$$\mathbb{P}^{(-i)}\left(\sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) \leqslant C' h^d\right) \leqslant e^{-\frac{C'^2 n h^{2d}}{2\sigma^2 + 2C' h^d/3}} \leqslant e^{-C'' n h^d},$$

and $e^{-C'' n h^d} \leqslant n^{-1}$ if $h \gtrsim \left(\frac{\log n}{n}\right)^{1/d}$ (with a sufficiently large hidden constant). Therefore, with probability at least $1 - n^{-2}$, it holds

$$\sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) \geqslant C' h^d.$$

Appendix C. Proof of Lemma 3

Fix any i from 1 to n and denote $\mathbb{E}^{(-i)}(\cdot) \equiv \mathbb{E}\left(\cdot|(X_i, Y_i)\right)$ and $\mathbb{P}^{(-i)}(\cdot) \equiv \mathbb{P}\left(\cdot|(X_i, Y_i)\right)$. Also, let $\mathscr{P}_i(\Delta h/\varkappa)$ be a set of projectors $\mathbf{\Pi}$ onto d-dimensional space, such that $\|\mathbf{\Pi} - \mathbf{\Pi}(X_i)\| \leqslant \Delta h/\varkappa$. First, we study the supremum of the empirical process

$$\sup_{\mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)} \left\| \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i)(Y_j - X_i) \right\| = \sup_{\substack{u \in \mathcal{B}(0,1) \\ \mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)}} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (Y_j - X_i)\,.$$

The rest of the proof can be summarized as follows. First, we fix $u \in \mathcal{B}(0,1)$ and $\mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)$ and bound the supremum of the expectation

$$\sup_{\substack{u \in \mathcal{B}(0,1) \\ \mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)}} \mathbb{E}^{(-i)} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (Y_j - X_i)\,.$$

Then we provide uniform bounds on

$$\mathbb{E}^{(-i)} \sup_{\substack{u \in \mathcal{B}(0,1) \\ \mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)}} \left(\sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (Y_j - X_i) - \mathbb{E}^{(-i)} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (Y_j - X_i) \right).$$

Finally, we derive large deviation results for

$$\sup_{\substack{u \in \mathcal{B}(0,1) \\ \mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)}} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (Y_j - X_i) - \mathbb{E}^{(-i)} \sup_{\substack{u \in \mathcal{B}(0,1) \\ \mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)}} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (Y_j - X_i).$$

As it was said earlier, we start with bounds on the expectation. The rigorous result is given in the next proposition.

Proposition 2 *Under conditions of Theorem 1 and Lemma 3, for any $u \in \mathcal{B}(0,1)$ and $\mathbf{\Pi}_i \in \mathscr{P}_i(\Delta h/\varkappa)$, it holds*

$$\mathbb{E}^{(-i)} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (X_j - X_i) \lesssim \left(M(\Delta + b/h) \vee h \vee \frac{\Delta^2 h^2}{\varkappa} \right) \frac{h^{d+1}}{\varkappa}, \tag{a}$$

$$\mathbb{E}^{(-i)} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T (\boldsymbol{I} - \mathbf{\Pi}(X_i))(X_j - X_i) \lesssim \frac{n h^{d+2}}{\varkappa}, \tag{b}$$

$$\mathbb{E}^{(-i)} \sum_{j=1}^{n} w_{ij}(\mathbf{\Pi}_i) u^T \varepsilon_j \lesssim \Phi_{M,b,h,\varkappa,\Delta} \frac{n h^{d+2}}{\varkappa}, \tag{c}$$

where

$$\Phi_{M,b,h,\varkappa,\Delta} = \frac{M^3(1+\Delta+b/h)^2}{h^2 \varkappa} + \frac{M^2(\Delta + b/h + \sqrt{\log h^{-1}})}{\varkappa h} + \frac{(1+\Delta^4) M h^2}{\varkappa^3},$$

and the hidden constants do not depend on Δ.

The proof of Proposition 2 relies on Taylor's expansion but it is quite technical. Therefore, it is moved to Appendix E. We continue with a uniform bound on the expectation

$$\mathbb{E} \sup_{\substack{u \in \mathcal{B}(0,1), \\ \|\mathbf{\Pi}_i - \mathbf{\Pi}(X_i)\| \leqslant \Delta h/\varkappa}} \sum_{j=1}^{n} \left(w_{ij}(\mathbf{\Pi}_i) u^T Y_j - \mathbb{E} w_{ij}(\mathbf{\Pi}_i) u^T Y_j \right).$$

Introduce the class of functions

$$\mathcal{F}_i = \Big\{ f(y) = \mathcal{K}\left(\frac{\|\mathbf{\Pi}_i (Y_i - y)\|^2}{h^2} \right) \mathbb{1}(\|Y_i - y\| \leqslant \tau) u^T (y - X_i) : \\ \|\mathbf{\Pi}_i - \mathbf{\Pi}(X_i)\| \leqslant \Delta h/\varkappa,\, Y_i \in \mathcal{B}(X_i, M),\, u \in \mathcal{B}(0,1) \Big\}.$$

We use the same trick as in (Giné and Koltchinskii, 2006, Section 4). Note that the class

$$\mathcal{F}_i^{(1)} = \big\{ f_1(y) = \|\mathbf{\Pi}_i (Y_i - y)\| : \\ \|\mathbf{\Pi}_i - \mathbf{\Pi}(X_i)\| \leqslant \Delta h/\varkappa,\, Y_i \in \mathcal{B}(X_i, M) \big\} \tag{15}$$

is VC subgraph, because the stripe $\{y : \|\mathbf{\Pi}(Y_i - y)\| \leqslant t\}$ is an intersection of a finite number of halfspaces. According to (van der Vaart and Wellner, 1996, Theorem 2.6.18 (viii)), the class

$$\widetilde{\mathcal{F}}_i^{(1)} = \left\{ f_1(y) = \mathcal{K}\left(\frac{\|\mathbf{\Pi}_i (Y_i - y)\|^2}{h^2} \right) : \right. \\ \left. \|\mathbf{\Pi}_i - \mathbf{\Pi}(X_i)\| \leqslant \Delta h/\varkappa,\, Y_i \in \mathcal{B}(X_i, M) \right\}$$

is also VC subgraph, since $\mathcal{K}(\cdot)$ monotonously decreases. The class of balls

$$\mathcal{F}_i^{(2)} = \big\{ f_2(y) = \mathbb{1}(\|Y_i - y\| \leqslant \tau) : Y_i \in \mathcal{B}(X_i, M) \big\} \tag{16}$$

and the class of hyperplanes

$$\mathcal{F}_i^{(3)} = \left\{ f_3(y) = u^T(y - X_i) : u \in \mathcal{B}(0,1) \right\}$$

are VC subgraph. The functions from the classes $\widetilde{\mathcal{F}}_i^{(1)}, \mathcal{F}_i^{(2)}$ and $\mathcal{F}_i^{(3)}$ are bounded by $1, 1$ and $R + M$ respectively. Then there exist constants $\mathcal{A}$ and ν, depending only on the VC characteristics of the classes $\widetilde{\mathcal{F}}_i^{(1)}, \mathcal{F}_i^{(2)}$ and $\mathcal{F}_i^{(3)}$, such that

$$\mathcal{N}(\mathcal{F}_i, L_2(\mathbb{P}_n^{(-i)}), \delta) \leqslant \left(\frac{\mathcal{A}}{\delta}\right)^{\nu},$$

where $\mathcal{N}(\mathcal{F}_i, L_2(\mathbb{P}_n^{(-i)}), \delta)$ is the δ-covering number of $\mathcal{F}_i$ with respect to the $L_2(\mathbb{P}_n^{(-i)})$ metric. Theorem 6 in Recht et al. (2011) (see also Szarek (1998)) implies that we can take $\nu \lesssim Dd$ and $\mathcal{A}$ to be an absolute constant, which does not depend on D, d or $\varkappa$. Corollary 2.2 from Giné and Guillou (2002) implies

$$\mathbb{E}^{(-i)} \sup_{f \in \mathcal{F}_i} \sum_{j=1}^{n} \left(f(Y_j) - \mathbb{E}^{(-i)} f(Y_j) \right) \leqslant \mathcal{R}\sigma \left(\sqrt{Dn \log \frac{\mathcal{A}}{\sigma}} \vee D \log \frac{\mathcal{A}}{\sigma} \right), \tag{17}$$

with an absolute constant $\mathcal{R}$ and $\sigma^2 \geqslant \sup_{f \in \mathcal{F}} \mathrm{Var} f(Y_1)$. Lemma 1 yields

$$\|X_i - X_j\|^2 \leqslant 8\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2 + \frac{32M^2(\Delta + b/h)^2 h^2}{\varkappa^2}.$$

Using this, we can derive

$$\begin{aligned}
&\mathbb{E}^{(-i)} e^{-\frac{2\|\mathbf{\Pi}_i(Y_j - Y_i)\|^2}{h^2}} (u^T(Y_j - X_i))^2 \\
&\leqslant \mathbb{E}^{(-i)} e^{\frac{8M^2(\Delta+\alpha)^2}{\varkappa^2}} e^{-\frac{\|X_j - X_i\|^2}{4h^2}} \|Y_j - X_i\|^2 \\
&\leqslant 2e^{1/2} \mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{4h^2}} \|X_j - X_i\|^2 + 2e^{1/2} \mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{2h^2}} \|\varepsilon_j\|^2 \\
&\leqslant 2e^{1/2} \mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{4h^2}} \|X_j - X_i\|^2 + 2e^{1/2} \mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{4h^2}} M^2.
\end{aligned}$$

Here we used the fact that $4M(\Delta + \alpha) \leqslant \varkappa$. Next, due to Lemma 9, there exist absolute constants B_1 and B_2, such that

$$\begin{aligned}
&\mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{4h^2}} \|X_j - X_i\|^2 \leqslant B_1 h^{d+2}, \\
&\mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{4h^2}} \leqslant B_2 h^d.
\end{aligned}$$

Therefore, we can take $\sigma^2 = B^2(h^2 \vee M^2)h^d$ with an absolute constant B. Thus, there exists a constant $C_{R,M,d}$, depending on R, M and d only (but not on Δ), such that

$$\begin{aligned}
&\mathbb{E}^{(-i)} \sup_{f \in \mathcal{F}_i} \sum_{j=1}^{n} \left(f(Y_j) - \mathbb{E}^{(-i)} f(Y_j) \right) \\
&\leqslant \mathcal{R}B \sqrt{D(h^2 \vee M^2) n h^d \log \frac{\mathcal{A}}{B^2(h^2 \vee M^2)h^d}}.
\end{aligned}$$

Finally, we use the Talagrand's concentration inequality (Talagrand, 1996) and obtain bounds on large deviations of

$$\sup_{\substack{u\in\mathcal{B}(0,1),\\ \|\boldsymbol{\Pi}_i-\boldsymbol{\Pi}(X_i)\|\leqslant\Delta h/\varkappa}}\sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)u^T(Y_j-X_i).$$

More precisely, we use the version of Talagrand's inequality from (Bousquet, 2002), where a deviation bound with nice constants was derived. Denote

$$Z_i=\sup_{f\in\mathcal{F}_i}\sum_{j=1}^{n}\left(f(Y_j)-\mathbb{E}f(Y_j)\right).$$

Then (Bousquet, 2002, Theorem 2.3) claims that, on an event with probability $1-n^{-2}$, it holds that

$$Z_i\leqslant\mathbb{E}Z_i+\sqrt{4v\log n}+\frac{2\log n}{3},$$

with $v=n\sigma^2+2\mathbb{E}Z_i$ and the same σ as in (17). This, together with (a) and (17), yields

$$\sup_{\substack{u\in\mathcal{B}(0,1),\\ \|\boldsymbol{\Pi}_i-\boldsymbol{\Pi}(X_i)\|\leqslant\Delta h/\varkappa}}\sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)u^T(Y_j-X_i)\lesssim\left(M(\Delta+b/h)\vee h\vee\frac{\Delta^2h^2}{\varkappa}\right)\frac{h^{d+1}}{\varkappa}$$
$$+\Phi_{M,b,h,\varkappa,\Delta}\frac{nh^{d+2}}{\varkappa}+\sqrt{D(h^2\vee M^2)nh^d\log n}$$

on an event with probability at least $1-n^{-2}$. The union bound implies that, on an event with probability at least $1-n^{-1}$, it holds that

$$\max_{1\leqslant i\leqslant n}\left\|\sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)(Y_j-X_i)\right\|\lesssim\left(M(\Delta+b/h)\vee h\vee\frac{\Delta^2h^2}{\varkappa}\right)\frac{h^{d+1}}{\varkappa}$$
$$+\Phi_{M,b,h,\varkappa,\Delta}\frac{nh^{d+2}}{\varkappa}+\sqrt{D(h^2\vee M^2)nh^d\log n}.$$

The bound

$$\max_{1\leqslant i\leqslant n}\left\|\sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)\big((\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)\big)\right\|$$
$$\lesssim(1+\Phi_{M,b,h,\varkappa,\Delta})nh^{d+2}/\varkappa+\sqrt{D(h^4/\varkappa^2\vee M^2)nh^d\log n}$$

is proven in a completely similar way. Proposition 2 yields

$$\sup_{\substack{u\in\mathcal{B}(0,1),\\ \|\boldsymbol{\Pi}_i-\boldsymbol{\Pi}(X_i)\|\leqslant\Delta h/\varkappa}}\mathbb{E}^{(-i)}\sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)u^T(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)\lesssim(1+\Phi_{M,b,h,\varkappa,\Delta})\frac{nh^{d+2}}{\varkappa}.$$

Again, applying the VC subgraph argument and using (Giné and Guillou, 2002, Corollary 2.2), we obtain

$$\mathbb{E}^{(-i)} \sup_{\substack{u\in\mathcal{B}(0,1),\\ \|\boldsymbol{\Pi}_i-\boldsymbol{\Pi}(X_i)\|\leqslant\Delta h/\varkappa}} \sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)u^T(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)$$
$$\lesssim (1+\Phi_{M,b,h,\varkappa,\Delta})\frac{nh^{d+2}}{\varkappa} + \mathcal{R}\sigma'\left(\sqrt{Dn\log\frac{\mathcal{A}}{\sigma'}}\vee D\log\frac{\mathcal{A}}{\sigma'}\right).$$

The only difference is that we can take $(\sigma')^2 \asymp h^d(h^4/\varkappa^2 \vee M^2)$ in this case. The reason for that is (Federer, 1959, Theorem 4.18), which implies

$$\|(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(X_j-X_i)\| \leqslant \frac{\|X_j-X_i\|^2}{2\varkappa},$$

and then

$$\begin{aligned}
&\mathbb{E}^{(-i)} w_{ij}^2(\boldsymbol{\Pi}_i)\left(u^T(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)\right)^2\\
&\leqslant \mathbb{E}^{(-i)} w_{ij}^2(\boldsymbol{\Pi}_i)\|(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)\|^2\\
&\leqslant 2\mathbb{E}^{(-i)} w_{ij}^2(\boldsymbol{\Pi}_i)\|(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(X_j-X_i)\|^2\\
&+ 2\mathbb{E}^{(-i)} w_{ij}^2(\boldsymbol{\Pi}_i)\|\varepsilon_j\|^2 \lesssim nh^d\left(h^4/\varkappa^2\vee M^2\right).
\end{aligned}$$

Thus,

$$\mathbb{E}^{(-i)} \sup_{\substack{u\in\mathcal{B}(0,1),\\ \|\boldsymbol{\Pi}_i-\boldsymbol{\Pi}(X_i)\|\leqslant\Delta h/\varkappa}} \sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)u^T(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)$$
$$\lesssim (1+\Phi_{M,b,h,\varkappa,\Delta})\frac{nh^{d+2}}{\varkappa} + \sqrt{Dnh^d(h^4/\varkappa^2\vee M^2)\log h^{-1}}.$$

Finally, applying the Talagrand's concentration inequality (Bousquet, 2002, Theorem 2.3)), we obtain that, for a fixed i, with probability at least $1-n^{-2}$,

$$\left\|\sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)\right\|$$
$$\lesssim \left((1+\Phi_{M,b,h,\varkappa,\Delta})nh^{d+2}/\varkappa + \sqrt{D(h^4/\varkappa^2\vee M^2)nh^d\log n}\right).$$

Applying the union bound, we conclude

$$\max_{1\leqslant i\leqslant n}\left\|\sum_{j=1}^{n} w_{ij}(\boldsymbol{\Pi}_i)(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(Y_j-X_i)\right\|$$
$$\lesssim (1+\Phi_{M,b,h,\varkappa,\Delta})nh^{d+2}/\varkappa + \sqrt{D(h^4/\varkappa^2\vee M^2)nh^d\log n}.$$

Appendix D. Proof of Lemma 4

Throughout the proof of Lemma 4, we work on the event Ω_k, on which (10) holds. Consider

$$\widehat{\mathbf{\Sigma}}_i^{(k)} = \sum_{j=1}^{n} (\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})(\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)})^T \mathbb{1}\left(\|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\| \leqslant \gamma h_k\right).$$

Denote $v_{ij} = \mathbb{1}\left(\|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\| \leqslant \gamma h_k\right)$. Let

$$\begin{aligned} Z_{ij} &= \pi_{\{X_i\}\oplus\mathcal{T}_{X_i}\mathcal{M}^*}(X_j), \quad 1 \leqslant j \leqslant n, \\ \widehat{Z}_{ij}^{(k)} &= \pi_{\{X_i\}\oplus\mathcal{T}_{X_i}\mathcal{M}^*}\left(\widehat{X}_j^{(k)}\right), \quad 1 \leqslant j \leqslant n, \end{aligned}$$

and introduce a matrix

$$\widehat{\mathbf{\Xi}}_i^{(k)} = \sum_{j=1}^{n} v_{ij} (\widehat{Z}_{ij}^{(k)} - \widehat{Z}_{ii}^{(k)})(\widehat{Z}_{ij}^{(k)} - \widehat{Z}_{ii}^{(k)})^T.$$

From the conditions of Lemma 4, we have

$$\max_{1\leqslant j\leqslant n} \|\widehat{Z}_{ij}^{(k)} - \widehat{X}_j^{(k)}\| \leqslant \beta_2 \left(\frac{h_k^2}{\varkappa} + \sqrt{\frac{h_k^4/\varkappa^2 \vee M^2}{nh_k^d} D \log n}\right) \leqslant \frac{2\beta_2 h_k^2}{\varkappa}.$$

This yields

$$\begin{aligned} &\|\widehat{\mathbf{\Sigma}}_i^{(k)} - \widehat{\mathbf{\Xi}}_i^{(k)}\| \\ &= \sup_{u\in\mathcal{B}(0,1)} \left| \sum_{j=1}^{n} v_{ij} \left[(u^T(\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}))^2 - (u^T(\widehat{Z}_{ij}^{(k)} - \widehat{Z}_{ii}^{(k)}))^2 \right] \right| \\ &\leqslant \sum_{j=1}^{n} v_{ij} \left(\|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\| + \|\widehat{Z}_{ij}^{(k)} - \widehat{Z}_{ii}^{(k)}\| \right) \\ &\qquad \cdot \left(\|\widehat{X}_j^{(k)} - \widehat{Z}_{ij}^{(k)}\| + \|\widehat{X}_i^{(k)} - \widehat{Z}_{ii}^{(k)}\| \right). \end{aligned}$$

Since $T_{X_i}\mathcal{M}^*$ is a convex set, then

$$\begin{aligned} \|\widehat{Z}_{ij}^{(k)} - \widehat{Z}_{ii}^{(k)}\| &= \left\| \pi_{\{X_i\}\oplus\mathcal{T}_{X_i}\mathcal{M}^*}\left(\widehat{X}_j^{(k)}\right) - \pi_{\{X_i\}\oplus\mathcal{T}_{X_i}\mathcal{M}^*}\left(\widehat{X}_i^{(k)}\right) \right\| \\ &\leqslant \|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\|. \end{aligned}$$

Thus,

$$\|\widehat{\mathbf{\Sigma}}_i^{(k)} - \widehat{\mathbf{\Xi}}_i^{(k)}\| \leqslant \frac{8\beta_2 h_k^2}{\varkappa} \sum_{j=1}^{n} v_{ij} \|\widehat{X}_j^{(k)} - \widehat{X}_i^{(k)}\| \leqslant \frac{8\gamma\beta_2 h_k^3}{\varkappa} \sum_{j=1}^{n} v_{ij}.$$

Next, we are going to prove that, with probability at least $1 - n^{-2}$,

$$\begin{aligned} \sum_{j=1}^{n} v_{ij} &\leqslant \sum_{j=1}^{n} \mathbb{1}\left(\|X_j - X_i\| \leqslant (\gamma + 4\beta_1)h_k\right) \\ &\leqslant 2C' n(\gamma + 4\beta_1)^d h_k^d. \end{aligned} \tag{18}$$

The first inequality follows from the fact that $\|\widehat{X}_i^{(k)} - \widehat{X}_j^{(k)}\| \leqslant \gamma h_k$ implies $\|X_i - X_j\| \leqslant (\gamma + 4\beta_1)h_k$. Next, we have

$$\mathbb{P}^{(-i)}\left(\|X_j - X_i\| \leqslant (\gamma + 4\beta_1)h_k\right) = \int\limits_{\mathcal{M}^* \cap \|x - X_i\| \leqslant (\gamma+4\beta_1)h_k} p(x)dW(x)$$
$$\leqslant p_1 \int\limits_{\|v\| \leqslant (\gamma+4\beta_1)h_k} \sqrt{\det g(v)}dv.$$

Using the inequality $|\sqrt{\det g(v)} - 1| \lesssim d\|v\|^2/\varkappa^2$ (see (Trillos et al., 2019, Equation 2.1)), we have that $\sqrt{\det g(v)} \lesssim 1$, provided that $\|v\| \leqslant (\gamma + 4\beta_1)h_k$. Then

$$\mathbb{P}^{(-i)}\left(\|X_j - X_i\| \leqslant (\gamma + 4\beta_1)h_k\right) \lesssim \mathrm{Vol}(\mathcal{B}(0, (\gamma + 4\beta_1)h_k)) \lesssim (\gamma + 4\beta_1)^d h_k^d.$$

Thus, there exists a constant C, such that

$$\mathbb{P}^{(-i)}\left(\|X_j - X_i\| \leqslant (\gamma + 4\beta_1)h_k\right) \leqslant C(\gamma + 4\beta_1)^d h_k^d\,.$$

Denote $C' = C \vee 16/3$. The Bernstein's inequality yields

$$\mathbb{P}^{(-i)}\left(\sum_{j=1}^n \mathbb{1}\left(\|X_j - X_i\| \leqslant 2\gamma h_k\right) > 2C'nh_k^d\right) \leqslant e^{-\frac{(C'nh_k^d)^2}{2\cdot(C'nh_k^d) + 2/3\cdot(C'nh_k^d)}}$$
$$= e^{-\frac{3C'nh_k^d}{8}} \leqslant e^{-2nh_k^d} \leqslant e^{-2nh_K^d} \leqslant \frac{1}{n^2},$$

and then (18) holds. From now on, we are working on the event, on which (18) holds. On this event, we have

$$\|\widehat{\boldsymbol{\Sigma}}_i^{(k)} - \widehat{\boldsymbol{\Xi}}_i^{(k)}\| \leqslant 8\gamma(\gamma + 4\beta_1)^d\beta_2 C'nh_k^{d+3}\,. \tag{19}$$

Consider the matrix $\widehat{\boldsymbol{\Xi}}_i^{(k)}$. According to Lemma 7, we have the following guarantee on the spectral gap of $\widehat{\boldsymbol{\Xi}}_i^{(k)}$:

$$\lambda_d(\widehat{\boldsymbol{\Xi}}_i) - \lambda_{d+1}(\widehat{\boldsymbol{\Xi}}_i)$$
$$\geqslant \frac{c}{4}\left(1 - \frac{2}{c(\gamma - 4\beta_1)^d} - \sqrt{\frac{6C'}{c^2(\gamma - 4\beta_1)^d}}\right)(\gamma - 4\beta_1)^{d+2}nh_k^{d+2}$$
$$- 9C'n^{-2/d}(\gamma + 4\beta_1)^{d+2}nh_k^{d+2} - 16C'\beta_1^2(\gamma + 4\beta_1)^d nh_k^{d+2}$$
$$- \frac{C'(\gamma + 4\beta_1)^{d+4}nh_k^{d+4}}{\varkappa^2}.$$

with probability at least $1 - n^{-2}$. Take γ satisfying the inequalities

$$\begin{aligned}
&(\gamma - 4\beta_1)^d \geqslant \frac{8}{c},\\
&(\gamma - 4\beta_1)^d \geqslant \frac{96C'}{c^2},\\
&\frac{c}{32}(\gamma - 4\beta_1)^{d+2} \geqslant C' n^{-2/d}(\gamma + 4\beta_1)^{d+2},\\
&\frac{c}{32}(\gamma - 4\beta_1)^{d+2} \geqslant 16C'\beta_1^2(\gamma + 4\beta_1)^d,\\
&\frac{c}{32}(\gamma - 4\beta_1)^{d+2} \geqslant \frac{C'(\gamma + 4\beta_1)^{d+4}h_0^2}{\varkappa^2}.
\end{aligned} \tag{20}$$

Note that such γ always exists if $n^{-2/d}$ and h_0 are sufficiently small. Then

$$\lambda_d(\widehat{\boldsymbol{\Xi}}_i) - \lambda_{d+1}(\widehat{\boldsymbol{\Xi}}_i) \geqslant \frac{c}{8}nh_k^{d+2} - \frac{3c}{32}nh_k^{d+2} = \frac{c}{32}nh_k^{d+2}\,. \tag{21}$$

The Davis-Kahan $\sin\theta$ theorem (Davis and Kahan, 1970) and the inequalities (19), (21) imply that for a fixed i from 1 to n with probability at least $1 - 2n^{-2}$ it holds

$$\|\widehat{\boldsymbol{\Pi}}_i^{(k+1)} - \boldsymbol{\Pi}(X_i)\| \leqslant \frac{256\gamma(\gamma + 4\beta_1)^d\beta_2 C' nh_k^{d+3}}{cnh_k^{d+2}} = \widetilde{C}h_k$$

with $\widetilde{C} = (256\gamma(\gamma + 4\beta_1)^d\beta_2 C')/c$. Applying the union bound, we have that

$$\max_{1\leqslant i\leqslant n} \|\widehat{\boldsymbol{\Pi}}_i^{(k+1)} - \boldsymbol{\Pi}(X_i)\| \leqslant \widetilde{C}h_k$$

with probability at least $1 - 2n^{-1}$, and the proof of Lemma 4 is finished.

Appendix E. Proof of Proposition 2

The proof of Proposition 2 is divided into three parts for the sake of convenience. On each step we prove one of the inequalities (a), (b), (c).

E.1 Proof of Proposition 2a

First, consider the expression

$$\mathbb{E}^{(-i)}w_{ij}(\boldsymbol{\Pi}_i)u^T(X_j - X_i).$$

Let $r_d = 4h\sqrt{(d+2)\log h^{-1}}$. Then

$$\begin{aligned}
\mathbb{E}^{(-i)}w_{ij}(\boldsymbol{\Pi}_i)u^T(X_j - X_i) = {}& \mathbb{E}^{(-i)}w_{ij}(\boldsymbol{\Pi}_i)u^T(X_j - X_i)\mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right)\\
&+ \mathbb{E}^{(-i)}w_{ij}(\boldsymbol{\Pi}_i)u^T(X_j - X_i)\mathbb{1}\left(X_j \notin \mathcal{B}(X_i, r_d)\right).
\end{aligned}$$

Due to Lemma 1,

$$\|X_i - X_j\|^2 \leqslant 8\|\boldsymbol{\Pi}_i(Y_i - Y_j)\|^2 + \frac{32M^2(\Delta + b/h)^2h^2}{\varkappa^2}, \tag{22}$$

and, if $X_j \notin \mathcal{B}(X_i, r_d)$, we conclude

$$w_{ij}(\mathbf{\Pi}_i) \leqslant e^{-\frac{\|X_i - X_j\|^2}{8h^2} + \frac{4M^2(\Delta + b/h)^2}{\varkappa^2}} \leqslant e^{\frac{4M^2(\Delta+b/h)^2}{\varkappa^2} - \frac{\|X_i - X_j\|^2 + r_d^2}{16h^2}} \leqslant e^{\frac{1}{4} - \frac{\|X_i - X_j\|^2}{16h^2}} h^{d+2}.$$

Here we used the fact that, due to the conditions of Theorem 1, $M(\Delta + b/h) \leqslant \varkappa/4$. Using the equality

$$\max_{t>0} t e^{-\frac{t^2}{16h^2}} = 2h\sqrt{2}e^{-1/2},$$

we conclude

$$\mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) u^T (X_j - X_i) \mathbb{1}\left(X_j \notin \mathcal{B}(X_i, r_d)\right) \lesssim h^{d+3}. \tag{23}$$

We see that outside the ball $\mathcal{B}(X_i, r_d)$, the weights $w_{ij}(\mathbf{\Pi}_i)$ become very small. It remains to consider the event $\{X_j \in \mathcal{B}(X_i, r_d)\}$. We assume that h_0 is sufficiently small, so it holds $r_d \leqslant 2h_0\sqrt{2(d+2)\log h_0^{-1}} \leqslant \tau/4$. On this event $\|Y_i - Y_j\| \leqslant 2M + r_d < \tau$, which yields

$$\begin{aligned}
&\mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) u^T (X_j - X_i) \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right) \\
&= \mathbb{E}^{(-i)} e^{-\frac{\|\mathbf{\Pi}_i (Y_j - Y_i)\|^2}{h^2}} u^T (X_j - X_i) \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right).
\end{aligned}$$

Using the Taylor's expansion, one has

$$\begin{aligned}
e^{-\frac{\|\mathbf{\Pi}_i (Y_j - Y_i)\|^2}{h^2}} &= e^{-\frac{\|X_j - X_i\|^2}{h^2}} \\
&+ e^{-\frac{\|\xi\|^2}{h^2}} \left(\frac{\|\mathbf{\Pi}_i (Y_j - Y_i)\|^2 - \|X_j - X_i\|^2}{h^2} \right),
\end{aligned} \tag{24}$$

where $\xi = \theta(X_j - X_i) + (1-\theta)\mathbf{\Pi}_i(Y_i - Y_j)$ for some $\theta \in (0,1)$.

Consider the expectation

$$\mathbb{E}^{(-i)} e^{-\frac{\|\mathbf{\Pi}_i (X_j - X_i)\|^2}{h^2}} u^T (X_j - X_i) \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right).$$

According to Lemma 8, it does not exceed

$$\mathbb{E}^{(-i)} e^{-\frac{\|\mathbf{\Pi}_i (X_j - X_i)\|^2}{h^2}} u^T (X_j - X_i) \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right) \lesssim \frac{d h^{d+2}}{\varkappa}.$$

Next, consider the second term in (24). Note that

$$\begin{aligned}
\|\xi\| &= \|\theta \mathbf{\Pi}_i (Y_j - Y_i) + (1-\theta)(X_j - X_i)\| \\
&\geqslant \|X_i - X_j\| - \|\mathbf{\Pi}_i (Y_j - Y_i) - (X_j - X_i)\|.
\end{aligned}$$

From the proof of Lemma 1 (see Equation 14) we know that

$$\begin{aligned}
\|X_j - X_i - \mathbf{\Pi}_i (Y_j - Y_i)\| &\leqslant \frac{\|X_j - X_i\|^2}{2\varkappa} + \frac{\Delta h \|X_j - X_i\|}{\varkappa} \\
&+ \frac{2M(\Delta h + b)}{\varkappa} + \frac{3M\|X_i - X_j\|}{\varkappa}.
\end{aligned}$$

Then

$$\begin{aligned}\|\xi\| &\geqslant \left(1-\frac{3M+\Delta h}{\varkappa}\right)\|X_i-X_j\|-\frac{\|X_i-X_j\|^2}{2\varkappa}-\frac{2M(\Delta h+b)}{\varkappa}\\ &\geqslant \left(1-\frac{3M+\Delta h+r_d}{\varkappa}\right)\|X_i-X_j\|-\frac{2M(\Delta h+b)}{\varkappa}\\ &\geqslant \left(1-\left(\frac{3}{16}+\frac{1}{4}+\frac{1}{8}\right)\right)\|X_i-X_j\|-\frac{2M(\Delta h+b)}{\varkappa}\\ &> \frac{1}{4}\|X_i-X_j\|-\frac{2M(\Delta h+b)}{\varkappa}.\end{aligned}$$

This yields

$$\|X_i-X_j\|^2 \leqslant 16\|\xi\|^2+128M^2(\Delta+b/h)^2h^2/\varkappa^2 \leqslant 16\|\xi\|^2+8h^2. \tag{25}$$

Then

$$\begin{aligned}&\mathbb{E}^{(-i)}e^{-\frac{\|\xi\|^2}{h^2}}\left(\frac{\|\boldsymbol{\Pi}_i(Y_j-Y_i)\|^2-\|X_j-X_i\|^2}{h^2}\right)\\ &\quad\cdot u^T(X_j-X_i)\mathbb{1}\left(X_j\in\mathcal{B}(X_i,r_d)\right)\\ &\leqslant e^{1/2}\mathbb{E}^{(-i)}e^{-\frac{\|X_i-X_j\|^2}{16h^2}}\left|\frac{\|\boldsymbol{\Pi}_i(Y_j-Y_i)\|^2-\|X_j-X_i\|^2}{h^2}\right|\\ &\quad\cdot\|X_j-X_i\|\mathbb{1}\left(X_j\in\mathcal{B}(X_i,r_d)\right)\end{aligned} \tag{26}$$

Note that

$$\begin{aligned}&\|\boldsymbol{\Pi}_i(Y_j-Y_i)\|^2-\|X_j-X_i\|^2=\|\boldsymbol{\Pi}_i(\varepsilon_j-\varepsilon_i)\|^2\\ &+2(\varepsilon_j-\varepsilon_i)^T\boldsymbol{\Pi}_i(X_j-X_i)-\|(\boldsymbol{I}-\boldsymbol{\Pi}_i)(X_j-X_i)\|^2.\end{aligned}$$

Due to (Federer, 1959, Theorem 4.18),

$$\begin{aligned}\|(\boldsymbol{I}-\boldsymbol{\Pi}_i)(X_j-X_i)\| &\leqslant \|\boldsymbol{\Pi}_i-\boldsymbol{\Pi}(X_i)\|\|X_j-X_i\|+\|(\boldsymbol{I}-\boldsymbol{\Pi}(X_i))(X_j-X_i)\|\\ &\leqslant \frac{\Delta h\|X_j-X_i\|}{\varkappa}+\frac{\|X_j-X_i\|^2}{2\varkappa}.\end{aligned}$$

Using (13), we obtain

$$\begin{aligned}&\left|\|\boldsymbol{\Pi}_i(Y_j-Y_i)\|^2-\|X_j-X_i\|^2\right|\\ &\leqslant\left(\frac{2M(\Delta h+b)}{\varkappa}+\frac{3M\|X_j-X_i\|}{\varkappa}\right)^2\\ &+2\|X_j-X_i\|\left(\frac{2M(\Delta h+b)}{\varkappa}+\frac{3M\|X_j-X_i\|}{\varkappa}\right)\\ &+\left(\frac{\Delta h\|X_j-X_i\|}{\varkappa}+\frac{\|X_j-X_i\|^2}{2\varkappa}\right)^2.\end{aligned} \tag{27}$$

Next, it is useful to control the expectations of the form

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{16h^2}} \|X_j - X_i\|^q \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right).$$

Lemma 9 yields

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{16h^2}} \|X_j - X_i\|^q \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right) \lesssim h^{q+d}.$$

The inequalities (26) and (27) and Lemma 9 yield that, up to a multiplicative constant, the left-hand side of (26) is bounded by

$$\begin{aligned}
&\frac{M^2(\Delta + b/h)^2 h^{d+1}}{\varkappa^2} + \frac{M^2 h^{d+1}}{\varkappa^2} + \frac{M(\Delta + b/h)h^{d+1}}{\varkappa} + \frac{Mh^{d+3}}{\varkappa} + \frac{\Delta^2 h^{d+3}}{\varkappa^2} + \frac{h^{d+3}}{\varkappa^2} \\
&\lesssim \frac{M(\Delta + b/h + 1)h^{d+1}}{\varkappa} + \frac{(\Delta^2 + 1)h^{d+3}}{\varkappa^2}.
\end{aligned}$$

This and Lemma 8 imply

$$\mathbb{E}^{(-i)} w_{ij}(\boldsymbol{\Pi}_i) u^T (X_j - X_i) \lesssim \left(M(\Delta + b/h) \vee h \vee \frac{\Delta^2 h^2}{\varkappa}\right) \frac{h^{d+1}}{\varkappa} \tag{28}$$

with a hidden constant, which does not depend on Δ.

E.2 Proof of Proposition 2b

Consider the expectation

$$\mathbb{E}^{(-i)} \sum_{j=1}^n w_{ij}(\boldsymbol{\Pi}_i) u^T (\boldsymbol{I} - \boldsymbol{\Pi}(X_i))(X_j - X_i).$$

Since for each $j \neq i$ the summand has the same conditional distribution with respect to (X_i, Y_i), it is enough to prove that

$$\mathbb{E}^{(-i)} w_{ij}(\boldsymbol{\Pi}_i) u^T (\boldsymbol{I} - \boldsymbol{\Pi}(X_i))(X_j - X_i) \lesssim \frac{h^{d+2}}{\varkappa}$$

for any distinct j.

Again, we use the decomposition

$$\begin{aligned}
&\mathbb{E}^{(-i)} \sum_{j=1}^n w_{ij}(\boldsymbol{\Pi}_i) u^T (\boldsymbol{I} - \boldsymbol{\Pi}(X_i))(X_j - X_i) \\
&= \mathbb{E}^{(-i)} \sum_{j=1}^n w_{ij}(\boldsymbol{\Pi}_i) u^T (\boldsymbol{I} - \boldsymbol{\Pi}(X_i))(X_j - X_i) \mathbb{1}(X_j \in \mathcal{B}(X_i, r_d)) \\
&+ \mathbb{E}^{(-i)} \sum_{j=1}^n w_{ij}(\boldsymbol{\Pi}_i) u^T (\boldsymbol{I} - \boldsymbol{\Pi}(X_i))(X_j - X_i) \mathbb{1}(X_j \notin \mathcal{B}(X_i, r_d)).
\end{aligned}$$

From (23), the second term is of order $h^{d+3} \ll h^{d+2}/\varkappa$. On the event $\{X_j \in \mathcal{B}(X_i, r_d)\}$, we can use (Federer, 1959, Theorem 4.18):

$$\begin{aligned}
&\mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) u^T (\boldsymbol{I} - \mathbf{\Pi}(X_i))(X_j - X_i) \mathbb{1}(X_j \in \mathcal{B}(X_i, r_d)) \\
&\leqslant \mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) \, \|(\boldsymbol{I} - \mathbf{\Pi}(X_i))(X_j - X_i)\| \, \mathbb{1}(X_j \in \mathcal{B}(X_i, r_d)) \\
&\leqslant \frac{1}{2\varkappa} \mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) \|X_j - X_i\|^2 \mathbb{1}(X_j \in \mathcal{B}(X_i, r_d)).
\end{aligned}$$

Using (22), we obtain

$$w_{ij}(\mathbf{\Pi}_i) = e^{-\frac{\|\mathbf{\Pi}_i(Y_j - Y_i)\|^2}{h^2}} \leqslant e^{-\frac{\|X_j - X_i\|^2}{8h^2} + \frac{4M^2(\Delta + b/h)^2}{h^2}} \leqslant e^{\frac{1}{2} - \frac{\|X_j - X_i\|^2}{8h^2}}.$$

The assertion of Proposition 2b now follows from Lemma 9.

E.3 Proof of Proposition 2c

To complete the proof of Proposition 2, it remains to bound the expectation

$$\mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) u^T \varepsilon_j.$$

Outside the ball $\mathcal{B}(X_i, r_d)$, we have

$$\mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) u^T \varepsilon_j \mathbb{1}(X_j \notin \mathcal{B}(X_i, r_d) \lesssim M h^{d+2},$$

so it remains to control the expectation

$$\mathbb{E}^{(-i)} w_{ij}(\mathbf{\Pi}_i) u^T \varepsilon_j \mathbb{1}(X_j \in \mathcal{B}(X_i, r_d).$$

Again, use the Taylor's expansion:

$$\begin{aligned}
e^{-\frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2}{h^2}} &= e^{-\frac{\|X_i - X_j\|^2}{h^2}} + e^{-\frac{\|X_i - X_j\|^2}{h^2}} \left(\frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2 - \|X_i - X_j\|^2}{h^2} \right) \\
&\quad + e^{-\frac{\|\zeta\|^2}{h^2}} \frac{\left(\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2 - \|X_i - X_j\|^2 \right)^2}{2h^4},
\end{aligned}$$

where $\zeta = \vartheta \mathbf{\Pi}_i(Y_j - Y_i) + (1 - \vartheta)(X_j - X_i)$ for some $\vartheta \in (0, 1)$. On the event $\{\|X_i - X_j\| \leqslant r_d\}$ it holds $\|Y_i - Y_j\| \leqslant 2M + r_d \leqslant \tau$. This yields

$$\begin{aligned}
&\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \mathbb{1}(\|Y_i - Y_j\| \leqslant \tau) \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j \\
&= \mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j = 0.
\end{aligned} \tag{29}$$

Now, consider the term

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \left(\frac{\|\mathbf{\Pi}_i(Y_i - Y_j)\|^2 - \|X_i - X_j\|^2}{h^2} \right) \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j.$$

It is equal to

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j \cdot \left(\frac{\|\boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j)\|^2 + 2(X_i - X_j)^T \boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j) - \|(\boldsymbol{I} - \boldsymbol{\Pi}_i)(X_i - X_j)\|^2}{h^2} \right).$$

First, note that

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \frac{\|(\boldsymbol{I} - \boldsymbol{\Pi}_i)(X_i - X_j)\|^2}{h^2} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j = 0. \tag{30}$$

Next, (13) implies

$$\|\boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j)\|^2 \leqslant \left(\frac{2M(\Delta h + b)}{\varkappa} + \frac{3M\|X_j - X_i\|}{\varkappa} \right)^2.$$

Then, using the inequality $\|\varepsilon_j\| \leqslant M$ (due to (A3)) and Lemma 9, we obtain

$$\begin{aligned} &\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j \frac{\|\boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j)\|^2}{h^2} \\ &\lesssim \frac{M^3(\Delta + b/h)^2 h^d}{\varkappa^2}. \end{aligned} \tag{31}$$

Finally, consider the expectation

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \cdot \frac{2(X_i - X_j)^T \boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j)}{h^2} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j.$$

Denote

$$v_{ij} = 2\mathbb{E}\left(\boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j) u^T \varepsilon_j \,|\, X_i, X_j, Y_i \right).$$

According to (13), the norm of the vector v_{ij} is bounded by

$$\|v_{ij}\| \leqslant 2M \|\boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j)\| \leqslant 2M \left(\frac{2M(\Delta h + b)}{\varkappa} + \frac{3M\|X_j - X_i\|}{\varkappa} \right).$$

On the event $\{\|X_i - X_j\| \leqslant r_d\}$, we have

$$\|v_{ij}\| \leqslant 2M \left(\frac{2M(\Delta h + b)}{\varkappa} + \frac{3Mr_d}{\varkappa} \right).$$

Applying Lemma 8 with the vector $u = v_{ij}/\|v_{ij}\|$, we obtain

$$\begin{aligned} &\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \cdot \frac{2(X_i - X_j)^T \boldsymbol{\Pi}_i(\varepsilon_i - \varepsilon_j)}{h^2} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j \\ &\lesssim \frac{Mh^d}{\varkappa} \left(\frac{M(\Delta h + b)}{\varkappa} + \frac{Mr_d}{\varkappa} \right) \\ &\lesssim \frac{Mh^d}{\varkappa} \left(\frac{M(\Delta h + b)}{\varkappa} + \frac{Mh\sqrt{\log h^{-1}}}{\varkappa} \right). \end{aligned} \tag{32}$$

Taking (30), (31) and (32) together, one obtains

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{h^2}} \left(\frac{\|\mathbf{\Pi}_i (Y_i - Y_j)\|^2 - \|X_i - X_j\|^2}{h^2} u^T \varepsilon_j \right) \tag{33}$$
$$\lesssim \frac{M^3 (\Delta + b/h)^2 h^d}{\varkappa^2} + \frac{M h^d}{\varkappa} \left(\frac{M(\Delta h + b)}{\varkappa} + \frac{M h \sqrt{\log h^{-1}}}{\varkappa} \right).$$

To complete the proof of Proposition 2, it remains to bound

$$\mathbb{E}^{(-i)} e^{-\frac{\|\zeta\|^2}{h^2}} \cdot \frac{\left(\|\mathbf{\Pi}_i (Y_i - Y_j)\|^2 - \|X_i - X_j\|^2 \right)^2}{2h^4} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j,$$

where $\zeta = \vartheta \mathbf{\Pi}_i (Y_j - Y_i) + (1 - \vartheta)(X_j - X_i)$ for some $\vartheta \in (0, 1)$. The same argument, as in the analysis of the vector ξ (see the proof of Proposition 1a, Inequality 25), yields

$$\|X_i - X_j\|^2 \leqslant 16\|\zeta\|^2 + 128 M^2 (\Delta + b/h)^2 h^2 / \varkappa^2 \leqslant 16\|\zeta\|^2 + 8h^2$$

and then

$$e^{-\frac{\|\zeta\|^2}{h^2}} \leqslant e^{1/2} e^{-\frac{\|X_i - X_j\|^2}{16h^2}}.$$

Due to (27),

$$\left(\|\mathbf{\Pi}_i (Y_i - Y_j)\|^2 - \|X_i - X_j\|^2 \right)^2 \leqslant \left[\left(\frac{2M(\Delta h + b)}{\varkappa} + \frac{3M\|X_j - X_i\|}{\varkappa} \right)^2 \right.$$
$$+ 2\|X_j - X_i\| \left(\frac{2M(\Delta h + b)}{\varkappa} + \frac{3M\|X_j - X_i\|}{\varkappa} \right)$$
$$\left. + \left(\frac{\Delta h \|X_j - X_i\|}{\varkappa} + \frac{\|X_j - X_i\|^2}{2\varkappa} \right)^2 \right]^2$$

Applying the inequality $(a + b + c)^2 \leqslant 3a^2 + 3b^2 + c^2$ and Lemma 9, we obtain

$$\mathbb{E}^{(-i)} e^{-\frac{\|\zeta\|^2}{h^2}} \cdot \frac{\left(\|\mathbf{\Pi}_i (Y_i - Y_j)\|^2 - \|X_i - X_j\|^2 \right)^2}{2h^4} \mathbb{1}(\|X_i - X_j\| \leqslant r_d) u^T \varepsilon_j \tag{34}$$
$$\lesssim M h^{d-4} \left(\frac{M^4 (\Delta + b/h)^4 h^4}{\varkappa^4} + \frac{M^4 h^4}{\varkappa^4} + \frac{M^2 (\Delta + b/h)^2 h^4}{\varkappa^2} + \frac{M^2 h^4}{\varkappa^2} + \frac{\Delta^4 h^8}{\varkappa^4} + \frac{h^8}{\varkappa^4} \right)$$
$$\lesssim \frac{M^5 (1 + \Delta + b/h)^4 h^d}{\varkappa^4} + \frac{M^3 (1 + \Delta + b/h)^2 h^d}{\varkappa^2} + \frac{(1 + \Delta^4) M h^{d+4}}{\varkappa^4}.$$

The assertion of Proposition 2c follows from Inequalities 29, 33, 34 and the fact that, due to conditions of Theorem 1,

$$\frac{M(1 + \Delta + b/h)}{\varkappa} \leqslant \frac{1}{16} + \frac{1}{4} = \frac{5}{16}.$$

Appendix F. Proofs Related to Theorem 3

F.1 Proof of Lemma 5

It is enough to consider the case $D = d+1$. For any $x_0 \in \mathbb{R}^{d+1}$ and $r > 0$, denote a sphere of radius r centered at x_0 by $\partial\mathcal{B}(x, r) = \{x \in \mathbb{R}^{d+1} : \|x - x_0\| = r\}$. Consider $\mathcal{M}_0 = \partial B(0, \varkappa)$. Let a random element X have a uniform distribution on $\mathcal{M}_0$. Clearly, $\mathcal{M}_0$ satisfies (A1) and the distribution of X satisfies (A2) with $L = 0$, $p_0 = p_1 = ((d+1)\omega_{d+1}\varkappa^d)^{-1}$, where ω_{d+1} is the volume of the Euclidean ball in $\mathbb{R}^{d+1}$ with radius 1. Given X, let ε have a uniform distribution on $\mathcal{T}_X\mathcal{M}_0 \cap \partial\mathcal{B}(X, Mb/\varkappa)$. Then

$$\mathbb{E}(\varepsilon \,|\, X) = 0, \quad \|\varepsilon\| = \frac{Mb}{\varkappa} \quad \mathbb{P}(\cdot \,|\, X)\text{-almost surely},$$

and the assumption (A3) is fulfilled. Consider $Y = X + \varepsilon$. Since X is orthogonal to ε by the construction, we have

$$\|Y\|^2 = \|X\|^2 + \|\varepsilon\|^2 = \varkappa^2 + \frac{M^2b^2}{\varkappa^2} \quad \text{almost surely}.$$

Consequently, the random element Y is supported on the sphere $\partial\mathcal{B}(0, R)$ with $R = \sqrt{\varkappa^2 + M^2b^2/\varkappa^2}$. By the spherical symmetry construction, Y has a uniform distribution on $\partial\mathcal{B}(0, R)$.

Let X' have a uniform distribution on $\mathcal{M}_1 = \partial\mathcal{B}(0, R)$ and, for any X', let $(\varepsilon' \,|\, X') = 0$ $\mathbb{P}(\cdot \,|\, X')$-almost surely. Then $X' + \varepsilon'$ and $X + \varepsilon$ have the same distribution. However,

$$d_H(\mathcal{M}_1, \mathcal{M}_0) = R - \varkappa = \varkappa\left(\sqrt{1 + \frac{M^2b^2}{\varkappa^4}} - 1\right) > \frac{M^2b^2}{3\varkappa^3}.$$

Here we used the fact that, due to the concavity of $\sqrt{1+x} - 1$, it holds that

$$\sqrt{1+x} - 1 \geqslant (\sqrt{2} - 1)x > \frac{x}{3}, \quad \forall x \in [0, 1].$$

Thus, for any estimate $\widehat{\mathcal{M}}$, we have

$$\sup_{\mathcal{M}^* \in \mathscr{M}_\varkappa^d} d_H(\widehat{\mathcal{M}}, \mathcal{M}) \geqslant \frac{1}{2} d_H(\mathcal{M}_1, \mathcal{M}_0) > \frac{M^2b^2}{6\varkappa^3}.$$

F.2 Proof of Lemma 6

Without loss of generality, we assume $D = d + 1$. We write a $(d+1)$-dimensional vector as (u, v), where $u \in \mathbb{R}^d$, $v \in \mathbb{R}$. Let $\mathcal{Z}^{(0)} \subset \mathbb{R}^D$ be a d-dimensional $\mathcal{C}^\infty$-manifold without a boundary with reach greater than 1 such that

$$\{(z, 0) : z \in \mathbb{R}^d, \|z\| \leqslant 1/2\} \subset \mathcal{Z}^{(0)}.$$

In (Aamari and Levrard, 2019), the authors claim that such a manifold can be constructed by flattering smoothly a unit sphere in $\mathbb{R}^D$. Let $\mathcal{Z} = 4\varkappa\mathcal{Z}^{(0)}$. Then $\mathcal{Z}$ is a $\mathcal{C}^\infty$-manifold without a boundary. Moreover, its reach is at least $4\varkappa$ and

$$\{(z, 0) : z \in \mathbb{R}^d, \|z\| \leqslant 2\varkappa\} \subset \mathcal{Z}.$$

We construct manifolds in the following way. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a smooth function, such that $\max_u \psi(u) = \psi(0) = 1$, $\psi(u) = 0$ for any $u \notin \mathcal{B}(0,1)$ and $\sup_u \|\nabla^2 \psi(u)\| \leqslant \Lambda$ for an absolute constant Λ. Let $(z_1, 0), \ldots, (z_N, 0)$, where $z_1, \ldots, z_N \in \mathbb{R}^d$, be a $2h$-packing of d-dimensional ball $\mathcal{Z} \cap B(0, \varkappa/2)$, $N = (4h/\varkappa)^{-d}$, $h < \varkappa/4$. For any $j \in \{1, \ldots, N\}$, introduce a manifold

$$\mathcal{M}_j = \left\{ \begin{pmatrix} z \\ 0 \end{pmatrix} + \frac{h^2}{\varkappa L} \psi\left(\frac{z - z_j}{h}\right) e_{d+1} : (z,0) \in \mathcal{Z} \cap \mathcal{B}(0,\varkappa) \right\} \cup \left(\mathcal{Z} \backslash \mathcal{B}(0,\varkappa)\right),$$

where the vector e_i is the i-th vector of the canonical basis in $\mathbb{R}^{d+1}$ with the components $e_i^{(j)} = \mathbb{1}(i = j)$. Let $\mathcal{M}_0$ be equal to $\mathcal{Z}$. Notice that $\mathcal{M}_j$, $j \in \{1, \ldots, N\}$ differs from $\mathcal{M}_0$ only on the set $\mathcal{B}((z_j, 0), h)$, and for any $k \neq j$ the balls $\mathcal{B}((z_j, 0), h)$ and $\mathcal{B}((z_k, 0), h)$ do not intersect. In other words, we consider a family of manifolds with a small bump in one of the points $(z_1, 0), \ldots, (z_N, 0)$.

Show that the family of manifolds $\mathscr{M}_\varkappa^\circ = \{\mathcal{M}_j : 1 \leqslant j \leqslant N\}$ with

$$h = c_0 \left(\frac{M^2 \varkappa^2 \log n}{n}\right)^{1/(d+4)},$$

where c_0 is a constant to be chosen later, is contained in the class $\mathscr{M}_\varkappa^d$ introduced in (A1). It is clear that $\mathcal{M}_j$ is a compact, connected, smooth d-dimensional manifold without a boundary for any j from 1 to N. The most important part is to check that the reach of $\mathcal{M}_j$ is not less than $\varkappa$. For this purpose, we use Theorem 4.18 from Federer (1959), which states that $\operatorname{reach}(\mathcal{M}) \geqslant \varkappa$ if and only if for any $x, x' \in \mathcal{M}$ it holds $d(x', \{x\} \oplus \mathcal{T}_x \mathcal{M}) \leqslant \|x - x'\|^2/(2\varkappa)$. Fix arbitrary $j \in \{1, \ldots, N\}$ and introduce

$$f_j(z) = \begin{pmatrix} z \\ 0 \end{pmatrix} + \frac{h^2}{L\varkappa} \psi\left(\frac{z - z_j}{h}\right) e_{d+1}, \quad z \in \mathcal{Z}.$$

Then for any $x \in \mathcal{M}_j \cap \mathcal{B}(0, \varkappa)$ the exists unique $(z, 0) \in \mathcal{Z}$, such that $x = f_j(z)$. By the construction, the inverse function to $f_j(z)$, $(z, 0) \in \mathcal{Z} \cap \mathcal{B}(0, \varkappa)$, is given by

$$f_j^{-1}(x) = \left(x^{(1)}, \ldots, x^{(d)}\right)^T,$$

where $x \in \mathcal{M}_j \cap \mathcal{B}(0, \varkappa)$ and $x^{(j)}$ is the j-th component of the vector x. Moreover, the unit normal to $\mathcal{M}_j$ at the point $x = f_j(z)$ is given by

$$\nu_j(z) = C_h^{-1} \left(-\frac{h}{\varkappa L} \nabla \psi\left(\frac{z - z_j}{h}\right)^T, 1\right)^T, \tag{35}$$

where

$$C_h = \sqrt{1 + \left(\frac{h}{\varkappa L}\right)^2 \left\|\nabla \psi\left(\frac{z - z_j}{h}\right)\right\|^2}.$$

Fix arbitrary $x = f_j(z), x_0 = f_j(z_0) \in \mathcal{M}_j$ and check that

$$\left|\nu_j(z_0)^T (x - x_0)\right| = \left|\nu_j(z_0)^T (f_j(z) - f_j(z_0))\right| \leqslant \frac{\|z - z_0\|^2}{2\varkappa} \leqslant \frac{\|x - x_0\|^2}{2\varkappa}.$$

The last inequality is obvious, since $(z - z_0)$ is a subvector of $(x - x_0)$. It remains to check the second inequality. It holds that

$$\begin{aligned}
&\left|\nu_j(z_0)^T(f_j(z) - f_j(z_0))\right| \\
&= C_h^{-1}\left|-\frac{h}{\varkappa\Lambda}\nabla\psi^T\left(\frac{z_0 - z_j}{h}\right)(z - z_0) + \frac{h^2}{\varkappa\Lambda}\left(\psi\left(\frac{z - z_j}{h}\right) - \psi\left(\frac{z_0 - z_j}{h}\right)\right)\right| \\
&\leqslant \left|-\frac{h}{\varkappa\Lambda}\nabla\psi^T\left(\frac{z_0 - z_j}{h}\right)(z - z_0) + \frac{h^2}{\varkappa\Lambda}\left(\psi\left(\frac{z - z_j}{h}\right) - \psi\left(\frac{z_0 - z_j}{h}\right)\right)\right| \\
&\leqslant \frac{h^2}{\varkappa\Lambda}\cdot\frac{\Lambda\|z - z_0\|^2}{2h^2} = \frac{\|z - z_0\|^2}{2\varkappa}.
\end{aligned}$$

Here we used Taylor's expansion of ψ up to the second order and the fact that $\|\nabla^2\psi\| \leqslant \Lambda$.

Now, we are going to describe distributions of X and ε in the model (1). Let a random element Z have a uniform distribution on $\mathcal{Z}$. For any fixed $j \in \{1, \ldots, N\}$, we take $X = f_j(Z)$, where

$$f_j(z) = \begin{pmatrix} z \\ 0 \end{pmatrix} + \frac{h^2}{\varkappa\Lambda}\psi\left(\frac{z - z_j}{h}\right)e_{d+1}, \quad z \in \mathcal{Z}.$$

Denote a volume of the set $\mathcal{Z}$ by $V_{\mathcal{Z}}$. Then for any x, such that $f_j^{-1}(x) \notin \mathcal{B}(z_j, h)$, the density $p_j(x)$ is just $V_{\mathcal{Z}}^{-1}$. Otherwise, the density $p_j(x)$ of X is defined by the formula

$$\begin{aligned}
p_j(x) &= \frac{1}{V_{\mathcal{Z}}}\left(\det\nabla f_j(f_j^{-1}(x))\right)^{-1} \\
&= \frac{1}{V_{\mathcal{Z}}}\left(\det\left(I + \frac{h}{\varkappa\Lambda}\nabla\psi\left(\frac{f_j^{-1}(x) - z_j}{h}\right)e_{d+1}^T\right)\right)^{-1}.
\end{aligned}$$

Since for any two vectors $u, v \in \mathbb{R}^{d+1}$ it holds $\det(I + uv^T) = 1 + u^Tv$, we have

$$p_j(x) = \frac{1}{V_{\mathcal{Z}}}\left(1 + \frac{h}{\varkappa\Lambda}e_{d+1}^T\nabla\psi\left(\frac{f_j^{-1}(x) - z_j}{h}\right)\right)^{-1}. \tag{36}$$

Note that $\nabla\psi(0) = 0$ by construction. Taking into account that $\sup_u\|\nabla^2\psi(u)\| \leqslant \Lambda$, we conclude that

$$\left\|\nabla\psi\left(\frac{f_j^{-1}(x) - z_j}{h}\right)\right\| \leqslant \frac{\Lambda\|f_j^{-1}(x) - z_j\|}{h} \leqslant \Lambda \quad \forall\, x : f_j^{-1}(x) \in \mathcal{B}(z_j, h).$$

This and the fact that $h < \varkappa/4$ yield

$$p_0 = \frac{4}{5V_{\mathcal{Z}}} \leqslant \frac{1}{V_{\mathcal{Z}}\left(1 + \frac{h}{\varkappa}\right)} \leqslant p_j(x) \leqslant \frac{1}{V_{\mathcal{Z}}\left(1 - \frac{h}{\varkappa}\right)} \leqslant \frac{4}{3V_{\mathcal{Z}}} = p_1.$$

Thus, the density of X is bounded from above and below by p_1 and p_0 respectively.

Show that $p_j(x)$ has $4p_1/(3\varkappa)$-Lipschitz derivative. Differentiating (36), we obtain

$$\|\nabla p_j(x)\| = \frac{V_{\mathcal{Z}}p_j^2(x)}{\varkappa\Lambda}\left\|\boldsymbol{I}_{d,d+1}\nabla^2\psi\left(\frac{f_j^{-1}(x) - z_j}{h}\right)e_{d+1}\right\| \leqslant \frac{V_{\mathcal{Z}}p_1^2}{\varkappa} \leqslant \frac{4p_1}{3\varkappa},$$

where $\boldsymbol{I}_{d,d+1} \in \mathbb{R}^{d\times(d+1)}$ is the matrix of the first d rows of the identity matrix $\boldsymbol{I}_{d+1}$. Thus, for each j, the density $q_j(x)$ fulfils (A2).

Next, we describe the conditional distribution of Y given X. We generate Y from the model

$$Y = X + \xi e_{d+1} \mathbb{1}(X \in \mathcal{B}(0,\varkappa)), \quad X \in \mathcal{M}_j, \tag{37}$$

where $\mathbb{P}(\xi = 0.5M - X^{(d+1)}|X) = \eta(X), \mathbb{P}(\xi = -0.5M - X^{(d+1)}) = 1-\eta(X)$, $\eta = \eta(X) = 1/2 + X^{(d+1)}/M$, and $X^{(d+1)}$ is the $(d+1)$-th component of X. Note that Y belongs either to the set $\mathcal{M}^+ = \{(z, 0.5M) : (z,0) \in \mathcal{B}(0,\varkappa) \subset \mathbb{R}^{d+1}\}$, or to the set $\mathcal{M}^- = \{(z, 0.5M) : (z,0) \in \mathcal{B}(0,\varkappa) \subset \mathbb{R}^{d+1}\}$, or to the set $\mathcal{Z}\backslash B(0, \mathcal{B}(0,\varkappa))$. It remains to check the condition (A3). Take

$$h = c_0 \left(\frac{M^2\varkappa^2 \log n}{n}\right)^{1/(d+4)},$$

where c_0 is such that the condition $h^2/(M\varkappa\Lambda) \leqslant 1/2$ is fulfilled. Such c_0 exists since $M \gtrsim (\log n/n)^{2/d}$. First, note that the noise magnitude is not greater than $0.5M + h^2/(\varkappa\Lambda)$, which is less than M. Second, using the expression (35) of the unit normal to $\mathcal{M}_j$ at the point $x = f_j(z)$, we have

$$\begin{aligned}
&\|\xi \boldsymbol{\Pi}(f_j(z)) e_{d+1}\|^2 = |\xi|^2 - |\xi e_{d+1}^T \nu_j(z)|^2 \\
&= |\xi|^2 - \frac{|\xi|^2}{1 + h^2\|\nabla\psi(z/h)\|^2/(\varkappa\Lambda)^2} \\
&= \frac{|\xi|^2 h^2 \|\nabla\psi(z/h)\|^2}{\varkappa^2\Lambda^2 + h^2\|\nabla\psi\left(\frac{z-z_j}{h}\right)\|^2} \leqslant \Lambda^2|\xi|^2 \cdot \frac{h^2}{\varkappa^2\Lambda^2} \leqslant \frac{M^2h^2}{\varkappa^2},
\end{aligned}$$

and (A3) holds with

$$b = c_0 \left(\frac{M^2\varkappa^2 \log n}{n}\right)^{1/(d+4)}.$$

Here we used the fact that for any $u \in \mathcal{B}(0,1)$

$$\|\nabla\psi(u)\| = \|\nabla\psi(u) - \nabla\psi(0)\| \leqslant \max_{u' \in \mathcal{B}(0,1)} \|\nabla^2\psi(u')\|\|u\| \leqslant L\,.$$

We use (Tsybakov, 2009, Theorem 2.5) to prove the lower bound in Theorem 3. Let P_j, $0 \leqslant j \leqslant N$, be the probability measure, generated by $Y = X + \varepsilon$, where $X \in \mathcal{M}_j$. Then P_0 is a dominating measure, i.e. $P_j \ll P_0$ for all j from 1 to N, and for any $j \neq k$ we have

$$d_H(\mathcal{M}_j, \mathcal{M}_k) \geqslant \frac{h^2}{\varkappa L}.$$

Prove that, for sufficiently large n, it holds

$$\frac{1}{N}\sum_{j=1}^{N} \mathcal{KL}(P_j^{\otimes n}, P_0^{\otimes n}) \leqslant \alpha \log N, \tag{38}$$

where $\mathcal{KL}(P, Q)$ is the Kullback-Leibler divergence between P and Q, α is a constant from the interval $(0, 1/8)$. Then Theorem 2.5 in Tsybakov (2009) yields

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

It remains to check (38). For any j from 1 to N, it holds

$$\mathcal{KL}(P_j^{\otimes n}, P_0^{\otimes n}) = n \int \log \frac{dP_j(y)}{dP_0(y)} dP_j(y).$$

The density of P_j with respect to the Hausdorff measure on $(\mathcal{Z} \backslash B(0, \varkappa)) \cup \mathcal{M}^+ \cup \mathcal{M}^-$ is given by the formula

$$\log p_j(y) = \begin{cases} \log\left(1 + \frac{2h^2}{M\varkappa\Lambda} \psi\left(\frac{\widetilde{y} - z_j}{h}\right)\right) - \log(2V_{\mathcal{Z}}), & y = (\widetilde{y}, 0.5M) \in \mathcal{M}^+, \\ \log\left(1 - \frac{2h^2}{M\varkappa\Lambda} \psi\left(\frac{\widetilde{y} - z_j}{h}\right)\right) - \log(2V_{\mathcal{Z}}), & y = (\widetilde{y}, -0.5M) \in \mathcal{M}^-, \\ -\log(V_{\mathcal{Z}}), & y \in \mathcal{Z} \backslash \mathcal{B}(0, \varkappa). \end{cases}$$

The density of P_0 with respect to the same measure is just

$$\log p_0(y) = \begin{cases} -\log(2V_{\mathcal{Z}}), & y = (\widetilde{y}, \pm M) \in \mathcal{M}^+ \cup \mathcal{M}^-, \\ -\log(V_{\mathcal{Z}}), & y \in \mathcal{Z} \backslash \mathcal{B}(0, \varkappa). \end{cases}$$

Then

$$\begin{aligned} &\mathcal{KL}(P_j^{\otimes n}, P_0^{\otimes n}) \\ &= \frac{n}{2V_{\mathcal{Z}}} \int\limits_{\|z - z_j\| \leqslant h} \log\left(1 + \frac{2h^2}{M\varkappa\Lambda} \psi\left(\frac{z - z_j}{h}\right)\right) \left(1 + \frac{2h^2}{M\varkappa\Lambda} \psi\left(\frac{z - z_j}{h}\right)\right) dz \\ &+ \frac{n}{2V_{\mathcal{Z}}} \int\limits_{\|z - z_j\| \leqslant h} \log\left(1 - \frac{2h^2}{M\varkappa\Lambda} \psi\left(\frac{z - z_j}{h}\right)\right) \left(1 - \frac{2h^2}{M\varkappa\Lambda} \psi\left(\frac{z - z_j}{h}\right)\right) dz. \end{aligned}$$

Using the inequality

$$(1 + t) \log(1 + t) \leqslant t + \frac{t^2}{2}, \quad t \in (-1, 1),$$

we obtain that

$$(1 + t) \log(1 + t) + (1 - t) \log(1 - t) \leqslant t^2, \quad t \in (-1, 1).$$

Then, since $2h^2 < M\varkappa\Lambda$, we have

$$\mathcal{KL}(P_j^{\otimes n}, P_0^{\otimes n}) \leqslant \frac{n}{2V_{\mathcal{Z}}} \int\limits_{\|z\| \leqslant h} \frac{4h^4}{M^2 \varkappa^2 \Lambda^2} \psi^2\left(\frac{z}{h}\right) dz = \frac{C_{\Lambda, \mathcal{Z}} n h^{d+4}}{M^2 \varkappa^2},$$

where $C_{\Lambda,\mathcal{Z}} > 0$ is a constant, depending on $\mathcal{Z}$ and Λ. Substituting h by $c_0 \left(\frac{M^2\varkappa^2 \log n}{n}\right)^{1/(d+4)}$, we obtain

$$\frac{C_{\Lambda,\mathcal{Z}} n h^{d+4}}{M^2\varkappa^2} = c_0^{d+4} C_{\Lambda,\mathcal{Z}} \log n.$$

On the other hand,

$$\begin{aligned}\log N = \log\left(\frac{4h}{\varkappa}\right)^{-d} &= d\log\frac{\varkappa}{4} - d\log c_0 + \frac{d}{d+4}\log\frac{n}{M^2\varkappa^2\log n} \\ &\geqslant -d\log c_0 + \frac{d}{2(d+4)}\log n,\end{aligned}$$

where in the last inequality we assumed that n is large, so it holds $M^2\varkappa^2 \log n \leqslant n^{1/4}$ and $\varkappa \geqslant 4n^{-1/(4d+16)}$. Choose any constant $c_0 > 0$, satisfying the inequality

$$8c_0^{d+4} C_{L,\mathcal{Z}} \log n + d\log c_0 < \frac{d}{2(d+4)}\log n.$$

Such constant always exists. Thus, (38) is fulfilled, and (Tsybakov, 2009, Theorem 2.5) yields the claim of Theorem 3.

Appendix G. Auxiliary Results

This section contains some auxiliary results, which are used in the proofs. The results below often use technique concerning integration over manifolds. Therefore, we would like to start with a short background, which will help a reader follow the proofs.

Given $x \in \mathcal{M}$ and $v \in \mathcal{T}_x\mathcal{M}$, let $\gamma(t, x, v)$ be a geodesic starting at x, such that

$$\left.\frac{d\gamma(t,x,v)}{dt}\right|_{t=0} = v.$$

The exponential map of $\mathcal{M}$ at the point x $\mathcal{E}_x : \mathcal{T}_x\mathcal{M} \to \mathcal{M}$ is defined as $\mathcal{E}_x(p) = \gamma(1, x, v)$. Note that $d_{\mathcal{M}}(\mathcal{E}_x(v), x) = \|v\|$ for $v \leqslant \varkappa/4$, where $d_{\mathcal{M}}(x, x')$ is the length of the shortest path on $\mathcal{M}$ between x and x'. We extensively use integration over manifolds. In these cases, the exponential map is useful to apply the change of variables formula. For a small open set U, $x \in U \subset \mathcal{M}$, it holds

$$\int\limits_U f(x)dW(x) = \int\limits_{\mathcal{E}_x^{-1}(U)} f(\mathcal{E}_x(p))\sqrt{\det g(p)}dp,$$

where $g(p)$ is the metric tensor. Without going deep into details, we just mention that the metric tensor allows a nice decomposition (see, for instance, (Trillos et al., 2019, Equation 2.1)), which is enough for our purposes:

$$\left|\sqrt{\det g(v)} - 1\right| \lesssim \frac{d\|v\|^2}{\varkappa^2}.$$

Proposition 3 *Let $\mathcal{M} \in \mathscr{M}_{\varkappa}^{d}$ and $x, x' \in \mathcal{M}$, $\|x - x'\| \leqslant 2\varkappa$. Let $\mathbf{\Pi}(x)$ and $\mathbf{\Pi}(x')$ be the projectors onto the tangent spaces $\mathcal{T}_x\mathcal{M}$ and $\mathcal{T}_{x'}\mathcal{M}$ respectively. Then the following inequalities hold:*

$$\|x - x'\| \leqslant d_{\mathcal{M}}(x, x') \leqslant 2\|x - x'\|, \tag{a}$$

$$\|\mathbf{\Pi}(x) - \mathbf{\Pi}(x')\| \leqslant \frac{\|x - x'\|}{\varkappa}. \tag{b}$$

Proof Proposition 3a follows from (Boissonnat et al., 2019, Lemma 2.5) and the inequality

$$\arcsin \frac{2t}{\pi} \leqslant t, \quad \forall\, t \in (0, \pi/2).$$

To prove Proposition 3b, we use (Golub and Van Loan, 2013, Theorem 2.5.1):

$$\|\mathbf{\Pi}(x) - \mathbf{\Pi}(x')\| = \|(\boldsymbol{I} - \mathbf{\Pi}(x))\mathbf{\Pi}(x')\| = \sin \angle(\mathcal{T}_x\mathcal{M}, \mathcal{T}_{x'}\mathcal{M}).$$

Then the claim of the proposition follows from (Boissonnat et al., 2019, Corollary 3.6):

$$\|\mathbf{\Pi}(x) - \mathbf{\Pi}(x')\| = \sin \angle(\mathcal{T}_x\mathcal{M}, \mathcal{T}_{x'}\mathcal{M}) \leqslant 2 \sin \frac{\angle(\mathcal{T}_x\mathcal{M}, \mathcal{T}_{x'}\mathcal{M})}{2} \leqslant \frac{\|x - x'\|}{\varkappa}.$$

■

Lemma 7 *Fix any i from 1 to n. There are absolute constants c and C', such that, with probability at least $1 - n^{-2}$, it holds*

$$\begin{aligned}
&\lambda_d(\widehat{\mathbf{\Xi}}_i) - \lambda_{d+1}(\widehat{\mathbf{\Xi}}_i) \\
&\geqslant \frac{c}{4}\left(1 - \frac{2}{c(\gamma - 4\beta_1)^d} - \sqrt{\frac{6C'}{c^2(\gamma - 4\beta_1)^d}}\right)(\gamma - 4\beta_1)^{d+2} n h_k^{d+2} \\
&- 9C' n^{-2/d}(\gamma + 4\beta_1)^{d+2} n h_k^{d+2} - 16C'\beta_1^2(\gamma + 4\beta_1)^d n h_k^{d+2} \\
&- \frac{C'(\gamma + 4\beta_1)^{d+4} n h_k^{d+4}}{\varkappa^2}.
\end{aligned}$$

Proof

Now, consider the matrix $\widehat{\mathbf{\Xi}}_i^{(k)}$. It is clear that all the eigenvectors of $\widehat{\mathbf{\Xi}}_i^{(k)}$ belong to the linear space $\mathcal{T}_{X_i}\mathcal{M}^*$. Thus, $\widehat{\mathbf{\Xi}}_i^{(k)}$ has at most d non-zero eigenvalues. In what follows, we show that the d-th largest eigenvalue of $\widehat{\mathbf{\Xi}}_i^{(k)}$ is non-zero and give a lower bound on the spectral gap $\lambda_d(\widehat{\mathbf{\Xi}}_i^{(k)}) - \lambda_{d+1}(\widehat{\mathbf{\Xi}}_i^{(k)})$. It holds that

$$\lambda_d(\widehat{\mathbf{\Xi}}_i^{(k)}) - \lambda_{d+1}(\widehat{\mathbf{\Xi}}_i^{(k)}) = \min_{u \in \mathcal{T}_{X_i}\mathcal{M}^*, \|u\| = 1} \sum_{j=1}^{n} v_{ij}(u^T(\widehat{Z}_{ij}^{(k)} - \widehat{Z}_{ii}^{(k)}))^2\,.$$

Using the inequality

$$\begin{aligned}
\|\widehat{Z}_{ij}^{(k)} - X_j\| &\leqslant \|\widehat{Z}_{ij}^{(k)} - Z_{ij}\| + \|X_j - Z_{ij}\| \\
&\leqslant \|\widehat{X}_j^{(k)} - X_j\| + \|X_j - Z_{ij}\| \leqslant 2\beta_1 h_k + \frac{\|X_j - X_i\|^2}{2\varkappa},
\end{aligned}$$

we obtain that for any u it holds

$$(u^T(\widehat{Z}_{ij}^{(k)} - \widehat{Z}_{ii}^{(k)}))^2 \geqslant \frac{1}{2}(u^T(X_j - X_i))^2 - 8\beta_1^2 h_k^2 - \frac{\|X_j - X_i\|^4}{2\varkappa^2},$$

which yields

$$\begin{aligned}
\lambda_d(\widehat{\boldsymbol{\Xi}}_i^{(k)}) - \lambda_{d+1}(\widehat{\boldsymbol{\Xi}}_i^{(k)}) &\geqslant \frac{1}{2} \min_{u \in \mathcal{B}(0,1) \cap \mathcal{T}_{X_i}\mathcal{M}^*} \sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2 \\
&\quad - 8\beta_1^2 h_k^2 \sum_{j=1}^n v_{ij} - \sum_{j=1}^n v_{ij} \frac{\|X_j - X_i\|^4}{2\varkappa^2} \qquad (39) \\
&\geqslant \frac{1}{2} \min_{u \in \mathcal{B}(0,1) \cap \mathcal{T}_{X_i}\mathcal{M}^*} \sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2 \\
&\quad - 16C'\beta_1^2(\gamma + 4\beta_1)^d n h_k^{d+2} - \frac{C'(\gamma + 4\beta_1)^{d+4} n h_k^{d+4}}{\varkappa^2}.
\end{aligned}$$

In the last inequality we used (18) and the fact that $\|X_i - X_j\| \leqslant (\gamma + 4\beta_1)h_k$, if $\|\widehat{X}_i^{(k)} - \widehat{X}_j^{(K)}\| \leqslant \gamma h_k$.

It remains to provide a lower bound for the sum

$$\min_{u \in \mathcal{B}(0,1) \cap \mathcal{T}_{X_i}\mathcal{M}^*} \sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2.$$

Let $\mathcal{N}_\varepsilon$ stand for a ε-net of the set $\mathcal{B}(0,1) \cap \mathcal{T}_{X_i}\mathcal{M}^*$. It is known that $|\mathcal{N}_\varepsilon| \leqslant (3/\varepsilon)^d$. Here and further in this proof we will assume $\varepsilon = \varepsilon_n = 3n^{-1/d}$. Then, for any $t > 0$, it holds

$$\begin{aligned}
&\left\{ \min_{u \in \mathcal{B}(0,1) \cap \mathcal{T}_{X_i}\mathcal{M}^*} \sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2 < t \right\} \\
&\subseteq \left\{ \min_{u \in \mathcal{N}_\varepsilon} \sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2 < 2t + 2\varepsilon^2 \sum_{j=1}^n v_{ij}\|X_i - X_j\|^2 \right\} \qquad (40) \\
&\subseteq \bigcup_{u \in \mathcal{N}_\varepsilon} \left\{ \sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2 < 2t + 4C'\varepsilon^2(\gamma + 4\beta_1)^{d+2} n h_k^{d+2} \right\}.
\end{aligned}$$

Fix any $u \in \mathcal{N}_\varepsilon$ and consider

$$\mathbb{P}^{(-i)} \left(\sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2 < 2t + 4C'\varepsilon^2(\gamma + 4\beta_1)^{d+2} n h_k^{d+2} \right).$$

Note that

$$\begin{aligned}
&\sum_{j=1}^n v_{ij}(u^T(X_j - X_i))^2 \\
&\geqslant \sum_{j=1}^n \mathbb{1}\left(\|X_i - X_j\| \leqslant (\gamma - 4\beta_1)h_k\right)(u^T(X_j - X_i))^2,
\end{aligned}$$

so it holds

$$\mathbb{P}^{(-i)}\left(\sum_{j=1}^{n} v_{ij}(u^T(X_j - X_i))^2 < 2t + 4C'\varepsilon^2(\gamma + 4\beta_1)^{d+2} n h_k^{d+2}\right)$$
$$\leqslant \mathbb{P}^{(-i)}\Bigg(\sum_{j=1}^{n} \mathbb{1}\left(\|X_i - X_j\| \leqslant (\gamma - 4\beta_1)h_k\right)(u^T(X_j - X_i))^2$$
$$< 2t + 4C'\varepsilon^2(\gamma + 4\beta_1)^{d+2} n h_k^{d+2}\Bigg).$$

Given X_i, the random variables $\mathbb{1}\left(\|X_i - X_j\| \leqslant h_k\right)(u^T(X_j - X_i))^2$, $1 \leqslant j \leqslant n$, are conditionally independent and identically distributed, and expectation of each of them can be bounded below by

$$\mathbb{E}^{(-i)}\mathbb{1}\left(\|X_i - X_1\| \leqslant (\gamma - 4\beta_1)h_k\right)(u^T(X_1 - X_i))^2$$
$$\geqslant \frac{p_0}{4} \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, h_k) \cap \{|u^T(X_i - x)| \geqslant \frac{1}{2}\}} \|x - X_i\|^2 dW(x) \geqslant c(\gamma - 4\beta_1)^{d+2} h_k^{d+2}.$$

At the same time, the variance of these random variables does not exceed

$$\mathbb{E}^{(-i)}\mathbb{1}\left(\|X_i - X_1\| \leqslant (\gamma - 4\beta_1)h_k\right)(u^T(X_1 - X_i))^4$$
$$\leqslant (\gamma - 4\beta_1)^4 h_k^4 \mathbb{P}^{(-i)}\left(\|X_i - X_1\| \leqslant (\gamma - 4\beta_1)h_k\right)$$
$$\leqslant C(\gamma - 4\beta_1)^{d+4} h_k^{d+4} \leqslant C'(\gamma - 4\beta_1)^{d+4} h_k^{d+4},$$

where $C' = C \vee 16/3$. Again, using the Bernstein's inequality, we obtain that for any $\tilde{t}$ it holds

$$\mathbb{P}^{(-i)}\Bigg(\sum_{j=1}^{n} \mathbb{1}\left(\|X_i - X_j\| \leqslant (\gamma - 4\beta_1)h_k\right)(u^T(X_j - X_i))^2$$
$$< n\mathbb{E}^{(-i)}\mathbb{1}\left(\|X_i - X_j\| \leqslant (\gamma - 4\beta_1)h_k\right)(u^T(X_j - X_i))^2 - \tilde{t}\Bigg)$$
$$\leqslant \exp\left\{-\frac{\tilde{t}^2}{2nC'(\gamma - 4\beta_1)^{d+4} h_k^{d+4} + 2(\gamma - 4\beta_1)^2 h_k^2 \tilde{t}/3}\right\}.$$

Take $\tilde{t} = \delta n \mathbb{E}^{(-i)}\mathbb{1}\left(\|X_i - X_j\| \leqslant (\gamma - 4\beta_1)h_k\right)(u^T(X_j - X_i))^2$. Then

$$\exp\left\{-\frac{\tilde{t}^2}{2nC'(\gamma - 4\beta_1)^{d+4} h_k^{d+4} + 2(\gamma - 4\beta_1)^2 h_k^2 \tilde{t}/3}\right\}$$
$$\leqslant \exp\left\{-\frac{c^2 n^2 \delta^2 (\gamma - 4\beta_1)^{2d+4} h_k^{2d+4}}{2nC'(\gamma - 4\beta_1)^{d+4} h_k^{d+4} + \frac{2cn\delta}{3}(\gamma - 4\beta_1)^{d+4} h_k^{d+4}}\right\}$$
$$= \exp\left\{-\frac{nc^2\delta^2(\gamma - 4\beta_1)^d h_k^d}{2C' + \frac{2c\delta}{3}}\right\}.$$

Choose δ satisfying the inequality

$$\frac{c^2\delta^2(\gamma-4\beta_1)^d}{2C'+\dfrac{2c\delta}{3}} \geqslant 3\,.$$

In particular,

$$\delta = \frac{2}{c(\gamma-4\beta_1)^d} + \sqrt{\frac{6C'}{c^2(\gamma-4\beta_1)^d}}$$

is a suitable choice. Then

$$\begin{aligned}
&\exp\left\{-\frac{nc^2\delta^2(\gamma-4\beta_1)^d h_k^d}{2C'+\frac{2c\delta}{3}}\right\} \leqslant e^{-3nh_k^d} \\
&\leqslant e^{-3\log n} \leqslant e^{-2\log n - \log|\mathcal{N}_\varepsilon|} = \frac{1}{|\mathcal{N}_\varepsilon| n^2}\,.
\end{aligned}$$

Thus, with probability at least $1-(|\mathcal{N}_\varepsilon|n^2)^{-1}$, it holds

$$\begin{aligned}
&\sum_{j=1}^n \mathbb{1}\left(\|X_i - X_j\| \leqslant (\gamma-4\beta_1)h_k\right)(u^T(X_j-X_i))^2 \\
&\geqslant \left(1 - \frac{2}{c(\gamma-4\beta_1)^d} - \sqrt{\frac{6C'}{c^2(\gamma-4\beta_1)^d}}\right) c(\gamma-4\beta_1)^{d+2}h_k^{d+2}\,.
\end{aligned}$$

By the union bound, on an event with probability at least $1-n^{-2}$ it holds

$$\begin{aligned}
&\min_{u\in\mathcal{N}_\varepsilon}\sum_{j=1}^n \mathbb{1}\left(\|X_i - X_j\| \leqslant (\gamma-4\beta_1)h_k\right)(u^T(X_j-X_i))^2 \\
&\geqslant \left(1 - \frac{2}{c(\gamma-4\beta_1)^d} - \sqrt{\frac{6C'}{c^2(\gamma-4\beta_1)^d}}\right) cn(\gamma-4\beta_1)^{d+2}h_k^{d+2}
\end{aligned} \tag{41}$$

Then, due to (40) and (41), on this event

$$\begin{aligned}
&\min_{u\in\mathcal{B}(0,1)\cap\mathcal{T}_{X_i}\mathcal{M}^*}\sum_{j=1}^n v_{ij}(u^T(X_j-X_i))^2 \\
&\geqslant \frac{c}{2}\left(1 - \frac{2}{c(\gamma-4\beta_1)^d} - \sqrt{\frac{6C'}{c^2(\gamma-4\beta_1)^d}}\right)(\gamma-4\beta_1)^{d+2}nh_k^{d+2} \\
&- 2C'\varepsilon^2(\gamma+4\beta_1)^{d+2}nh_k^{d+2},
\end{aligned}$$

and, together with (39), this yields

$$\begin{aligned}
&\lambda_d(\widehat{\mathbf{\Xi}}_i) - \lambda_{d+1}(\widehat{\mathbf{\Xi}}_i) \\
&\geqslant \frac{c}{4}\left(1 - \frac{2}{c(\gamma-4\beta_1)^d} - \sqrt{\frac{6C'}{c^2(\gamma-4\beta_1)^d}}\right)(\gamma-4\beta_1)^{d+2} n h_k^{d+2} \\
&- C'\varepsilon^2(\gamma+4\beta_1)^{d+2} n h_k^{d+2} - 16C'\beta_1^2(\gamma+4\beta_1)^d n h_k^{d+2} \\
&- \frac{C'(\gamma+4\beta_1)^{d+4} n h_k^{d+4}}{\varkappa^2}.
\end{aligned}$$

The choice $\varepsilon = 3n^{-1/d}$ yields the claim of Lemma 7. ■

Lemma 8 *Let $\mathbb{E}^{(-i)}$ denote the conditional expectation $\mathbb{E}(\cdot|(X_i, Y_i))$ and let $r_d = 4h\sqrt{(d+2)\log h^{-1}}$. Then, for any i from 1 to n, it holds*

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{h^2}} u^T (X_j - X_i) \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right) \lesssim \frac{d h^{d+2}}{\varkappa}.$$

Proof

We have

$$\begin{aligned}
&\mathbb{E}^{(-i)} e^{-\frac{\|X_j - X_i\|^2}{h^2}} u^T (X_j - X_i) \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right) \\
&= \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|x - X_i\|^2}{h^2}} u^T (x - X_i) p(x) dW(x).
\end{aligned}$$

Due to (A2), the last expression does not exceed

$$\begin{aligned}
&\leqslant p(X_i) \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|x - X_i\|^2}{h^2}} u^T (x - X_i) dW(x) \\
&+ \frac{L}{\varkappa} \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|x - X_i\|^2}{h^2}} \|x - X_i\|^2 dW(x).
\end{aligned}$$

Due to Lemma 9,

$$\frac{L}{\varkappa} \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|x - X_i\|^2}{h^2}} \|x - X_i\|^2 dW(x) \lesssim \frac{h^{d+2}}{\varkappa},$$

so it remains to prove that

$$p(X_i) \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|x - X_i\|^2}{h^2}} u^T (x - X_i) dW(x) \lesssim \frac{h^{d+2}}{\varkappa}.$$

Let $\mathcal{E}_{X_i}(\cdot)$ be the exponential map of $\mathcal{M}^*$ at X_i and denote $\widetilde{\mathcal{B}}(X_i, r_d) = \mathcal{E}^{-1}(\mathcal{M}^* \cap \mathcal{B}(X_i, r_d))$. Note that $\mathcal{E}_{X_i}(\cdot)$ is a bijection on $\widetilde{\mathcal{B}}(X_i, r_d)$ (see, for instance, (Aamari and Levrard, 2019, Lemma 1)), because $r_d \leqslant \varkappa/4$. Then

$$\begin{aligned}
&\int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|x - X_i\|^2}{h^2}} u^T (x - X_i) dW(x) \\
&= \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\|^2}{h^2}} u^T (\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)) \sqrt{\det g(v)} dv.
\end{aligned}$$

Introduce functions

$$\psi_{X_i}(v) = \mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0) - v$$

and

$$\varphi_{X_i}(v) = \|\psi_{X_i}(v)\|^2 + 2v^T \psi_{X_i}(v).$$

Due to (Aamari and Levrard, 2019, Lemma 1), it holds that

$$\|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0) - v\| = \|\psi_{X_i}(v)\| \leqslant \frac{5\|p\|^2}{4\varkappa}, \tag{42}$$

which yields $\psi_{X_i}(v) = O\left(\frac{\|v\|^2}{\varkappa}\right)$, $\varphi_{X_i}(v) = O\left(\frac{\|v\|^3}{\varkappa}\right)$. Now, consider $\sqrt{\det g(v)}$. It is known (see, for instance, (Trillos et al., 2019, Equation 2.1)) that there exists an absolute constant $\overline{C}$, such that

$$\left|\sqrt{\det g(v)} - 1\right| \leqslant \frac{\overline{C} d \|v\|^2}{\varkappa^2}. \tag{43}$$

Taking (42) and (43) into account, we obtain

$$\begin{aligned}
&\int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\|^2}{h^2}} u^T (\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)) \sqrt{\det g(v)} dv \\
&= \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2 + \varphi_{X_i}(v)}{h^2}} u^T (v + \psi_{X_i}(v)) \sqrt{\det g(v)} dv \\
&= \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} \left(1 + O\left(\frac{\|v\|^3}{h^2 \varkappa}\right)\right) \left(u^T v + O\left(\frac{\|v\|^2}{\varkappa}\right)\right) \left(1 + O\left(\frac{\|v\|^2}{\varkappa^2}\right)\right) dv \\
&= \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} u^T v dv + \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} O\left(\frac{\|v\|^4}{h^2 \varkappa}\right) dv \\
&+ \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} O\left(\frac{\|v\|^2}{\varkappa}\right) dv + \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} O\left(\frac{\|v\|^3}{\varkappa^2}\right) dv.
\end{aligned}$$

For the last three terms, we get

$$\int\limits_{\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|v\|^2}{h^2}} O\left(\frac{\|v\|^4}{h^2\varkappa}\right) dv \lesssim \frac{h^{d+2}}{\varkappa},$$
$$\int\limits_{\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|v\|^2}{h^2}} O\left(\frac{\|v\|^2}{\varkappa}\right) dv \lesssim \frac{h^{d+2}}{\varkappa},$$
$$\int\limits_{\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|v\|^2}{h^2}} O\left(\frac{\|v\|^3}{\varkappa^2}\right) dv \lesssim \frac{h^{d+3}}{\varkappa^2} \lesssim \frac{h^{d+2}}{\varkappa}.$$

Thus,

$$\int\limits_{\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|\mathcal{E}_{X_i}(v)-\mathcal{E}_{X_i}(0)\|^2}{h^2}} u^T(\mathcal{E}_{X_i}(v)-\mathcal{E}_{X_i}(0))\sqrt{\det g(v)}dv \tag{44}$$
$$= \int\limits_{\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|v\|^2}{h^2}} u^T v dv + O\left(\frac{h^{d+2}}{\varkappa}\right),$$

and, in order to complete the proof, we have to show that

$$\int\limits_{\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|v\|^2}{h^2}} u^T v dv = O\left(\frac{h^{d+2}}{\varkappa}\right).$$

Note that, for any fixed $u \in \mathbb{R}^d$, it holds

$$\int\limits_{\mathbb{R}^d} e^{-\frac{\|v\|^2}{h^2}} u^T v dv = 0.$$

Then

$$\int\limits_{\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|v\|^2}{h^2}} u^T v dv = -\int\limits_{\mathbb{R}^d\setminus\widetilde{\mathcal{B}}(X_i,r_d)} e^{-\frac{\|v\|^2}{h^2}} u^T v dv$$

Remind that $\widetilde{\mathcal{B}}(X_i, r_d) = \mathcal{E}_{X_i}^{-1}(\mathcal{B}(X_i, r_d)) = \{v : \|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\| \leqslant r_d\}$. By the definition of the exponential map,

$$\|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\| \leqslant d_{\mathcal{M}^*}(\mathcal{E}_{X_i}(v), \mathcal{E}_{X_i}(0)) = \|v\|.$$

Then we conclude that $\widetilde{\mathcal{B}}(X_i, r_d) \supseteq \mathcal{B}(0, r_d)$. This yields

$$\begin{aligned}
&\left| \int\limits_{\mathbb{R}^d \setminus \widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} u^T v dv \right| \leqslant \int\limits_{\mathbb{R}^d \setminus \widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} \|v\| dv \\
&\leqslant \int\limits_{\mathbb{R}^d \setminus \mathcal{B}(0, r_d)} e^{-\frac{\|v\|^2}{h^2}} \|v\| dv \leqslant e^{-\frac{r_d^2}{2h^2}} \int\limits_{\mathbb{R}^d \setminus \mathcal{B}(0, r_d)} e^{-\frac{\|v\|^2}{2h^2}} \|v\| dv \\
&\leqslant e^{-\frac{r_d^2}{2h^2}} \int\limits_{\mathbb{R}^d} e^{-\frac{\|v\|^2}{2h^2}} \|v\| dv.
\end{aligned}$$

By definition, $r_d = 2h\sqrt{2(d+2)\log h^{-1}}$. This implies $e^{-\frac{r_d^2}{2h^2}} = h^{4(d+2)}$. Moreover,

$$\int\limits_{\mathbb{R}^d} e^{-\frac{\|v\|^2}{2h^2}} \|v\| dv \lesssim h^{d+1}.$$

Thus, we conclude

$$\int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{h^2}} u^T v dv \lesssim h^{d+1+4(d+2)} \lesssim \frac{h^{d+2}}{\varkappa}, \tag{45}$$

and (45) finishes the proof of Lemma 8. ■

Lemma 9 *Let $\mathbb{E}^{(-i)}$ denote the conditional expectation $\mathbb{E}(\cdot|(X_i, Y_i))$ and let $r_d = 4h\sqrt{(d+2)\log h^{-1}}$. Then, for any i from 1 to n, it holds*

$$\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{16h^2}} \|X_j - X_i\|^q \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right) \lesssim h^{q+d}.$$

Proof

Using (A2), we obtain

$$\begin{aligned}
&\mathbb{E}^{(-i)} e^{-\frac{\|X_i - X_j\|^2}{16h^2}} \|X_j - X_i\|^q \mathbb{1}\left(X_j \in \mathcal{B}(X_i, r_d)\right) \\
&= \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|X_i - x\|^2}{16h^2}} \|x - X_i\|^q p(x) dW(x) \\
&\leqslant p_1 \int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|X_i - x\|^2}{16h^2}} \|x - X_i\|^q dW(x).
\end{aligned}$$

Using the exponential map, we get

$$\int\limits_{\mathcal{M}^* \cap \mathcal{B}(X_i, r_d)} e^{-\frac{\|X_i - x\|^2}{16h^2}} \|x - X_i\|^q dW(x)$$
$$= \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\|^2}{16h^2}} \|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\|^q \sqrt{\det g(v)} dv$$

Taking into account that $\|v\| = d_{\mathcal{M}^*}(\mathcal{E}_{X_i}(v), \mathcal{E}_{X_i}(0))$ and applying (Boissonnat et al., 2019, Lemma 2.5), we conclude

$$\frac{\|v\|}{2} \leqslant \|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\| \leqslant \|v\|.$$

On the other hand, (43) yields $\sqrt{\det g(v)} \lesssim 1$ for all $v \in \widetilde{\mathcal{B}}(X_i, r_d)$. Thus, we obtain

$$\int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\|^2}{16h^2}} \|\mathcal{E}_{X_i}(v) - \mathcal{E}_{X_i}(0)\|^q \sqrt{\det g(v)} dv$$
$$\lesssim \int\limits_{\widetilde{\mathcal{B}}(X_i, r_d)} e^{-\frac{\|v\|^2}{32h^2}} \|v\|^q dv \leqslant \int\limits_{\mathbb{R}^d} e^{-\frac{\|v\|^2}{32h^2}} \|v\|^q dv \lesssim h^{d+q}.$$

■

Appendix H. Pseudocode of the Manifold Blurring Mean Shift Algorithm

This section contains a pseudocode of the manifold blurring mean shift algorithm (Wang and Carreira-Perpinan, 2010, MBMS).

Algorithm 2 Manifold blurring mean shift algorithm (with full graph), Wang and Carreira-Perpinan (2010)

The sample of noisy observations $\mathbb{Y}_n = (Y_1, \dots, Y_n)$, a bandwidth $\sigma > 0$, and positive integers k and d are given.

Initialize $\widehat{X}_1 = Y_1, \dots, \widehat{X}_n = Y_n$.

repeat

Compute the increments

$$\partial \widehat{X}_i = -\widehat{X}_i + \frac{\sum_{j=1}^{n} \mathcal{K}\left(\frac{\|\widehat{X}_j - \widehat{X}_i\|^2}{2\sigma^2}\right) \widehat{X}_j}{\sum_{j=1}^{n} \mathcal{K}\left(\frac{\|\widehat{X}_j - \widehat{X}_i\|^2}{2\sigma^2}\right)}, \quad 1 \leqslant i \leqslant n,$$

where $\mathcal{K}(t) = e^{-t}$.

For each i from 1 to n, find k nearest neighbors $\mathcal{N}_i$ of $\widehat{X}_i$.

For all i from 1 to n, perform local PCA, that is compute

$$\mu_i = \frac{1}{\mathsf{k}} \sum_{j \in \mathcal{N}_i} \widehat{X}_j, \quad 1 \leqslant i \leqslant n,$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\mathsf{k}} \sum_{j \in \mathcal{N}_i} (\widehat{X}_j - \mu_i)(\widehat{X}_j - \mu_i)^T, \quad 1 \leqslant i \leqslant n,$$

and put $\boldsymbol{\Pi}_i$ a projector onto a linear span of eigenvectors of $\widehat{\boldsymbol{\Sigma}}_i$, corresponding to the largest d eigenvalues.

Update the increments

$$\partial \widehat{X}_i \leftarrow (\boldsymbol{I} - \boldsymbol{\Pi}_i) \partial \widehat{X}_i, \quad 1 \leqslant i \leqslant n.$$

Update the estimates

$$\widehat{X}_i \leftarrow \widehat{X}_i + \partial X_i, \quad 1 \leqslant i \leqslant n.$$

until stop

return the estimates $\widehat{X}_1, \dots, \widehat{X}_n$.

References

E. Aamari and C. Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete Comput. Geom.*, 59(4):923–971, 2018.

E. Aamari and C. Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.*, 47(1):177–204, 2019.

E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *J. Mach. Learn. Res.*, 17(43): 1–28, 2016.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

J.-D. Boissonnat, A. Lieutier, and M. Wintraecken. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *J. Appl. and Comput. Topology*, 3(1): 29–58, 2019.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

M. A. Carreira-Perpinan. Gaussian mean-shift is an em algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(5):767–776, 2007.

M. A. Carreira-Perpinan. A review of mean-shift algorithms for clustering. *ArXiv*, abs/1503.00687, 2015.

O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

M.-Y. Cheng and H.-T. Wu. Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc.*, 108(504):1421–1434, 2013.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, 1995.

D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.

D. Eberly, R. Gardner, B. Morse, S. Pizer, and C. Scharlach. Ridges for image analysis. *J. Math. Imaging Vis.*, 4(4):353–373, 1994.

H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.

C. Fefferman, S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan. Fitting a putative manifold to noisy data. *COLT, Proc. Mach. Learn. Res.*, 75:688–720, 2018.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory*, 21:32–40, 1975.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13:1263–1291, 2012a.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40(2):941–963, 2012b.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *Ann. Statist.*, 42(4):1511–1545, 2014.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 38(6):907–921, 2002.

E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. *IMS Lecture Notes Monogr. Ser.*, 51:238–259, 2006.

G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 4th edition, 2013.

D. Gong, F. Sha, and G. Medioni. Locally linear denoising on image manifolds. *AISTATS, Proc. Mach. Learn. Res.*, 9:265–272, 2010.

M. Hein and M. Maier. Manifold denoising. *NIPS*, 19:561–568, 2006.

M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Ann. Statist.*, 29(6):1537–1566, 2001a.

M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623, 2001b.

A. K. H. Kim and H. H. Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. *Electron. J. Stat.*, 9(1):1562–1582, 2015.

J. Li, S. Ray, and B. G. Lindsay. A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.*, 8(59):1687–1723, 2007.

M. Maggioni, S. Minsker, and N. Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *J. Mach. Learn. Res.*, 17(2):1–51, 2016.

S. Osher, Z. Shi, and W. Zhu. Low dimensional manifold model for image processing. *SIAM J. Img. Sci.*, 10:1669–1690, 2017.

U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *J. Mach. Learn. Res.*, 12:1249–1286, 2011.

B. Recht, W. Xu, and B. Hassibi. Null space conditions and thresholds for rank minimization. *Math. Program.*, 127(1, Ser. B):175–202, 2011.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

Z. Shi and J. Sun. Convergence of the point integral method for laplace–beltrami equation on point cloud. *Res. Math. Sci.*, 4(1):22, 2017.

S. J. Szarek. Metric entropy of homogeneous spaces. *Banach Center Publ.*, 43:395–410, 1998.

M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3): 505–563, 1996.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

N. G. Trillos, D. Sanz-Alonso, and R. Yang. Local regularization of noisy point clouds: Improved global geometric estimates and data analysis. *J. Mach. Learn. Res.*, 20(136): 1–37, 2019.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9: 2579–2605, 2008.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

W. Wang and M. A. Carreira-Perpinan. Manifold blurring mean shift algorithms for manifold denoising. *CVPR*, pages 1759–1766, 2010.

Y. Xia, H. Tong, W. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 64(3):363–410, 2002.

Z. Zhang and J. Wang. MLLE: Modified locally linear embedding using multiple weights. *NIPS*, 19:1593–1600, 2006.