Journal of Machine Learning Research 20 (2019) 1-37 Submitted 4/19; Revised 8/19; Published 8/19

Local Regularization of Noisy Point Clouds: Improved Global Geometric Estimates and Data Analysis

Nicolás García Trillos GARCIATRILLO@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Daniel Sanz-Alonso SANZALONSO@UCHICAGO.EDU
Ruiyi Yang YRY@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Editor: Sayan Mukherjee

Abstract

Several data analysis techniques employ similarity relationships between data points to uncover the intrinsic dimension and geometric structure of the underlying data-generating mechanism. In this paper we work under the model assumption that the data is made of random perturbations of feature vectors lying on a low-dimensional manifold. We study two questions: how to define the similarity relationships over noisy data points, and what is the resulting impact of the choice of similarity in the extraction of global geometric information from the underlying manifold. We provide concrete mathematical evidence that using a local regularization of the noisy data to define the similarity improves the approximation of the hidden Euclidean distance between unperturbed points. Furthermore, graph-based objects constructed with the locally regularized similarity function satisfy better error bounds in their recovery of global geometric ones. Our theory is supported by numerical experiments that demonstrate that the gain in geometric understanding facilitated by local regularization translates into a gain in classification accuracy in simulated and real data.

Keywords: manifold denoising, metric estimation, spectral convergence, graph Laplacian

1. Introduction

Several techniques for the analysis of high dimensional data build on the observation that data-generating mechanisms can often be described by few degrees of freedom. In this paper we study graph-based methods that employ similarity relationships between data points to uncover the low intrinsic dimension and geometric structure of datasets. Graph-based learning provides a well-balanced compromise between accuracy and interpretability Coifman and Lafon (2006), and is popular in a variety of unsupervised and semi-supervised tasks Zhu (2005); Von Luxburg (2007). These methods have been extensively analyzed in the idealized setting where the data is sampled from a low-dimensional manifold and

†. All authors contributed equally to this work.

similarities are computed using the ambient Euclidean distance or the geodesic distance, see e.g. Coifman and Lafon (2006); Singer (2006); Burago et al. (2014); García Trillos et al. (2018). The manifold setting is truthful in spirit to the presupposition that data arising from structured systems may be described by few degrees of freedom, but it is not so in that the data are typically noisy. The aim of this paper is to provide new mathematical theory under the more general and realistic model assumption that the data consist of random perturbations of low-dimensional features lying on a manifold.

By relaxing the manifold assumption we bring forward two fundamental questions that are at the heart of graph-based learning but have not been accounted for by previous theory. First, how to define the inter-point similarities between noisy data points in order to approximate the Euclidean distances between unperturbed data-points? Second, is it possible to recover global geometric features of the manifold from suitably-defined similarities between noisy data points? We will show by rigorous mathematical reasoning that:

(i) Denoising inter-point distances leads to an improved approximation of the hidden Euclidean distance between unperturbed points. We illustrate this general idea by analyzing a simple, easily-computable similarity defined in terms of a local-regularization of the noisy dataset.

(ii) Graph-based objects defined via locally regularized similarities can be guaranteed to satisfy improved error bounds in the recovery of global geometric properties. We illustrate this general idea by showing the spectral approximation of an unnormalized ε-graph Laplacian to a Laplace operator defined on the underlying manifold.

In addition to giving theoretical support for the denoising of point clouds, we study the practical use of local regularization in classification problems. Our analytically tractable local-regularization depends on a parameter that modulates the amount of localization, and our analysis suggests the appropriate scaling of said parameter with the level noise level. In our numerical experiments we show that in semi-supervised classification problems this parameter may be chosen by cross-validation, ultimately producing classification rules with improved accuracy. Finally, we propose two alternative denoising methods with similar empirical performance that are sometimes easier to implement. In short, the improved recovery of the geometric structure of the underlying point cloud facilitated by (local) regularization translates into improved graph-based data analysis, and the results seem to be robust to the choice of methodology.

1.1 Framework

We assume a data model

$$y_i = x_i + z_i, \tag{1}$$

where the *unobserved* points x_i are sampled from an unknown m-dimensional manifold $\mathcal{M}$, the vectors $z_i \in \mathbb{R}^d$ represent noise, and $\mathcal{Y}_n = \{y_1, \dots, y_n\} \subseteq \mathbb{R}^d$ is the observed data. Further geometric and probabilistic structure will be imposed to prove our main results—see Assumptions 1 and 2 below. Our analysis is motivated by the case, often found in applications, where the number n of data points and the ambient space dimension d are large, but the underlying intrinsic dimension m is small or moderate. Thus, the data-generating mechanism is described (up to a noisy perturbation) by $m \ll d$ degrees of

freedom. We aim to uncover geometric properties of the underlying manifold $\mathcal{M}$ from the observed data $\mathcal{Y}_n$ by using *similarity* graphs. The set of vertices of these graphs will be identified with the set $[n] := \{1, \dots, n\}$—so that the i-th node corresponds to the i-th data point—and the weight $W(i,j)$ between the i-th and j-th data-point will be defined in terms of a similarity function $\delta : [n] \times [n] \to [0, \infty)$.

The first question that we consider is how to choose the similarity function so that $\delta(i,j)$ approximates the hidden Euclidean distance $\delta_{\mathcal{X}_n}(i,j) := |x_i - x_j|$. Full knowledge of the Euclidean distance between the latent variables x_i would allow to recover, in the large n limit, global geometric features of the underlying manifold. This motivates the idea of denoising the observed point cloud $\mathcal{Y}_n$ to approximate the hidden similarity function $\delta_{\mathcal{X}_n}$. Here we will study a family of similarity functions based on the Euclidean distance between local averages of points in $\mathcal{Y}_n$, i.e. averages of the local measures. We define a denoised dataset $\bar{\mathcal{Y}}_n = \{\overline{y}_1, \dots, \overline{y}_n\}$ by locally averaging the original dataset, and we then define an associated similarity function

$$\delta_{\bar{\mathcal{Y}}_n}(i,j) := |\overline{y}_i - \overline{y}_j|.$$

In its simplest form, $\overline{y}_i$ is defined by averaging all points in $\mathcal{Y}_n$ that are inside the ball of radius $r > 0$ centered around y_i, that is,

$$\overline{y}_i := \frac{1}{N_i} \sum_{j \in \mathcal{A}_i} y_j, \tag{2}$$

where N_i is the cardinality of $\mathcal{A}_i := \{j \in [n] : y_j \in B(y_i, r)\}$. As discussed in Subsection 1.3, this corresponds to one step of the mean-shift algorithm Fukunaga and Hostetler (1975). Note that $\bar{\mathcal{Y}}_n$ (and the associated similarity function $\delta_{\bar{\mathcal{Y}}_n}$) depends on r, but we do not include said dependence in our notation for simplicity. Other possible local and non-local averaging approaches may be considered. We will only analyze the choice made in (2) and we will explore other constructions numerically. Introducing the notation

$$\delta_{\mathcal{X}_n}(i,j) = |x_i - x_j|, \quad \delta_{\mathcal{Y}_n}(i,j) = |y_i - y_j|,$$

the first question that we study may be formalized as understanding when, and to what extent, the similarity function $\delta_{\bar{\mathcal{Y}}_n}$ is a better approximation than $\delta_{\mathcal{Y}_n}$ (the standard choice) to the hidden similarity function $\delta_{\mathcal{X}_n}$. An answer is given in Theorem 1 below.

The second question that we investigate is how an improvement in the approximation of the hidden similarity function affects the approximation of the Laplace Beltrami operator on the underlying manifold $\mathcal{M}$. Specifically, we study how the spectral convergence of graph-Laplacians constructed with noisy data may be improved by local regularization of the point cloud. For concreteness, our theoretical analysis is focused on ε-graphs and unnormalized graph-Laplacians, but we expect our results to generalize to other graphs and graph-Laplacians—evidence to support this claim will be given through numerical experiments. We now summarize the necessary background to formalize this question. For a given similarity $\delta : [n] \times [n] \to [0, \infty)$ and a parameter $\varepsilon > 0$, we define a weighted graph $\Gamma_{\delta,\varepsilon} = ([n], W)$ by setting the weight between the i-th and j-th node to be

$$W(i,j) := \frac{2(m+2)}{\alpha_m \varepsilon^{m+2} n} \mathbb{1}\{\delta(i,j) < \varepsilon\}, \tag{3}$$

where α_m is the volume of the m-dimensional Euclidean unit ball. Associated to the graph $\Gamma_{\delta,\varepsilon}$ we define the unnormalized graph Laplacian matrix

$$\Delta_{\delta,\varepsilon} := D - W \in \mathbb{R}^{n\times n}, \tag{4}$$

where D is a diagonal matrix with diagonal entries

$$D(i,i) := \sum_{j=1}^{n} W(i,j).$$

The motivation for the scaling in (3) is so that $\Delta_{\delta,\epsilon}$ matches the scale of the Laplace-Beltrami operator (see for example Burago et al. (2014)). For the rest of the paper we shall denote $\Gamma_{\mathcal{X}_n,\varepsilon} := \Gamma_{\delta_{\mathcal{X}_n},\varepsilon}$ and $\Delta_{\mathcal{X}_n,\varepsilon} := \Delta_{\delta_{\mathcal{X}_n},\varepsilon}$. We use analogous notation for $\mathcal{Y}_n$ and $\bar{\mathcal{Y}}_n$. The second question that we consider may be formalized as understanding when, and to what extent, $\Delta_{\bar{\mathcal{Y}}_n}$ provides a better approximation (in the spectral sense) than $\Delta_{\mathcal{Y}_n}$ to a Laplace operator on the manifold $\mathcal{M}$. An answer is given in Theorem 3 below.

1.2 Main results

In this subsection we state our main theoretical results. We first impose some geometric conditions on the underlying manifold $\mathcal{M}$.

Assumption 1 *$\mathcal{M}$ is a smooth, oriented, compact manifold with no boundary and intrinsic dimension m, embedded in $\mathbb{R}^d$. Moreover, $\mathcal{M}$ has injectivity radius $\geq i_0$, maximum of the absolute value of sectional curvature $\leq K$, and reach $\geq R$. Finally, we assume that $\mathcal{M}$'s total volume is normalized and equal to one.*

Loosely speaking, the injectivity radius determines the range of the exponential map (which will be an important tool in our analysis and will be reviewed in the next section) and the sectional curvature controls the metric distortion induced by the exponential map, and thereby its Jacobian. The reach R can be thought of as an (inverse) conditioning number of the manifold and controls its second fundamental form; it can also be interpreted as a measure of extrinsic curvature—see, e.g. Aamari et al. (2019), Federer (1959) for technical background. The significance of these geometric quantities and their role in our analysis will be further discussed in Section 2.

Next we impose further probabilistic structure into the data model (1). We assume that the pairs (x_i, z_i) are i.i.d. samples of the random vector $(X, Z) \sim \boldsymbol{\mu} \in \mathcal{P}(\mathcal{M} \times \mathbb{R}^d)$. Let μ and μ_x be, respectively, the marginal distribution of X and the conditional distribution of Z given $X = x$. We assume that μ is absolutely continuous with respect to the Riemannian volume form of $\mathcal{M}$ with density $p(x)$, i.e.,

$$d\mu(x) = p(x) d\mathrm{vol}_{\mathcal{M}}(x). \tag{5}$$

Furthermore, we assume that μ_x is supported on $T_x\mathcal{M}^\perp$ (the orthogonal complement of the tangent space $T_x\mathcal{M}$) and that it is absolutely continuous with respect to the $(d-m)$-dimensional Hausdorff measure $\mathcal{H}^{d-m}$ restricted to $T_x\mathcal{M}^\perp$ with density $p(z|x)$, i.e.,

$$d\mu_x(z) = p(z|x) d\mathcal{H}^{d-m}(z).$$

To ease the notation we will write dz instead of $\mathcal{H}^{d-m}(dz)$. We make the following assumptions on these densities.

Assumption 2 *It holds that:*

(i) The density $p(x)$ is of class $C^2(\mathcal{M})$ and is bounded above and below by positive constants:

$$0 < p_{min} \leq p(x) \leq p_{max}, \quad \forall x \in \mathcal{M}.$$

(ii) For all $x \in \mathcal{M}$,

$$\int z p(z|x) dz = 0.$$

Moreover, there is $\sigma < R$ such that $p(z|x) = 0$ for all z with $|z| \geq \sigma$.

Note that the assumption on $p(z|x)$ ensures that the noise is centered and bounded by a constant σ. While the assumption that the noise is bounded and orthogonal to the manifold can be relaxed, we choose not to do so here to streamline our results and proofs.

In our first main theorem we study the approximation of the similarity function $\delta_{\mathcal{X}_n}$ by $\delta_{\bar{\mathcal{Y}}_n}$. We consider points x_i and x_j that are close with respect to the geodesic distance $d_{\mathcal{M}}$ on the manifold, and show that local regularization improves the approximation of the hidden similarity provided that n is large and the noise level σ is small. The local regularity parameter r needs to be suitably scaled with σ. We make the following standing assumption linking both parameters; we refer to Remark 2 below for a discussion on the optimal scaling of r with σ, and to our numerical experiments for practical guidelines.

Assumption 3 *The localization parameter r and the noise level σ satisfy*

$$\sigma \leq \frac{R}{16m}, \; r \leq \min\left\{ i_0, \frac{1}{\sqrt{K}}, \sqrt{\frac{\alpha_m}{2CmK}}, \sqrt{\frac{R}{32}} \right\}, \text{ and } \sigma \leq \frac{1}{3} r, \tag{6}$$

where C is a universal constant, α_m denotes the volume of the Euclidean unit ball in $\mathbb{R}^m$, and i_0, R, and K are as in Assumption 1.

In words, Assumption 3 requires both r and σ to be sufficiently small, and r to be larger than σ.

Now we are ready to state the first main result.

Theorem 1 *Under Assumptions 1, 2 and 3, with probability at least $1 - 4ne^{-cnr^{\max\{2m, m+4\}}}$, for all x_i and x_j with $d_{\mathcal{M}}(x_i, x_j) \leq r$ we have*

$$\left| \delta_{\mathcal{X}_n}(i,j) - \delta_{\bar{\mathcal{Y}}_n}(i,j) \right| \leq C_{\mathcal{M}} \left(r^3 + r\sigma + \frac{\sigma^2}{r} \right), \tag{7}$$

where $c = \min\left\{ \frac{\alpha_m^2 p_{min}^2}{4^{m+2}}, \frac{1}{16} \right\}$ and $C_{\mathcal{M}}$ is a constant depending on m, K, R, a uniform bound on the change in second fundamental form of $\mathcal{M}$, and on the regularity of the density p.

Remark 2 *Theorem 1 gives concrete evidence of the importance of the choice of similarity function. For the usual Euclidean distance between observed data, $\delta_{\mathcal{Y}_n}$, one can only guarantee that*

$$\big|\delta_{\mathcal{X}_n}(i,j) - \delta_{\mathcal{Y}_n}(i,j)\big| \leq 2\sigma,$$

which follows from

$$\big||x_i - x_j| - |y_i - y_j|\big| \leq |z_i - z_j| \leq 2\sigma.$$

However, if we choose $r \propto \sigma^{1/2}$, then the error in (7) is of order $\sigma^{3/2}$, which is a considerably smaller quantity in the small noise limit.

Our second main result translates the local similarity bound from Theorem 1 into a global geometric result concerning the spectral convergence of the graph Laplacian to the Laplace operator formally defined by

$$\Delta_{\mathcal{M}} f = -\frac{1}{p}\mathrm{div}\big(p^2 \nabla f\big), \tag{8}$$

where div and ∇ denote the divergence and gradient operators on the manifold and p is the sampling density of the hidden point cloud $\mathcal{X}_n$, as introduced in Equation (5). It is intuitively clear that the spectral approximation of the discrete graph-Laplacian to the continuum operator $\Delta_{\mathcal{M}}$ necessarily rests upon having a sufficient number of samples from μ (defined in (5)). In other words, the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ needs to be close to μ, the sampling density of the hidden dataset. We characterize the closeness between μ_n and μ by the ∞-OT transport distance, defined as

$$d_\infty(\mu_n, \mu) := \min_{T:T_\sharp \mu = \mu_n} \operatorname*{esssup}_{x \in \mathcal{M}} \quad d_{\mathcal{M}}\big(x, T(x)\big),$$

where $T_\sharp \mu$ denotes the push-forward of μ by T, that is, $T_\sharp \mu = \mu\big(T^{-1}(U)\big)$ for any Borel subset U of $\mathcal{M}$. Theorem 2 in García Trillos et al. (2018) shows that for every $\beta > 1$, with probability at least $1 - C_{\beta,\mathcal{M}} n^{-\beta}$,

$$d_\infty(\mu_n, \mu) \leq C_{\mathcal{M}} \frac{\log(n)^{p_m}}{n^{1/m}},$$

where $p_m = 3/4$ if $m = 2$ and $p_m = 1/m$ for $m \geq 3$. This is the high probability scaling of $d_\infty(\mu_n, \mu)$ in terms of n.

We introduce some notation before stating our second main result. Let $\lambda_\ell(\Gamma_{\delta,\varepsilon})$ be the ℓ-th smallest eigenvalue of the unnormalized graph-Laplacian $\Delta_{\delta,\varepsilon}$ defined in Equation (4), and let $\lambda_\ell(\mathcal{M})$ be the ℓ-th smallest eigenvalue of the continuum Laplace operator defined in Equation (8).

Theorem 3 *Suppose that Assumptions 1, 2, and 3 hold. Suppose further that ε is small enough (but not too small) so that*

$$\begin{aligned} \max\Big\{(m+5)d_\infty(\mu_n,\mu), 2Cm\eta\Big\} < \varepsilon < \min\Big\{1, \frac{i_0}{10}, \frac{1}{\sqrt{mK}}, \frac{R}{\sqrt{27m}}\Big\}, \\ \big(\sqrt{\lambda_\ell(\mathcal{M})} + 1\big)\varepsilon + \frac{d_\infty(\mu_n,\mu)}{\varepsilon} < \tilde{c}_p, \end{aligned} \tag{9}$$

where $\tilde{c}_p$ is a constant that only depends on m and the regularity of the density p, C is a universal constant, and

$$\eta = C_{\mathcal{M}} \left(r^3 + r\sigma + \frac{\sigma^2}{r} \right)$$

is the bound in (7)*. Then, with probability at least $1 - 4ne^{-cnr^{\max\{2m,m+4\}}}$, for all $\ell = 1, 2, 3, \ldots$,*

$$\frac{|\lambda_\ell(\Gamma_{\bar{\mathcal{Y}}_n,\varepsilon}) - \lambda_\ell(\mathcal{M})|}{\lambda_\ell(\mathcal{M})} \leq \tilde{C} \left(\frac{\eta}{\varepsilon} + \frac{d_\infty(\mu_n, \mu)}{\varepsilon} + \big(1 + \sqrt{\lambda_\ell(\mathcal{M})}\big)\varepsilon + \Big(K + \frac{1}{R^2}\Big)\varepsilon^2 \right),$$

where $\tilde{C}$ only depends on m and the regularity of p, and $c = \min\left\{ \frac{\alpha_m^2 p_{min}^2}{4^{m+2}}, \frac{1}{16} \right\}$.

Remark 4 *We will see in Section 3 that Theorem 3 follows by plugging the probabilitstic estimate* (7) *into a modification of a deterministic result from (García Trillos et al., 2018, Corollary 2), which we present for the convenience of the reader in Theorem 20. We remark that any improvement of Theorem 20 would immediately translate into an improvement of our Theorem 3. As discussed in Remark 2, local regularization enables a smaller η than if no regularization is performed. This in turn allows one to choose, for a given error tolerance, a smaller connectivity ε, leading to a sparser graph that is computationally more efficient. Note also that the bound in Theorem 3 does not depend on the ambient space dimension d, but only on the intrinsic dimension m of the data.*

Remark 5 *Theorem 3 concretely shows how an improvement in metric approximation translates into an improved estimation of global geometric quantities. We have restricted our attention to analyzing eigenvalues of a Laplacian operator, but we remark that the idea goes beyond this particular choice. For example, one can conduct an asymptotic analysis illustrating the effect of changing the similarity function in the approximation of other geometric quantities of interest like Cheeger cuts. Such analysis could be carried out using the variational convergence approach from García Trillos and Slepčev (2016). Finally, we remark that it is possible to study convergence of* eigenvectors *of graph Laplacians following the results in García Trillos et al. (2018).*

1.3 Related and Future Work

Graph-based learning algorithms include spectral clustering, total variation clustering, graph-Laplacian regularization for semi-supervised learning, graph based Bayesian semi-supervised learning. A brief and incomplete summary of methodological and review papers is Shi and Malik (2000); Ng et al. (2002); Belkin and Niyogi (2004); Zhou and Schölkopf (2005); Spielman and Teng (2007); Von Luxburg (2007); Zhu (2005); Bertozzi et al. (2018). These algorithms involve either a graph Laplacian, the graph total variation, or Sobolev norms involving the graph structure. The large sample $n \to \infty$ theory studying the behavior of some of the above methodologies has been analyzed without reference to the intrinsic dimension of the data Von Luxburg et al. (2008) and in the case of points laying *on* a low dimensional manifold, see e.g. Belkin et al. (2006); Garcia Trillos and Sanz-Alonso (2018); Garcia Trillos et al. (2017) and references therein. Some papers that account for *both* the noisy and low

intrinsic dimensional structure of data are Niyogi et al. (2008); Little and Maggioni (2017); Agapiou et al. (2017); Weed and Bach (2017); Genovese et al. (2012); Aamari and Levrard (2019). For example, Niyogi et al. (2008) studies the recovery of the homology groups of submanifolds from noisy samples. We use the techniques for the analysis of spectral convergence of graph-Laplacians introduced in Burago et al. (2014) and further developed in García Trillos et al. (2018). The results in the latter reference would allow to extend our analysis to other graph Laplacians, but we do not pursue this here for conciseness.

We highlight that the denoising by local regularization occurs at the level of the dataset. That is, rather than denoising each of the observed features individually, we analyze denoising by averaging different data points. In practice combining both forms of denoising may be advantageous. For instance, when each of the data points corresponds to an image, one can first denoise each image at the pixel level and then do regularization at the level of the dataset as proposed here. In this regard, our regularization at the level of the data-set is similar to applying a filter at the level of individual pixels Tukey and Tukey (1988). The success of non-local filter image denoising algorithms suggests that non-local methods may be also of interest at the level of the dataset, but we expect this to be application-dependent. Finally, while in this paper we only consider first-order regularization based on averages, a topic for further research is the analysis of local PCA regularization Little and Maggioni (2017), incorporating covariance information.

With the same motivation for our work, in Mémoli et al. (2018) a general construction of metrics on noisy datasets was proposed. The so called *Wasserstein transform* associates to each of the data points a "local" probability distribution, and defines a new metric on the data by computing the Wasserstein distance between the corresponding local measures. A particular construction of local measures closely related to the metric we study here assigns to each observation the empirical measure of the observations restricted to a ball of certain radius around the given data point. The authors of Mémoli et al. (2018) propose the Wasserstein transform as a way to generalize the *mean-shift algorithm* and they study how it alleviates the so called *chaining effect* in single linkage clustering. The aim of our work is to provide quantitative evidence of the effect that changing the metric on noisy datasets has on graph-based spectral clustering algorithms. The success of these algorithms hinges on their ability to capture the geometry of the underlying data generating model.

It is worth noting the parallel between the local regularization that we study here and mean-shift and mode seeking methods Chen et al. (2016); Fukunaga and Hostetler (1975). As a matter of fact the points $\overline{x}_i$ that we construct here correspond to one step in the standard mean shift algorithm. However, we notice that our goal is not to run mean shift for mode seeking, but rather, as a way to construct a metric that better captures the underlying "true" geometric structure of the data that was blurred by noise. This paralellism with mean-shift techniques (or the more general Wasserstein transform in Mémoli et al. (2018)) suggests the idea of doing local averaging iteratively. Of course, it is important to notice that unless one prevents points to move tangentially to $\mathcal{M}$ (as discussed in Wang and Carreira-Perpinán (2010)), a large number of iterations would result in points collapsing to a finite number of local modes.

Local regularization may be also interpreted as a form of dictionary learning, where each data-point is represented in terms of its neighbors. For specific applications it may be of interest to restrict (or extend) the dictionary used to represent each data point Haddad et al.

(2014). Finally we refer to Hein and Maier (2007) for alternative techniques on manifold denoising.

1.4 Outline

The paper is organized as follows. In Section 2 we formalize the geometric setup and prove Theorem 1. Section 3 contains the proof of Theorem 3 and a lemma that may be of independent interest. Finally, Section 4 includes several numerical experiments. In the Appendix we prove a technical lemma that serves as a key ingredient in proving Theorem 1.

2. Distance Approximation

In this section we prove Theorem 1. We start with Subsection 2.1 by giving some intuition on the geometric conditions imposed in Assumption 1 and introducing the main geometric tools in our analysis. In Subsection 2.2 we decompose the approximation error between the similarity functions $\delta_{\bar{\mathcal{Y}}_n}$ and $\delta_{\mathcal{X}_n}$ into three terms, which are bounded in Subsections 2.3, 2.4, and 2.5.

2.1 Geometric Preliminaries

In this subsection we set our notation and provide some background on geometric concepts used in the remainder of this paper.

2.1.1 Basic Notation

For each $x \in \mathcal{M}$ we let $T_x\mathcal{M}$ be the tangent plane of $\mathcal{M}$ at x centered at the origin. In particular, $T_x\mathcal{M}$ is a m-dimensional subspace of $\mathbb{R}^d$, and we denote by $T_x\mathcal{M}^\perp$ its orthogonal complement. We will use $\mathrm{vol}_{\mathcal{M}}$ to denote the Riemannian volume form of $\mathcal{M}$. We will denote by $|x - \tilde{x}|$ the Euclidean distance between arbitrary points in $\mathbb{R}^d$ and denote by $d_{\mathcal{M}}(x, \tilde{x})$ the geodesic distance between points in $\mathcal{M}$. We denote by B_x balls in $T_x\mathcal{M}$ and by $B_{\mathcal{M}}$ balls in the manifold $\mathcal{M}$ (with respect to the geodesic distance). Also, unless otherwise specified B, without subscripts will be used to denote balls in $\mathbb{R}^d$. We denote by α_m the volume of the unit Euclidean ball in $\mathbb{R}^m$. Throughout the rest of the paper we use R, i_0 and K to denote the reach, injectivity radius, and maximum absolute curvature of $\mathcal{M}$, as in Assumption 1. We now describe at an intuitive level the role that these quantities play in our analysis.

2.1.2 The Reach

The reach of a closed submanifold $\mathcal{M}$ is the largest value $t \in [0, \infty]$ such that the projection map onto $\mathcal{M}$ is well defined on $\{x \in \mathbb{R}^d : \inf_{\tilde{x}\in\mathcal{M}} |\tilde{x} - x| < t\}$, i.e., every point in the tubular neighborhood around $\mathcal{M}$ of width t has a unique closest point in $\mathcal{M}$. Our assumption that the noise level satisfies $\sigma < R$ guarantees that x_i is the (well-defined) projection of y_i onto the manifold. The reach can be thought of as an inverse conditioning number for the manifold Niyogi et al. (2008). We will use that the inverse of the reach provides a uniform upper bound on the second fundamental form (see Lemma 12).

2.1.3 Exponential Map, Injectivity Radius and Sectional Curvature

We will make use of the exponential map exp, which for every $x \in \mathcal{M}$ is a map

$$\exp_x : B_x(0, i_0) \to B_{\mathcal{M}}(x, i_0)$$

where i_0 is the injectivity radius for the manifold $\mathcal{M}$. We recall that the exponential map $\exp_x$ takes a vector $v \in T_x\mathcal{M}$ and maps it to the point $\exp_x(v) \in \mathcal{M}$ that is at geodesic distance $|v|$ from x along the unit speed geodesic that at time $t = 0$ passes through x with velocity $v/|v|$. The injectivity radius i_0 is precisely the maximum radius of a ball in $T_x\mathcal{M}$ centered at the origin for which the exponential map is a well defined diffeomorphism for every x. We denote by J_x the Jacobian of the exponential map $\exp_x$. Integrals with respect to $d\text{vol}_{\mathcal{M}}$ can then be written in terms of integrals on $T_x\mathcal{M}$ weighted by the function J_x. More precisely, for an arbitrary test function $\varphi : \mathcal{M} \to \mathbb{R}$,

$$\int_{B_{\mathcal{M}}(x,i_0)} \varphi(\tilde{x}) d\text{vol}_{\mathcal{M}}(\tilde{x}) = \int_{B_x(0,i_0)} \varphi\big(\exp_x(v)\big) J_x(v) dv.$$

For fixed $0 < r \leq \min\{i_0, 1/\sqrt{K}\}$ one can obtain bounds on the metric distortion by the exponential map $\exp_x \colon B_x(0, r) \subseteq T_x\mathcal{M} \to \mathcal{M}$ ((do Carmo, 1992, Chapter 10) and (Burago et al., 2014, Section 2.2)), and thereby guarantee the existence of a universal constant C such that, for $|v| \leq r$,

$$(1 + CmK|v|^2)^{-1} \leq J_x(v) \leq (1 + CmK|v|^2). \tag{10}$$

An immediate consequence of the previous inequalities is

$$|\text{vol}(B_{\mathcal{M}}(x, r)) - \alpha_m r^m| \leq CmKr^{m+2}, \tag{11}$$

where we recall α_m is the volume of the unit ball in $\mathbb{R}^m$. Equations (10) and (11) will be used in our geometric and probabilistic arguments and motivate our assumptions on the choice of local regularization parameter r in terms of the injectivity radius and the sectional curvature.

2.2 Local Distributions

Next we study the local behavior of (X, Z). To characterize its local distribution, it will be convenient to introduce the following family of probability measures.

Definition 6 *Let y be a vector in $\mathbb{R}^d$ whose distance to $\mathcal{M}$ is less than R. Let x be the projection of y onto $\mathcal{M}$. We say that the random variable $(\tilde{X}, \tilde{Z})$ has the distribution $\boldsymbol{\mu}_y$ provided that*

$$\mathbb{P}\big((\tilde{X}, \tilde{Z}) \in A_1 \times A_2\big) := \mathbb{P}\big((X, Z) \in A_1 \times A_2 | X + Z \in B(y, r)\big),$$

for all Borel sets $A_1 \subseteq \mathcal{M}$ $A_2 \subseteq \mathbb{R}^d$, where in the above (X, Z) is distributed according to $\boldsymbol{\mu}$.

In the remainder we use $\boldsymbol{\mu}_i$ as shorthand notation for $\boldsymbol{\mu}_{y_i}$. As for the original measure $\boldsymbol{\mu}$, we characterize $\boldsymbol{\mu}_i$ in terms of a marginal and conditional distribution. We introduce the density $\widetilde{p}_i : \mathcal{M} \to \mathbb{R}$ given by

$$\widetilde{p}_i(x) := \frac{\mathbb{P}_i\big(X + Z \in B(y_i, r) | X = x\big)}{\mathbb{P}_i\big(X + Z \in B(y_i, r)\big)} \cdot p(x), \tag{12}$$

and define

$$\widetilde{p}_i(z|x) = \frac{\mathbb{1}_{x+z \in B(y_i,r)}}{\mathbb{P}_i\big(X + Z \in B(y_i, r) | X = x\big)} \cdot p(z|x), \tag{13}$$

where in the above and in the remainder we use $\mathbb{E}_i$ and $\mathbb{P}_i$ to denote conditional expectation and conditional probability given (x_i, z_i). It can be easily shown that these functions correspond to the marginal density of $\tilde{X}_i$ and the conditional density of $\tilde{Z}_i$ given $\tilde{X}_i = x$, where $(\tilde{X}_i, \tilde{Z}_i) \sim \boldsymbol{\mu}_i$. The distribution $\boldsymbol{\mu}_i$ is of relevance because by definition of $\overline{y}_i$ one has

$$\mathbb{E}_i[\overline{y}_i] = \mathbb{E}_i[\tilde{X}_i + \tilde{Z}_i].$$

Now we are ready to introduce the main decomposition of the error between the similarity functions $\delta_{\bar{\mathcal{Y}}_n}$ and $\delta_{\mathcal{X}_n}$. Using the triangle inequality we can write

$$\begin{aligned} \big||x_i - x_j| - |\bar{y}_i - \bar{y}_j|\big| \leq \big|\mathbb{E}_i[\tilde{X}_i] - x_i - (\mathbb{E}_j[\tilde{X}_j] - x_j)\big| & \tag{14} \\ + \big|\mathbb{E}_j[\tilde{Z}_j]\big| + \big|\mathbb{E}_i[\tilde{Z}_i]\big| & \tag{15} \\ + \big|\mathbb{E}_i[\bar{y}_i] - \bar{y}_i\big| + \big|\mathbb{E}_j[\bar{y}_j] - \bar{y}_j\big|. & \tag{16} \end{aligned}$$

In the next subsections we bound each of the terms (15) (expected conditional noise), (14) (difference in *geometric* bias), and (16) (sampling error). As we will see in Subsection 2.5 we can control both terms in (16) with very high probability using standard concentration inequalities. The other three terms are deterministic quantities that can be written in terms of integrals with respect to the distributions $\tilde{\mu}_i$ and $\tilde{\mu}_j$. To study these integrals it will be convenient to introduce two quantities $r_- < r < r_+$ (independent of $i = 1, \ldots, n$) satisfying:

i) For all $x \in \mathcal{M}$ with $d_{\mathcal{M}}(x, x_i) > r_+$ we have

$$\mathbb{P}_i\big(X + Z \in B(y_i, r) | X = x\big) = 0.$$

Equivalently, the density $\tilde{p}_i(x)$ is supported in $\overline{B_{\mathcal{M}}(x_i, r_+)}$.

ii) For all x with $d_{\mathcal{M}}(x, x_i) < r_-$ we have

$$\mathbb{P}_i\big(X + Z \in B(y_i, r) | X = x\big) = 1.$$

It should be noted that the choice of both r_- and r_+ depends on r. In Appendix A we present the proof of the following lemma giving estimates for r_+ and r_-.

Lemma 7 (Bounds for r_+ and r_-) *Under Assumption 3, the quantities*

$$r_- := r\left(\sqrt{1 + \frac{4\sigma}{R} + \frac{16\sigma^2}{r^2}} + \frac{m\sigma}{R}\right)^{-1},$$

$$r_+ := r\left(\sqrt{1 - \frac{8r^2}{R} - \frac{4\sigma}{R}} - \frac{m\sigma}{R}\right)^{-1},$$

satisfy properties i) and ii). Furthermore,

$$r_+ - r_- \leq C_{m,R}\left(r^3 + r\sigma + \frac{\sigma^2}{r}\right),\ C_{m,R} := \max\left\{\frac{8m+32}{R}, 64\right\}$$

and

$$\frac{1}{2}r_+ \leq r \leq 2r_-. \tag{17}$$

2.3 Bounding Expected Conditional Noise

Proposition 8 *Suppose that Assumptions 1 and 2 hold. Then,*

$$\left|\mathbb{E}_i[\tilde{Z}_i]\right| \leq C_{m,p}\frac{\sigma}{r}(r_+ - r_-), \qquad C_{m,p} := \frac{4^{m+1}p_{max}}{mp_{min}}.$$

Proof Using the definition of r_+,

$$\begin{aligned}
\mathbb{E}_i[\tilde{Z}_i] &= \int_{B_{\mathcal{M}}(x_i,r_+)} \int z\tilde{p}_i(z|x)dz\ \tilde{p}_i(x)\, d\mathrm{vol}_{\mathcal{M}}(x) \\
&= \int_{B_{\mathcal{M}}(x_i,r_-)} \int z\tilde{p}_i(z|x)dz\ \tilde{p}_i(x)\, d\mathrm{vol}_{\mathcal{M}}(x) \\
&+ \int_{B_{\mathcal{M}}(x_i,r_+)\setminus B_{\mathcal{M}}(x_i,r_-)} \int z\tilde{p}_i(z|x)dz\, \tilde{p}_i(x) d\mathrm{vol}_{\mathcal{M}}(x).
\end{aligned}$$

The first integral is the zero vector because for $x \in B_{\mathcal{M}}(x_i, r_-)$, we have $\tilde{p}(z|x) \propto p(z|x)$ and $p(z|x)$ is assumed to be centered. Therefore,

$$\begin{aligned}
\left|\mathbb{E}_i[\tilde{Z}_i]\right| &\leq \sigma \int_{B_{\mathcal{M}}(x_i,r_+)\setminus B_{\mathcal{M}}(x_i,r_-)} \tilde{p}_i(x) d\mathrm{vol}_{\mathcal{M}}(x) \\
&= \frac{\sigma}{\mathbb{P}_i\big(X+Z\in B(y_i,r)\big)} \int_{B_{\mathcal{M}}(x_i,r_+)\setminus B_{\mathcal{M}}(x_i,r_-)} p(x) d\mathrm{vol}_{\mathcal{M}}(x) \\
&\leq \frac{\sigma p_{max}}{\mathbb{P}_i\big(X+Z\in B(y_i,r)\big)} \int_{B_{\mathcal{M}}(x_i,r_+)\setminus B_{\mathcal{M}}(x_i,r_-)} d\mathrm{vol}_{\mathcal{M}}(x) \\
&\leq \frac{\sigma p_{max}}{\mathbb{P}_i\big(X+Z\in B(y_i,r)\big)} \int_{B_{x_i}(0,r_+)\setminus B_{x_i}(0,r_-)} J_{x_i}(v) dv \\
&\leq \frac{2\alpha_m \sigma p_{max}}{\mathbb{P}_i\big(X+Z\in B(y_i,r)\big)} (r_+^m - r_-^m) \\
&\leq \frac{2\alpha_m \sigma p_{max}}{m\mathbb{P}_i\big(X+Z\in B(y_i,r)\big)} (r_+ - r_-) r_+^{m-1},
\end{aligned}$$

where we have used (10) and the assumptions on r to say (in particular) that $J_{x_i}(v) \leq 2$, and also the fact that, for $t > s > 0$,

$$t^m - s^m = \int_s^t \frac{u^{m-1}}{m} du \leq (t-s)\frac{t^{m-1}}{m}.$$

Finally, notice that

$$\mathbb{P}_i\big(X+Z \in B(y_i,r)\big) \geq \mathbb{P}_i\big(X \in B_{\mathcal{M}}(x_i,r_-)\big) = \int_{B_{x_i}(0,r_-)} p\big(\exp_{x_i}(v)\big) J_{x_i}(v) dv \geq \frac{1}{2} p_{min} \alpha_m r_-^m,$$

where again we have used (10) to conclude (in particular) that $J_{x_i}(v) \geq 1/2$. The result now follows by (17).

■

2.4 Bounding Difference in Geometric Bias

In terms of r_+ and r_-, the difference $\mathbb{E}_i[\tilde{X}_i] - x_i$ (and likewise $\mathbb{E}_j[\tilde{X}_j] - x_j$) can be written as:

$$\begin{aligned}
\mathbb{E}_i[\tilde{X}_i] - x_i =& \int_{B_{\mathcal{M}}(x_i,r_+)} (x - x_i)\tilde{p}_i(x) d\mathrm{vol}_{\mathcal{M}}(x) \\
=& \int_{B_{x_i}(0,r_+)} \big(\exp_{x_i}(v) - x_i\big)\tilde{p}_i\big(\exp_x(v)\big) J_{x_i}(v) dv \\
=& \int_{B_{x_i}(0,r_+)} \big(\exp_{x_i}(v) - x_i\big)\tilde{p}_i\big(\exp_x(v)\big) dv + \\
& \int_{B_{x_i}(0,r_+)} \big(\exp_{x_i}(v) - x_i\big)\tilde{p}_i\big(\exp_x(v)\big)\big(J_{x_i}(v) - 1\big) dv \\
=& \frac{1}{\mathbb{P}_i\big(X + Z \in B(y_i,r)\big)} \int_{B_{x_i}(0,r_-)} \big(\exp_{x_i}(v) - x_i\big) p\big(\exp_x(v)\big) dv + \\
& \int_{B_{x_i}(0,r_+)\setminus B_{x_i}(0,r_-)} \big(\exp_{x_i}(v) - x_i\big)\tilde{p}_i\big(\exp_x(v)\big) dv + \\
& \int_{B_{x_i}(0,r_+)} \big(\exp_{x_i}(v) - x_i\big)\tilde{p}_i\big(\exp_x(v)\big)\big(J_{x_i}(v) - 1\big) dv \\
:=& \frac{1}{\mathbb{P}_i\big(X + Z \in B(y_i,r)\big)} \int_{B_{x_i}(0,r_-)} \big(\exp_{x_i}(v) - x_i\big) p\big(\exp_x(v)\big) dv + \xi_i,
\end{aligned}$$

where the second to last equality follows from (12). To further simplify the expression for $x_i - \mathbb{E}_i[\tilde{X}_i]$ let us define

$$b_i := \int_{B_{x_i}(0,r_-)} \big(\exp_{x_i}(v) - x_i\big) p\big(\exp_x(v)\big) dv.$$

It follows that

$$\begin{aligned}
\big|\mathbb{E}_i[\tilde{X}_i] - x_i - (\mathbb{E}_j[\tilde{X}_j] - x_j)\big| &\leq \Big|\frac{b_i}{P_i} - \frac{b_j}{P_j}\Big| + \big|\xi_i\big| + \big|\xi_j\big| \\
&\leq \Big|\frac{1}{P_i} - \frac{1}{P_j}\Big|\big|b_i\big| + \frac{1}{P_j}\big|b_i - b_j\big| + \big|\xi_i\big| + \big|\xi_j\big|,
\end{aligned} \tag{18}$$

where in the above

$$P_i := \mathbb{P}_i\big(X + Z \in B(y_i,r)\big), \quad P_j := \mathbb{P}_j\big(X + Z \in B(y_j,r)\big).$$

Lemma 9 *The following hold.*

i) *The terms P_i satisfy*

$$\frac{1}{2}p_{min}\alpha_m r_-^m \leq P_i.$$

ii) *The terms ξ_i satisfy:*

$$\big|\xi_i\big| \leq C_1(r_+ - r_-) + C_2 r^3,$$

where, up to universal multiplicative constants,

$$C_1 = \frac{4^{m+1}p_{max}}{mp_{min}}, \quad C_2 = 4^{m+3}mK\frac{p_{max}}{p_{min}}.$$

iii) *Suppose that $d_{\mathcal{M}}(x_i, x_j) \leq r$. Then,*

$$|P_i - P_j| \leq C_3 r^{m+1} + C_4(r_+ - r_-)r^{m-1} + C_5 r^{m+2},$$

where, up to universal multiplicative constants,

$$C_3 = C_p\alpha_m, \quad C_4 = \frac{2^{m-1}\alpha_m p_{max}}{m}, \quad C_5 = mKp_{max}\alpha_m.$$

and C_p only depends on bounds on the first derivatives of the density p.

Proof The first inequality was already obtained at the end of the proof of Proposition 8. For the second inequality recall that

$$\begin{aligned}\xi_i = &\int_{B_{x_i}(0,r_+)\setminus B_{x_i}(0,r_-)} \big(x_i - \exp_{x_i}(v)\big)\tilde{p}_i\big(\exp_{x_i}(v)\big)dv + \\ &\int_{B_{x_i}(0,r_+)} \big(x_i - \exp_{x_i}(v)\big)\tilde{p}_i\big(\exp_{x_i}(v)\big)[J_{x_i}(v) - 1]dv := I_1 + I_2.\end{aligned}$$

For the first term we notice that $|x_i - \exp_{x_i}(v)| \leq d_{\mathcal{M}}(x_i, \exp_{x_i}(v)) \leq r_+$. Thus using i) and the definition of $\tilde{p}_i$ we have

$$|I_1| \leq \frac{r_+ p_{max}\alpha_m}{\mathbb{P}_i\big(X + Z \in B(y_i, r)\big)}(r_+^m - r_-^m) \leq \frac{4^{m+1}p_{max}}{mp_{min}}(r_+ - r_-).$$

For the second term we use i) and (10) to see that

$$|I_2| \leq \frac{CmKp_{max}\alpha_m}{\mathbb{P}_i\big(X + Z \in B(y_i, r)\big)}r_+^{m+3} \leq C4^{m+3}mK\frac{p_{max}}{p_{min}}r^3.$$

For iii) we notice that by definition of r_- and r_+ we can write

$$\mathbb{P}_i\big(X \in B_{x_i}(0, r_-)\big) - \mathbb{P}_j\big(X \in B_{x_j}(0, r_+)\big) \leq P_i - P_j \leq \mathbb{P}_i\big(X \in B_{x_i}(0, r_+)\big) - \mathbb{P}_j\big(X \in B_{x_j}(0, r_-)\big),$$

and in particular it is enough to bound $H_{ij} := \big|\mathbb{P}_i\big(X \in B_{\mathcal{M}}(x_i, r_+)\big) - \mathbb{P}_j\big(X \in B_{\mathcal{M}}(x_j, r_-)\big)\big|$. We can expand H_{ij} as follows.

$$\begin{aligned}
H_{ij} &= \int_{B_{x_i}(0,r_-)} p\big(\exp_{x_i}(v)\big)dv - \int_{B_{x_j}(0,r_-)} p\big(\exp_{x_j}(\tilde{v})\big)d\tilde{v} \\
&+ \int_{B_{x_i}(0,r_+)\setminus B_{x_i}(0,r_-)} p\big(\exp_{x_i}(v)\big)dv \\
&+ \int_{B_{x_i}(0,r_-)} p\big(\exp_{x_i}(v)\big)\big(J_{x_i}(v)-1\big)dv - \int_{B_{x_j}(0,r_-)} p\big(\exp_{x_j}(\tilde{v})\big)\big(J_{x_j}(\tilde{v})-1\big)d\tilde{v} \\
&:= \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3.
\end{aligned}$$

By a similar argument as above, we can bound $\mathcal{I}_2$ and $\mathcal{I}_3$ by

$$\begin{aligned}
|\mathcal{I}_2| &\leq p_{max}\alpha_m(r_+^m - r_-^m) \leq \frac{2^{m-1}}{m}\alpha_m p_{max}(r_+ - r_-)r^{m-1}, \\
|\mathcal{I}_3| &\leq 2CmKp_{max}\alpha_m r_-^{m+2} \leq 2CmKp_{max}\alpha_m r^{m+2}.
\end{aligned}$$

Finally, we notice that we can identify $B_{x_i}(0, r_-)$ with $B_{x_j}(0, r_-)$. From the assumed smoothness on p (which in particular is C^1) we see that for any $v \in B_{x_i}(0, r_-)$ we have

$$\big|p\big(\exp_{x_i}(v)\big) - p\big(\exp_{x_j}(v)\big)\big| \leq C_p d_{\mathcal{M}}\big(\exp_{x_i}(v), \exp_{x_j}(v)\big) \leq 3C_p r.$$

Then it follows that $|\mathcal{I}_1| \leq 3C_p\alpha_m r^{m+1}$ and we get the desired result. ∎

We now bound the difference $\big|b_i - b_j\big|$ for nearby points x_i, x_j, where we recall that

$$b_i := \int_{B_{x_i}(0,r_-)} \big(\exp_{x_i}(v) - x_i\big)p\big(\exp_{x_i}(v)\big)\,dv.$$

Proposition 10 *Suppose that x_i and x_j are such that $d_{\mathcal{M}}(x_i, x_j) \leq r$. Then,*

$$\big|b_i - b_j\big| \leq Cr^{m+3},$$

where the constant C can be written as

$$C = p_{max}\alpha_m\left(\frac{6\sqrt{m}}{R^2} + \Big(1 + \frac{4}{R}\Big)C_{\mathcal{M}}\right) + \frac{C_p}{R}\alpha_m,$$

where C_p is a constant that depends on bounds on first and second derivatives of the density p, and $C_{\mathcal{M}}$ is a constant that depends only on the change in second fundamental form along $\mathcal{M}$ (a third order term).

As we will see Proposition 10 can be proved combining several ideas from differential geometry. We present the required auxiliary results as we develop the proof of the proposition.

We start by conveniently writing b_i and b_j in a way that facilitates their direct comparison. Indeed, for any given $v \in B_{x_i}(0, r_-)$ let us consider the curves

$$\gamma_{v,i}(t) := \exp_{x_i}\left(t\frac{v}{|v|}\right), \quad t \in [0, |v|],$$

and

$$t \in [0, |v|] \mapsto x_i + t \in [0, |v|].$$

Thus, $\gamma_{v,i}$ is an arc-length parameterized geodesic on $\mathcal{M}$ that starts at the point x_i and at time $|v|$ passes though the point $\exp_{x_i}(v)$. Its initial velocity $\dot{\gamma}_{v,i}(0)$ is the vector $v/|v|$. On the other hand, while the second curve does not stay in $\mathcal{M}$ for $t > 0$, it does have the same starting point and velocity as $\gamma_{v,i}$. We can use the fundamental theorem of calculus to write:

$$\exp_{x_i}(v) - (x_i + v) = \int_0^{|v|} \left(\dot{\gamma}_{v,i} - \frac{v}{|v|} \right) dt,$$

as well as

$$\dot{\gamma}_{v,i}(t) - \frac{v}{|v|} = \int_0^t \ddot{\gamma}_{v,i}(s) ds, \quad \forall t \in [0, |v|]. \tag{19}$$

In particular, we have the second order representation

$$\exp_{x_i}(v) - x_i = v + \int_0^{|v|} \int_0^t \ddot{\gamma}_{v,i}(s) ds dt. \tag{20}$$

As a consequence of the previous formula we can rewrite b_i as

$$\begin{aligned} b_i &= \int_{B_{x_i}(0,r_-)} \big(\exp_{x_i}(v) - x_i \big) p\big(\exp_{x_i}(v) \big) dv \\ &= \int_{B_{x_i}(0,r_-)} p\big(\exp_{x_i}(v) \big) \int_0^{|v|} \int_0^t \ddot{\gamma}_{v,i}(s) ds dt dv + \int_{B_{x_i}(0,r_-)} v p\big(\exp_{x_i}(v) \big) dv. \end{aligned} \tag{21}$$

Completely analogous definitions and statements can be introduced to represent b_j.

With the objective of using the formula (21) to compare b_i and b_j we relate vectors in $T_{x_i}\mathcal{M}$ with vectors in $T_{x_j}\mathcal{M}$ by a convenient linear isometry $F_{ij} : T_{x_i}\mathcal{M} \mapsto T_{x_j}\mathcal{M}$ constructed using parallel transport.

Lemma 11 *Suppose that x_i and x_j are such that $d_{\mathcal{M}}(x_i, x_j) \leq r$. Let $\phi : t \in [0, d_{\mathcal{M}}(x_i, x_j)] \mapsto \phi(t) \in \mathcal{M}$, be the arc-length parameterized geodesic starting at x_i at time zero and passing through x_j at time $t = d_{\mathcal{M}}(x_i, x_j)$. For an arbitrary vector $v \in T_{x_i}\mathcal{M}$ let V_v be the (unique) vector field along ϕ that solves the ODE*

$$\begin{cases} \frac{D}{dt} V_v(t) = 0, \quad t \in \big(0, d_{\mathcal{M}}(x_i, x_j)\big), \\ V_v(0) = v, \end{cases},$$

where $\frac{D}{dt}$ denotes the covariant derivative (on $\mathcal{M}$) along the curve ϕ. Then, the map F_{ij} defined by

$$F_{ij} : v \longmapsto \tilde{v} := V_v\big(d_{\mathcal{M}}(x_i, x_j)\big)$$

is a linear isometry. Moreover,

$$|v - \tilde{v}| \leq \frac{1}{R} |v| d_{\mathcal{M}}(x_i, x_j), \quad \forall v \in T_{x_i}\mathcal{M}. \tag{22}$$

Proof First note that F_{ij} is a linear isometry since the ODE defining V_v is linear and the vector fields V_v are parallel to the curve ϕ by definition. To get the estimate (22) we can use the fundamental theorem of calculus and write

$$\tilde{v} = v + \int_0^t \dot{V}_v(s)ds,$$

where $t := d_{\mathcal{M}}(x_i, x_j)$. The fact that V_v is parallel along the curve ϕ implies that $\dot{V}_v(s) \in T_{\phi(s)}\mathcal{M}^{\perp}$ and furthermore that for arbitrary unit norm η with $\eta \in T_{\phi(s)}\mathcal{M}^{\perp}$ we have

$$|\langle \dot{V}_v(s), \eta\rangle| = |\langle S_\eta(V_v(s)), \dot{\phi}(s)\rangle| \leq \|S_\eta\||V_v(s)||\dot{\phi}(s)| = \|S_\eta\||v|,$$

where S_η is the so called *shape operator* representing the second fundamental form (see Proposition 2.3. Chapter 6 in do Carmo (1992)). The relevance of the previous inequality is that when combined with Proposition 6.1 in Niyogi et al. (2008) (which shows that the operator norm of the second fundamental form is bounded by $1/R$) it implies that

$$|\dot{V}_v(s)| \leq \frac{|v|}{R}, \quad \forall s \in [0, t].$$

Therefore,

$$|\tilde{v} - v| \leq \int_0^t |\dot{V}_v(s)|ds \leq \frac{|v|}{R} d_{\mathcal{M}}(x_i, x_j),$$

establishing in this way the desired bound. ■

From now on, for a given $v \in B_{x_i}(0, r_-)$ we let $\tilde{v} \in B_{x_j}(0, r_-)$ be its image under F_{ij}. We consider the curve

$$\gamma_{\tilde{v},j}(t) := \exp_{x_j}\left(t\frac{\tilde{v}}{|\tilde{v}|}\right), \quad t \in [0, |\tilde{v}|],$$

where we recall that $|v| = |\tilde{v}|$ because F_{ij} is a linear isometry. We can then make a change of variables and write b_j as

$$b_j = \int_{B_{x_i}(0,r_-)} p\big(\exp_{x_j}(\tilde{v})\big) \int_0^{|v|} \int_0^t \ddot{\gamma}_{\tilde{v},j}(s)dsdtdv + \int_{B_{x_i}(0,r_-)} \tilde{v}p\big(\exp_{x_j}(\tilde{v})\big)dv. \tag{23}$$

In the next lemma we find bounds for the norms of accelerations.

Lemma 12 *Let $v \in B_{x_i}(0, r_-)$ and let $\tilde{v}$ be as in Lemma 11. Then, for all $t \in [0, |v|]$ we have*

$$|\ddot{\gamma}_{v,i}(t)| \leq \frac{1}{R},$$

and

$$|\dot{\gamma}_{v,i}(t) - \dot{\gamma}_{\tilde{v},j}(t)| \leq 2\frac{|v|}{R} + \frac{d_{\mathcal{M}}(x_i, x_j)}{R}.$$

Proof The first inequality appears in the proof of Proposition 2 in Niyogi et al. (2008) and is obtained in a completely analogous way as we obtained the bound for $\dot{V}_v$ in the proof of Lemma 11 (given that unit speed geodesics are auto parallel).

To prove the second estimate, we notice that from the first bound and (19) it follows that

$$\left|\dot{\gamma}_{v,i}(t) - \frac{v}{|v|}\right| \leq \frac{|v|}{R}, \quad \forall t \in [0, |v|].$$

Naturally, a similar inequality holds for $\gamma_{\tilde{v},j}$. Using Lemma 11 we conclude that for all $t \in [0, |v|]$ (recall that $|v| = |\tilde{v}|$)

$$\begin{aligned}
|\dot{\gamma}_{v,i}(t) - \dot{\gamma}_{\tilde{v},j}(t)| &\leq \left|\frac{v}{|v|} - \frac{\tilde{v}}{|\tilde{v}|}\right| + \left|\dot{\gamma}_{v,i}(t) - \frac{v}{|v|}\right| + \left|\dot{\gamma}_{\tilde{v},i}(t) - \frac{\tilde{v}}{|\tilde{v}|}\right| \\
&\leq \frac{1}{|v|}|v - \tilde{v}| + 2\frac{|v|}{R} \\
&\leq \frac{d_{\mathcal{M}}(x_i, x_j)}{R} + 2\frac{|v|}{R}.
\end{aligned}$$

■

From our assumption that the density p was in $C^2(\mathcal{M})$ it follows that

$$\begin{aligned}
p\big(\exp_{x_i}(v)\big) &= p(x_i) + \langle \nabla p(x_i), v\rangle + R_i(v), \\
p\big(\exp_{x_j}(\tilde{v})\big) &= p(x_j) + \langle \nabla p(x_j), \tilde{v}\rangle + R_j(\tilde{v}),
\end{aligned}$$

where $\nabla p(x_i)$ is the gradient (in $\mathcal{M}$) of p at the point x_i, and the remainder terms satisfy

$$\max\{|R_i(v)|, |R_j(\tilde{v})|\} \leq C_p|v|^2,$$

for a constant C_p that depends on a uniform bound on second derivatives of p. Likewise,

$$\max\{|p(x_i) - p(x_j)|, |\nabla p(x_i) - \nabla p(x_j)|\} \leq C_p d_{\mathcal{M}}(x_i, x_j).$$

Plugging the previous identities in the expressions (21) and (23), using

$$\int_{B_{x_i}(0,r_-)} p(x_i) v dv = 0, \quad \int_{B_{x_i}(0,r_-)} p(x_j)\tilde{v} d\tilde{v} = 0,$$

inequality (22), the bound on accelerations from Lemma 12, and finally Assumption 2, we can conclude that

$$\begin{aligned}
\big|b_i - b_j\big| &\leq \int_{B_{x_i}(0,r_-)} \int_0^{|v|} \int_0^t \big|p(x_j)\ddot{\gamma}_{\tilde{v},j}(s) - p(x_i)\ddot{\gamma}_{v,i}(s)\big| ds dt dv + \frac{C_p}{R}\alpha_m r^{m+2}\big(r + d_{\mathcal{M}}(x_i, x_j)\big) \\
&\leq p_{max} \cdot \int_{B_{x_i}(0,r_-)} \int_0^{|v|} \int_0^t \big|\ddot{\gamma}_{\tilde{v},j}(s) - \ddot{\gamma}_{v,i}(s)\big| ds dt dv \\
&+ \int_{B_{x_i}(0,r_-)} \int_0^{|v|} \int_0^t \big|p(x_j) - p(x_i)\big|\big|\ddot{\gamma}_{v,i}(s)\big| ds dt dv + \frac{C_p}{R}\alpha_m r^{m+2}\big(r + d_{\mathcal{M}}(x_i, x_j)\big) \\
&\leq p_{max} \cdot \int_{B_{x_i}(0,r_-)} \int_0^{|v|} \int_0^t \big|\ddot{\gamma}_{\tilde{v},j}(s) - \ddot{\gamma}_{v,i}(s)\big| ds dt dv + \frac{C_p}{R}\alpha_m r^{m+2}\big(r + d_{\mathcal{M}}(x_i, x_j)\big).
\end{aligned} \tag{24}$$

In the above, C_p is a constant that depends on derivatives of p of order 1 and order 2 (and in particular is equal to zero when p is constant) and α_m is the volume of the m-dimensional unit ball.

Proposition 10 now follows from the next lemma where we bound the difference of accelerations.

Lemma 13 *Let $v \in B_{x_i}(0, r_-)$ and let $\tilde{v}$ be as in Lemma 11. Then, for all $t \in [0, |v|]$ we have*

$$\left|\ddot{\gamma}_{v,i}(t) - \ddot{\gamma}_{\tilde{v},j}(t)\right| \leq \left(2\frac{\sqrt{m}}{R^2} + C_{\mathcal{M}}\right)\left(2|v| + d_{\mathcal{M}}(x_i, x_j)\right) + 2C_{\mathcal{M}}\left(\frac{|v|}{R} + \frac{d_{\mathcal{M}}(x_i, x_j)}{R}\right),$$

where $C_{\mathcal{M}}$ is a constant that depends on $\mathcal{M}$ (a third order term).

Proof For a fixed $t \in [0, |v|]$ we let

$$x := \gamma_{v,i}(t), \quad \tilde{x} := \gamma_{\tilde{v},j}(t).$$

We start by constructing a convenient linear map

$$\eta \in T_x\mathcal{M}^{\perp} \mapsto \tilde{\eta} \in T_{\tilde{x}}\mathcal{M}^{\perp}.$$

For this purpose we use an *orthonormal frame* $E_1, \ldots, E_m$ on a neighborhood (in $\mathcal{M}$) of x containing the geodesic connecting x and $\tilde{x}$. The frame is constructed by parallel transporting an orthonormal basis $E_1(x), \ldots, E_m(x)$ of $T_x\mathcal{M}$ along geodesics emanating from x. Now, associated to $\eta \in T_x\mathcal{M}^{\perp}$ we define the (normal) vector field N_η by

$$N_\eta := \eta - \sum_{l=1}^{m} \langle E_l, \eta\rangle E_l.$$

Equivalently, N_η can be written as

$$N_\eta(z) = \Pi_z(\eta),$$

where for a point $z \in \mathcal{M}$, Π_z denotes the projection onto $T_z\mathcal{M}^{\perp}$ (the orthogonal complement of the tangent plane at z).

Let $\phi_{x\tilde{x}}$ be the arc-length parameterized geodesic with $\phi_{x\tilde{x}}(0) = x$ and $\phi_{x\tilde{x}}(\tilde{t}) = \tilde{x}$. We restrict the vector field N_η to the curve $\phi_{x,\tilde{x}}$ and abuse notation slightly to write $N_\eta(s)$ and $E_j(s)$ for the value of the vector fields at the point $\phi_{x\tilde{x}}(s)$. We let $\tilde{\eta} := N_\eta(\tilde{t})$ and notice that

$$|\eta - \tilde{\eta}| = \left(\sum_{l=1}^{m} \langle E_l(\tilde{t}), \eta\rangle^2\right)^{1/2} = \left(\sum_{l=1}^{m} \langle E_l(\tilde{t}) - E_l(0), \eta\rangle^2\right)^{1/2} \leq \frac{\sqrt{m} d_{\mathcal{M}}(x, \tilde{x})|\eta|}{R}, \qquad (25)$$

where in the last line we have used that $|E_l(\tilde{t}) - E_l(0)| \leq \frac{\tilde{t}}{R}$ (proved in the exact same way as (22)).

Let $\eta \in T_x\mathcal{M}^\perp$ be a unit norm vector and let $\tilde{\eta}$ be as constructed before. Since N_η is a normal vector field which locally extends η we can follow the characterization for the shape operator in Proposition 2.3 Chapter 6 in do Carmo (1992) and deduce that:

$$\langle \ddot{\gamma}_{v,i}(t), \eta \rangle = \langle S_\eta(\dot{\gamma}_{v,i}), \dot{\gamma}_{v,i}(t) \rangle = \langle \frac{d}{dt} N_\eta(\gamma_{v,i}(t)), \dot{\gamma}_{v,i}(t) \rangle.$$

Moreover, the smoothness of the manifold $\mathcal{M}$ allows us to extend N_η smoothly to a neighborhood in $\mathbb{R}^d$ of x and $\tilde{x}$ (we also use N_η to represent the extension). Indeed, for any point z in a tubular neighborhood of $\mathcal{M}$ of width smaller than R we can define

$$N_\eta(z) := N_\eta(Proj_{\mathcal{M}}(z)),$$

where $Proj_{\mathcal{M}}$ is the projection onto $\mathcal{M}$ (which is well defined for points within distance R from $\mathcal{M}$). The smoothness of N_η in particular implies that

$$\|DN_\eta(x) - DN_\eta(\tilde{x})\| \leq C_{\mathcal{M}} |x - \tilde{x}| \leq C_{\mathcal{M}} d_{\mathcal{M}}(x, \tilde{x}),$$

$$\|DN_\eta(\tilde{x})\| \leq C_{\mathcal{M}},$$

where DN_η is the matrix of derivatives of the vector field N_η, and where $C_{\mathcal{M}}$ is some constant that only depends on $\mathcal{M}$. With this extension at hand, we can then use the chain rule and write:

$$\langle \ddot{\gamma}_{v,i}(t), \eta \rangle = \langle DN(x)\dot{\gamma}_{v,i}(t), \dot{\gamma}_{v,i}(t) \rangle,$$

and in a similar fashion

$$\langle \ddot{\gamma}_{\tilde{v},j}(t), \eta \rangle = \langle \ddot{\gamma}_{\tilde{v},j}(t), \eta - \tilde{\eta} \rangle + \langle \ddot{\gamma}_{\tilde{v},j}(t), \tilde{\eta} \rangle = \langle \ddot{\gamma}_{\tilde{v},j}(t), \eta - \tilde{\eta} \rangle + \langle -DN(\tilde{x})\dot{\gamma}_{\tilde{v},j}(t), \dot{\gamma}_{\tilde{v},j}(t) \rangle.$$

Using the triangle and Cauchy-Schwarz inequalities we obtain

$$\begin{aligned} |\langle \ddot{\gamma}_{v,i}(t) - \ddot{\gamma}_{\tilde{v},j}(t), \eta \rangle| \leq & |\ddot{\gamma}_{\tilde{v},j}||\eta - \tilde{\eta}| + \|DN(x) - DN(\tilde{x})\||\dot{\gamma}_{v,i}|^2 + \\ & \|DN(\tilde{x})\||\dot{\gamma}_{v,i} - \dot{\gamma}_{\tilde{v},j}|(|\dot{\gamma}_{\tilde{v},j}| + |\dot{\gamma}_{v,i}|) \\ \leq & \frac{\sqrt{m}}{R^2} d_{\mathcal{M}}(x, \tilde{x}) + C_{\mathcal{M}} d_{\mathcal{M}}(x, \tilde{x}) + 2C_{\mathcal{M}} \left(\frac{|v|}{R} + \frac{d_{\mathcal{M}}(x_i, x_j)}{R} \right). \end{aligned}$$

Since the above inequality holds for all $\eta \in T_x\mathcal{M}^\perp$ with norm one, we conclude that

$$|\Pi_x(\ddot{\gamma}_{v,i}(t)) - \Pi_x(\ddot{\gamma}_{\tilde{v},j}(t))| \leq \frac{\sqrt{m}}{R^2} d_{\mathcal{M}}(x, \tilde{x}) + C_{\mathcal{M}} d_{\mathcal{M}}(x, \tilde{x}) + 2C_{\mathcal{M}} \left(\frac{|v|}{R} + \frac{d_{\mathcal{M}}(x_i, x_j)}{R} \right),$$

where we recall Π_x represents the projection onto $T_x\mathcal{M}^\perp$. Moreover, since $\ddot{\gamma}_{v,i}(t)$ is the acceleration of a unit speed geodesic passing through x, we know that $\ddot{\gamma}_{v,i}(t) \in T_x\mathcal{M}^\perp$, so that $\Pi_x(\ddot{\gamma}_{v,i}) = \ddot{\gamma}_{v,i}$. Similarly we have $\Pi_{\tilde{x}}(\ddot{\gamma}_{\tilde{v},j}) = \ddot{\gamma}_{\tilde{v},j}$ (where $\Pi_{\tilde{x}}$ represents projection onto $T_{\tilde{x}}\mathcal{M}^\perp$) . Hence

$$|\ddot{\gamma}_{v,i}(t) - \ddot{\gamma}_{\tilde{v},j}(t)| \leq |\Pi_x \ddot{\gamma}_{v,i}(t) - \Pi_x \ddot{\gamma}_{\tilde{v},j}(t)| + |\Pi_x \ddot{\gamma}_{\tilde{v},j}(t) - \ddot{\gamma}_{\tilde{v},j}(t)|, \tag{26}$$

and so it remains to find a bound for $|\Pi_x \ddot{\gamma}_{\tilde{v},j}(t) - \ddot{\gamma}_{\tilde{v},j}(t)|$. We can write

$$\Pi_x \ddot{\gamma}_{\tilde{v},j} = \ddot{\gamma}_{\tilde{v},j} - \sum_{l=1}^{m} \langle \ddot{\gamma}_{\tilde{v},j}, E_l(0)\rangle E_l(0).$$

Therefore,

$$|\ddot{\gamma}_{\tilde{v},j} - \Pi_x \ddot{\gamma}_{\tilde{v},j}| = \left(\sum_{l=1}^{m} \langle \ddot{\gamma}_{\tilde{v},j}, E_l(0)\rangle^2\right)^{1/2} = \left(\sum_{l=1}^{m} \langle \ddot{\gamma}_{\tilde{v},j}, E_l(0) - E_l(\tilde{t})\rangle^2\right)^{1/2} \leq \sqrt{m}\frac{d_{\mathcal{M}}(x,\tilde{x})}{R^2}.$$

Putting everything together we deduce that

$$\begin{aligned} |\ddot{\gamma}_{v,i} - \ddot{\gamma}_{\tilde{v},j}| &\leq (2\frac{\sqrt{m}}{R^2} + C_{\mathcal{M}}) d_{\mathcal{M}}(x,\tilde{x}) + 2C_{\mathcal{M}} \left(\frac{|v|}{R} + \frac{d_{\mathcal{M}}(x_i,x_j)}{R}\right) \\ &\leq (2\frac{\sqrt{m}}{R^2} + C_{\mathcal{M}})(2|v| + d_{\mathcal{M}}(x_i,x_j)) + 2C_{\mathcal{M}} \left(\frac{|v|}{R} + \frac{d_{\mathcal{M}}(x_i,x_j)}{R}\right), \end{aligned}$$

where in the last step we have used the triangle inequality

$$d_{\mathcal{M}}(x,\tilde{x}) \leq d_{\mathcal{M}}(x,x_i) + d_{\mathcal{M}}(x_i,x_j) + d_{\mathcal{M}}(x_j,\tilde{x}) \leq 2|v| + d_{\mathcal{M}}(x_i,x_j).$$

■

Remark 14 *Notice that the computations in the proof of Proposition 10 also show that*

$$|b_i| \leq C r^{m+2}, \qquad i = 1, \ldots, n.$$

Indeed, this can be seen directly from (21)*, Lemma 12 (which bounds the acceleration term), and the fact that the first term on the right-hand side of the following expression drops by symmetry:*

$$\int_{B_{x_i}(0,r_-)} p(\exp_{x_i}(v)) v dv = p(x_i) \int_{B_{x_i}(0,r_-)} v dv + \int_{B_{x_i}(0,r_-)} (\langle \nabla p(x_i), v\rangle + R_i(v)) v dv.$$

2.5 Bounding Sampling Error

We will make use of two concentration inequalities to bound the sampling error. We first recall Hoeffding's inequality.

Lemma 15 (Hoeffding's inequality) *Let $w_1, \ldots, w_n$ be i.i.d samples from a random variable w taking values in the interval [0,1] and let $\overline{w}$ be the sample average. Then,*

$$\mathbb{P}\left(|\overline{w} - \mathbb{E}[\overline{w}]| > t\right) \leq 2e^{-2nt^2}.$$

The next is a generalization for random vectors that follows directly from the simple and elegant work Pinelis (1992) (more precisely, Theorem 3).

Lemma 16 *Let $W_1, \dots, W_n$ be i.i.d samples from a random vector W such that $|W| \leq M$ for some constant M, and $\mathbb{E}[W] = 0$. Let $\overline{W}$ be the sample average. Then,*

$$\mathbb{P}\left(\left|\overline{W} - \mathbb{E}[\overline{W}]\right| > \sqrt{\frac{M^2}{n}}t\right) \leq 2e^{-t^2/16}.$$

Proposition 17 *Suppose Assumption 3 holds. Then,*

$$\mathbb{P}\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{2^{m+4}}{\alpha_m p_{min}}}r^3\right) \leq 4e^{-cnr^{\max\{2m,m+4\}}}, \text{ where } c = \min\left\{\frac{\alpha_m^2 p_{min}^2}{4^{m+2}}, \frac{1}{16}\right\}.$$

In particular, if $nr^{\max\{2m,m+4\}} \gg 1$, then $\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| \leq \sqrt{\frac{2^{m+4}}{\alpha_m p_{min}}}r^3$ with high probability.

Proof Let N_i be the number of points in $B(y_i, r)$. Notice that $\tilde{x}_i + \tilde{z}_i - \mathbb{E}_i[\tilde{X}_i + \tilde{Z}_i]$ is centered and bounded by $2r$ in norm, and $\overline{y}_i = \overline{\tilde{x}_i + \tilde{z}_i}$. Then Lemma 16 implies

$$\mathbb{P}_i\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{4r^2}{N_i}}t \middle| N_i\right) \leq 2e^{-t^2/16}.$$

By the law of iterated expectations it follows that

$$\mathbb{P}\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{4r^2}{N_i}}t\right) \leq 2e^{-t^2/16}.$$

Next note that N_i, the number of points y_j in $B(y_i, r)$, can be bounded below by $\widetilde{N}_i$, the number of points x_j that lie in the ball $B_{\mathcal{M}}(x_i, r_-)$. Thus,

$$\mathbb{P}\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{4r^2}{\widetilde{N}_i}}t\right) \leq 2e^{-t^2/16}. \tag{27}$$

Now we find probabilistic bound for $\widetilde{N}_i$. Let $w_j = \mathbb{1}\{x_j \in B(x_i, r_-)\}$. Then given x_i, the w_j are i.i.d samples from Bernoulli(q_i), where $q_i = \mu\big(B_{\mathcal{M}}(x_i, r_-)\big)$. Lemma 15 implies

$$\mathbb{P}_i\left(\left|\widetilde{N}_i - nq_i\right| > nt \middle| x_i\right) \leq 2e^{-2nt^2}.$$

Again by the law of iterated expectation and rearranging terms, we have

$$\mathbb{P}\left(\widetilde{N}_i < n(q_i - t)\right) \leq 2e^{-2nt^2}. \tag{28}$$

Combining (27) and (28), we obtain

$$\begin{aligned}
\mathbb{P}\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{4r^2}{n(q_i - s)}}t\right) &= \mathbb{P}\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{4r^2}{n(q_i - s)}}t, \widetilde{N}_i < n(q_i - s)\right) \\
&\quad + \mathbb{P}\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{4r^2}{n(q_i - s)}}t, \widetilde{N}_i \geq n(q_i - s)\right) \\
&\leq \mathbb{P}\left(\widetilde{N}_i < n(q_i - s)\right) + \mathbb{P}\left(\left|\overline{y}_i - \mathbb{E}_i[\overline{y}_i]\right| > \sqrt{\frac{4r^2}{\widetilde{N}_i}}t\right) \\
&\leq 2e^{-2ns^2} + 2e^{-t^2/16}.
\end{aligned}$$

Under Assumption 3, (11) implies $q_i \geq \frac{\alpha_m p_{min}}{2^{m+1}} r^m$. Taking $s = \frac{\alpha_m p_{min}}{2^{m+2}} r^m$ and $t = \sqrt{n r^{m+4}}$, we see that

$$\mathbb{P}\left(\left| \overline{y}_i - \mathbb{E}_i[\overline{y}_i] \right| > \sqrt{\frac{2^{m+4}}{\alpha_m p_{min}}} r^3 \right) \leq 2e^{-\frac{\alpha_m^2 p_{min}^2}{4^{m+2}} n r^{2m}} + 2e^{-n r^{m+4}/16} \leq 4e^{-cnr^{\max\{2m, m+4\}}},$$

where $c = \min\left\{ \frac{\alpha_m^2 p_{min}^2}{4^{m+2}}, \frac{1}{16} \right\}$. The result then follows. ■

Theorem 1 now follows by combining Lemma 9, Propositions 8, 10, and Proposition 17 together with a union bound.

3. From Local Regularization to Global Estimates

In this section we use the local estimates (7) to show spectral convergence of $\Delta_{\bar{\mathcal{Y}}_n, \varepsilon}$ towards the continuum Laplace-Beltrami operator. We first make some definitions. Recall that the graph $\Gamma_{\delta,\varepsilon} = ([n], W)$ has weights

$$W(i,j) = \frac{2(m+2)}{\alpha_m \varepsilon^{m+2} n} \mathbb{1}\{\delta(i,j) < \varepsilon\},$$

where m is the dimension of $\mathcal{M}$ and α_m is the volume of the $m-$dimensional Euclidean unit ball. For a function $u : [n] \to \mathbb{R}$, we denote its value on the i-th node as $u(i)$. We then define the discrete Dirichlet energy of u as

$$E_{\delta,\varepsilon}[u] = \frac{m+2}{\alpha_m \varepsilon^{m+2} n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}\{\delta(i,j) < \varepsilon\} |u(i) - u(j)|^2$$

and the L^2 norm of u as

$$\|u\|^2 = \frac{1}{n} \sum_{i=1}^{n} |u(i)|^2.$$

Given that $\Delta_{\delta,\varepsilon}$ is a positive semi-definite operator, we can use the minimax principle (see for example Lieb et al. (2001)) to write

$$\lambda_\ell(\Gamma_{\delta,\varepsilon}) = \min_{L} \max_{u \in L \setminus \{0\}} \frac{E_{\delta,\varepsilon}[u]}{\|u\|^2},$$

where $\lambda_\ell(\Gamma_{\delta,\varepsilon})$ is the ℓ-th smallest eigenvalue of $\Delta_{\Gamma_{\delta,\varepsilon}}$ and the minimum is taken over all subspaces L of dimension ℓ. The following lemma compares the eigenvalues of the discrete graphs constructed using $\delta_{\mathcal{X}_n}$ and $\delta_{\bar{\mathcal{Y}}_n}$.

Lemma 18 *Let η be the bound in (7) so that for all i, j with $d_{\mathcal{M}}(x_i, x_j) \leq r$ we have*

$$\left| \delta_{\mathcal{X}_n}(i,j) - \delta_{\bar{\mathcal{Y}}_n}(i,j) \right| \leq \eta.$$

Suppose that ε is chosen so that $\varepsilon \geq 2Cm\eta$, for some universal constant C. Then,

$$\left(1 - Cm\frac{\eta}{\varepsilon}\right) \lambda_\ell(\Gamma_{\mathcal{X}_n, \varepsilon - \eta}) \leq \lambda_\ell(\Gamma_{\bar{\mathcal{Y}}_n, \varepsilon}) \leq \left(1 + Cm\frac{\eta}{\varepsilon}\right) \lambda_\ell(\Gamma_{\mathcal{X}_n, \varepsilon + \eta}). \tag{29}$$

Proof We first compare the Dirichlet energies. Since $\delta_{\mathcal{X}_n}(i,j) < \delta_{\bar{\mathcal{Y}}_n}(i,j) + \eta$, we have

$$\begin{aligned} E_{\bar{\mathcal{Y}}_n,\varepsilon}[u] &= \frac{m+2}{\alpha_m \varepsilon^{m+2} n} \sum_i \sum_j \mathbb{1}\{\delta_{\bar{\mathcal{Y}}_n}(i,j) < \varepsilon\} |u_i - u_j|^2 \\ &\leq \frac{m+2}{\alpha_m \varepsilon^{m+2} n} \sum_i \sum_j \mathbb{1}\{\delta_{\mathcal{X}_n}(i,j) < \varepsilon + \eta\} |u_i - u_j|^2 \\ &= \Big(\frac{\varepsilon+\eta}{\varepsilon}\Big)^{m+2} E_{\mathcal{X}_n,\varepsilon+\eta}[u] \\ &\leq \Big(1 + Cm\frac{\eta}{\varepsilon}\Big) E_{\mathcal{X}_n,\varepsilon+\eta}[u]. \end{aligned} \tag{30}$$

Now we use the minimax principle to show the upper-bound on (29). Let $u_1, \ldots, u_\ell$ be the first l eigenvectors of $\Delta_{\mathcal{X}_n,\varepsilon+\eta}$ and let $L = \text{span}\{u_1, \ldots, u_\ell\}$. Then $\dim L = \ell$ and for any $u \in L$, $E_{\mathcal{X}_n,\varepsilon+\eta}[u] \leq \lambda_\ell(\Gamma_{\mathcal{X}_n,\varepsilon+\eta})\|u\|^2$. Then by (30), we have

$$\lambda_\ell(\Gamma_{\bar{\mathcal{Y}}_n,\varepsilon}) \leq \max_{L\backslash 0} \frac{E_{\bar{\mathcal{Y}}_n,\varepsilon}[u]}{\|u\|^2} \leq \Big(1 + Cm\frac{\eta}{\varepsilon}\Big) \max_{L\backslash 0} \frac{E_{\mathcal{X}_n,\varepsilon+\eta}[u]}{\|u\|^2} = \Big(1 + Cm\frac{\eta}{\varepsilon}\Big) \lambda_\ell(\Gamma_{\mathcal{X}_n,\varepsilon+\eta}).$$

By a similar argument applied to $\Gamma_{\mathcal{X}_n,\varepsilon-\eta}$ and $\Gamma_{\bar{\mathcal{Y}}_n,\varepsilon}$, we get the lower-bound in (29). ■

Remark 19 *With the convergence of eigenvalues and the relationship between the Dirichlet energies it is also possible to make statements about convergence of eigenvectors (or better yet, spectral projections).*

The spectral convergence towards the continuum (Theorem 3) is a consequence of the following theorem, proved in (García Trillos et al., 2018, Corollary 2).

Theorem 20 *Let d_∞ be the ∞-OT distance between μ_n and μ. Suppose ε satisfies the conditions in Equation* (9) *and that Assumptions 1 and 2 hold. Then*

$$\frac{|\lambda_\ell(\Gamma_{\mathcal{X}_n,\varepsilon}) - \lambda_\ell(\mathcal{M})|}{\lambda_\ell(\mathcal{M})} \leq \tilde{C}\left(\frac{d_\infty}{\varepsilon} + \big(1 + \sqrt{\lambda_\ell(\mathcal{M})}\big)\varepsilon + \Big(K + \frac{1}{R^2}\Big)\varepsilon^2\right),$$

where $\tilde{C}$ only depends on m and the regularity of p.

Combining Lemma 18 and Theorem 20 gives Theorem 3.

4. Numerical Experiments

In this section we present a series of numerical experiments where we conduct local regularization on three different datasets. In Subsection 4.1 we consider a toy example with artificial data generated by perturbing points sampled uniformly from the unit, two-dimensional sphere embedded in $\mathbb{R}^d$ with $d = 100$. We show that the approximation of the hidden Euclidean distances between unperturbed points is significantly improved by locally regularizing the data, and that this improvement translates into better spectral approximation of the spherical Laplacian. Our numerical findings corroborate the theory developed in the previous two sections. In Subsection 4.2 we consider the two-moon and MNIST datasets and show that graphs constructed with locally regularized data can be used to improve the performance of a simple graph-based optimization method for semi-supervised classification.

4.1 Distance & Spectrum

Here we study the effect of local regularization on distance approximation and spectral convergence, as an illustration of the results from Sections 2 and 3. In our toy model we consider uniform samples from the unit two-dimensional sphere $\mathcal{M} = \mathcal{S}$ embedded in $\mathbb{R}^d$, with $d = 100$. The motivation for such a choice is that the eigenvalues of the associated Laplace-Beltrami operator on $\mathcal{S}$ are known explicitly (see for example Olver). Indeed, after appropriate normalization, $\Delta_{\mathcal{S}}$ admits eigenvalues $\ell(\ell+1), \ell \in \mathbb{N}$, with corresponding multiplicity $2\ell + 1$.

The dataset is generated by sampling $n = 3000$ points x_i uniformly from the sphere and adding uniform noise z_i normal to the tangent plane, and bounded by σ in norm. To be more precise, the noise is normal to the sphere for the first three dimensions and uniform in all directions for the rest dimensions. Local regularization is performed by taking $r \propto \sqrt{\sigma}$ and the graph is constructed with $\varepsilon = 2n^{-1/4}$. The optimal proportion constant in r is not obvious from our theory and in the experiments below we choose $r = \sqrt{\sigma}/3$ for $\sigma = 0.1$ and $r = \sqrt{\sigma}$ for the rest of the σ's. We first show that the $\bar{y}_i$ give a better approximation of the pairwise distances of the x_i than the y_i do. We only consider those nodes i, j such that $\delta_{\mathcal{X}_n}(i,j) < \varepsilon$ (i.e. the nodes that are relevant for the construction of the graph Laplacians). More precisely, let $D_{\mathcal{X}_n}$ be the matrix whose ijth entry is $\delta_{\mathcal{X}_n}(i,j)\mathbb{1}\{\delta_{\mathcal{X}_n}(i,j) < \varepsilon\}$. Similarly, we define $[D_{\mathcal{Y}_n}]_{ij} = \delta_{\mathcal{Y}_n}(i,j)\mathbb{1}\{\delta_{\mathcal{X}_n}(i,j) < \varepsilon\}$ and $[D_{\bar{\mathcal{Y}}_n}]_{ij} = \delta_{\bar{\mathcal{Y}}_n}(i,j)\mathbb{1}\{\delta_{\mathcal{X}_n}(i,j) < \varepsilon\}$. In Table 1 we compare the entrywise ∞-norm of the $D_{\mathcal{X}_n} - D_{\mathcal{Y}_n}$ and $D_{\mathcal{X}_n} - D_{\bar{\mathcal{Y}}_n}$ for different values of σ. We see that the improvement is substantial.

	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$	$\sigma = 0.7$	$\sigma = 0.9$
$\|D_{\mathcal{X}_n} - D_{\mathcal{Y}_n}\|_\infty$	0.095	0.298	0.505	0.723	0.937
$\|D_{\mathcal{X}_n} - D_{\bar{\mathcal{Y}}_n}\|_\infty$	0.084	0.090	0.087	0.093	0.144

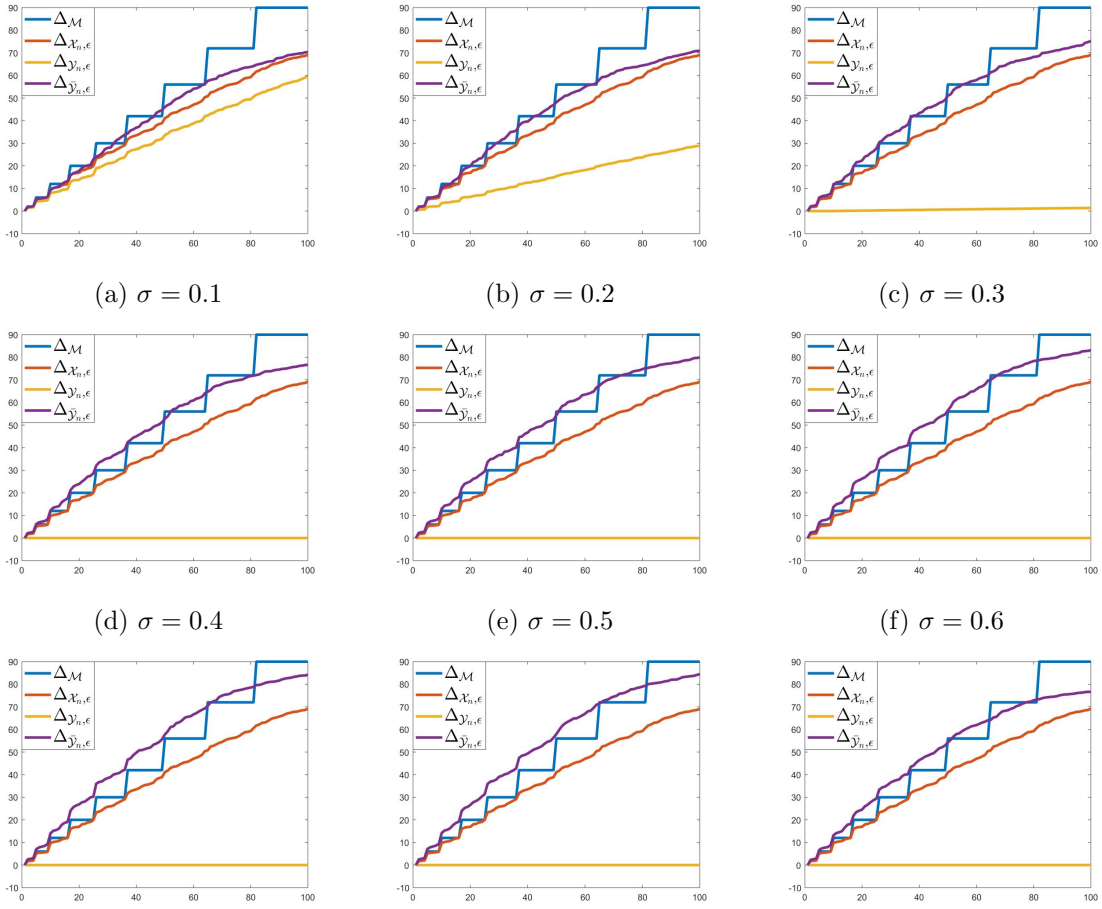
Table 1: Entrywise ∞-norm of $D_{\mathcal{X}_n} - D_{\mathcal{Y}_n}$ and $D_{\mathcal{X}_n} - D_{\bar{\mathcal{Y}}_n}$ on $\mathcal{S}$ for several σ's.

Next we study the spectral approximation of Laplacians by comparing the spectra of $\Delta_{\mathcal{X}_n,\varepsilon}$, $\Delta_{\mathcal{Y}_n,\varepsilon}$ with that of $\Delta_{\bar{\mathcal{Y}}_n,\varepsilon}$. Note that since the x_i are uniformly distributed, the density p on $\mathcal{S}$ that they are sampled from is constant and equal to $\frac{1}{\mathrm{vol}\mathcal{M}}$. So for the spectra of the graph Laplacians to match in scale with that of $\Delta_{\mathcal{S}}$, the weights should be rescaled according to

$$W(i,j) = \frac{2(m+2)\mathrm{vol}(\mathcal{M})}{\alpha_m \varepsilon^{m+2} n},$$

where $\mathrm{vol}(\mathcal{M})$ is the volume of the manifold and equals 4π in this case. In Figure 1 we compare the first 100 eigenvalues of $\Delta_{\mathcal{X}_n,\varepsilon}$, $\Delta_{\mathcal{Y}_n,\varepsilon}$, and $\Delta_{\bar{\mathcal{Y}}_n,\varepsilon}$ with the continuum spectrum. We see that when the noise size is large, the Euclidean graph Laplacian $\Delta_{\mathcal{Y}_n,\varepsilon}$ does not give a meaningful approximation of the continuum spectrum, while the locally regularized version $\Delta_{\bar{\mathcal{Y}}_n,\varepsilon}$ still performs well.

Remark 21 *While our theory in Section 2 suggests the choice $r \propto \sqrt{\sigma}$ in the small r and large n limit, for practical purposes some other scalings may give better results. Indeed for*



(a) $\sigma = 0.1$ (b) $\sigma = 0.2$ (c) $\sigma = 0.3$

(d) $\sigma = 0.4$ (e) $\sigma = 0.5$ (f) $\sigma = 0.6$

(g) $\sigma = 0.7$ (h) $\sigma = 0.8$ (i) $\sigma = 0.9$

Figure 1: Comparison of spectra of continuum Laplacian, $\Delta_{\mathcal{X}_n,\varepsilon}$, $\Delta_{\mathcal{Y}_n,\varepsilon}$ and $\Delta_{\bar{\mathcal{Y}}_n,\varepsilon}$ for different values of σ.

the above σ's, choosing $r = \sigma$ seems to give better spectral approximation. The choice of the local-regularization parameter will be further investigated in Subsection 4.2.2 in the context of a semi-supervised classification task, where a data-driven (cross-validation) approach can be used.

4.2 Classification

In this subsection we demonstrate the practical use of local regularization by applying it to classification problems. To show the potential benefits, we consider synthetic and real datasets, namely the two moons and MNIST datasets. Since in one of our experiments we study a real dataset, where in general the connectivity parameter in an ε graph is hard to tune, we instead consider fully connected graphs with self-tuning weights. Precisely, given a similarity $\delta : [n] \times [n] \to [0, \infty)$ we define, following Zelnik-Manor and Perona (2005), the

weights by

$$W(i,j) = \exp\left(-\frac{\delta(i,j)^2}{2\tau(i)\tau(j)}\right), \tag{31}$$

where $\tau(i)$ is the similarity between the i-th data point and its K-th nearest neighbor with respect to the distance δ. As before, we denote by $\Gamma_{\mathcal{X}_n}$, $\Gamma_{\mathcal{Y}_n}$ and $\Gamma_{\bar{\mathcal{Y}}_n}$ the graphs constructed with similarities $\delta_{\mathcal{X}_n}, \delta_{\mathcal{Y}_n}$, and $\delta_{\bar{\mathcal{Y}}_n}$. Instead of specifying a universal ε representing the connectivity length-scale, the neighborhood for each point is selected from using the local geometry which varies in space. It amounts to choosing different values of ε adaptively depending on the local scale, as proposed in Zelnik-Manor and Perona (2005). Since the $\tau(i)$ are defined by considering K-nearest neighbors, a natural variant of the above fully connected graph is to set the weights to be 0 whenever x_i and x_j are not among the K-nearest neighbors of each other. In other words, we can construct a (symmetrized) K-NN graph with the same K as in the definition of $\tau(i)$ and the nonzero weights are the same as above. It turns out that empirically this K-NN version can improve the classification performance substantially, but to illustrate the local regularization idea, we will present results for both graph constructions. We shall denote these two types of graphs as fully-connected and K-NN variants for brevity, or fully and K-NN for short.

In the following, we focus on the semi-supervised learning setting where we are given n data points with the first J being labeled. The classification is done by minimizing a probit functional as explained below. Let Δ_δ be a normalized graph Laplacian constructed on the dataset, which will be constructed using $\mathcal{X}_n$, $\mathcal{Y}_n$ and $\bar{\mathcal{Y}}_n$ and $\Delta_\delta = I - D^{-1/2}WD^{-1/2}$ as compared with (4). Let (λ_i, q_i), $i = 1, \ldots, n$ be the associated eigenvalue-eigenvector pairs, and let $U = \text{span}\{q_2, \ldots, q_n\}$. The classifier is set to be the sign of the minimizer u of the functional

$$\mathcal{J}(u) := \frac{1}{2c}\langle u, \Delta_\delta u\rangle - \sum_{j=1}^{J} \log\Big(\Phi(y(j)u(j);\gamma)\Big), \text{ with } c := n\Big(\sum_{i=2}^{n}\lambda_i^{-1}\Big)^{-1},$$

where $\{y(j)\}_{j=1}^J$ is the vector of labels and Φ is the cdf of $\mathcal{N}(0,\gamma^2)$. The functional $\mathcal{J}$ can be interpreted as the negative log posterior in a Bayesian setting, as discussed in Bertozzi et al. (2018). Throughout our experiments we set $\gamma = 0.1$.

4.2.1 Two Moons

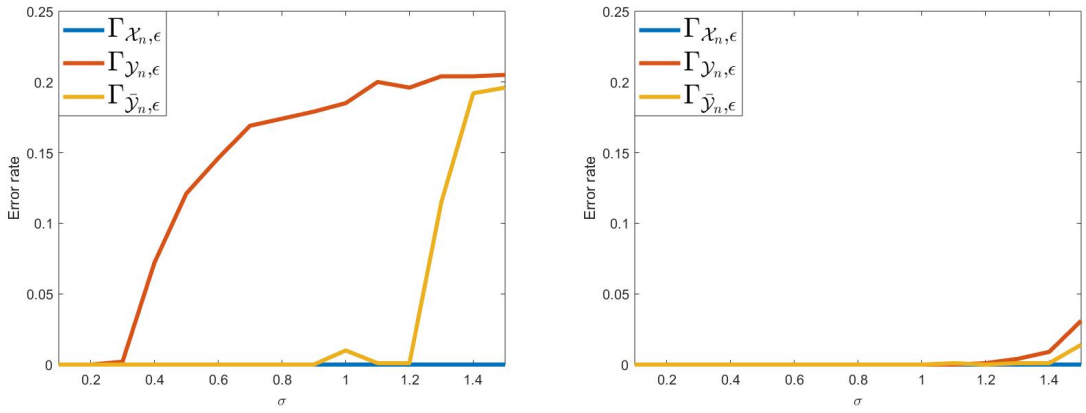
We first study the two moons dataset (Bühler and Hein (2009)), which is generated by sampling points uniformly from two semi-circles of unit radius centered at $(0,0)$ and $(1, 0.5)$ and then embedding the dataset in $\mathbb{R}^d$, with $d = 100$. We then perturb the data by adding uniform noise with norm bounded by σ. As before, the noise in the first two dimensions are normal to the semicirlces; the noise is taken to be uniform in the ambient space in the remaining dimensions. In addition to the semi-supervised setting, we also examine the unsupervised case.

We consider $n = 1000$ points 1% of which have labels and we set $K = 10$. As pointed out in Remark 21, we choose the regularization parameter r to be equal to σ. We compare the approximation of distance matrix and classification performance on $\mathcal{X}_n, \mathcal{Y}_n$, and $\bar{\mathcal{Y}}_n$'s,

as in Table 2 and Figure 4. Instead of comparing nodes that are within $\delta_{\mathcal{X}_n}$-distance ε, we consider nodes that are K-nearest neighbors of each other with respect to $\delta_{\mathcal{X}_n}$. As before, the regularized points $\bar{\mathcal{Y}}_n$ approximate the pairwise distances better and moreover, they improve the classification performance. Especially for the fully-connected case, we see that $\bar{\mathcal{Y}}_n$ is able to capture the exact correct labeling as the clean data does for moderate σ's, while the noisy data $\mathcal{Y}_n$ is making mistakes even when σ is as small as 0.3.

	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$	$\sigma = 0.7$	$\sigma = 0.9$
$\|D_{\mathcal{X}_n} - D_{\mathcal{Y}_n}\|_\infty$	0.109	0.344	0.589	0.795	0.996
$\|D_{\mathcal{X}_n} - D_{\bar{\mathcal{Y}}_n}\|_\infty$	0.064	0.164	0.240	0.372	0.431

Table 2: Entrywise ∞-norm of $D_{\mathcal{X}_n} - D_{\mathcal{Y}_n}$ and $D_{\mathcal{X}_n} - D_{\bar{\mathcal{Y}}_n}$ on two moons for different values of σ.

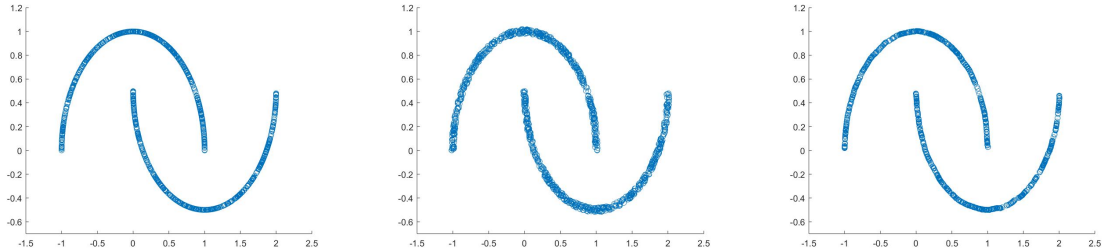


(a) Fully-connected. (b) K-NN variant.

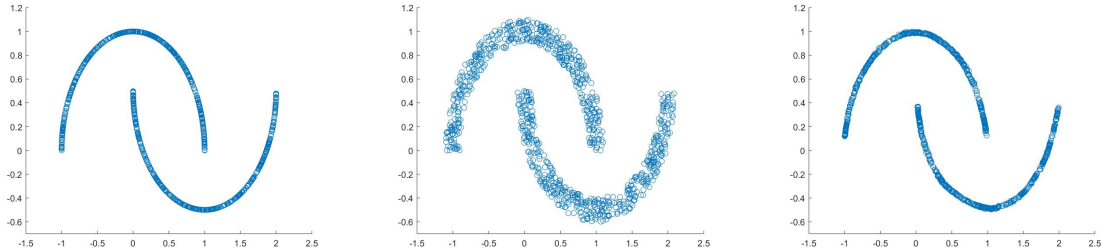
Figure 2: Classification error rates for $\Gamma_{\mathcal{X}_n}$, $\Gamma_{\mathcal{Y}_n}$ and $\Gamma_{\bar{\mathcal{Y}}_n}$ on two moons for different values of σ.

For further understanding, in Figure 3 we plot the first two coordinates of the points in $\mathcal{X}_n$, $\mathcal{Y}_n$ and $\bar{\mathcal{Y}}_n$ for large values of σ. We see that after local regularization, the first two coordinates of $\bar{\mathcal{Y}}_n$ lie almost on the underlying manifold. The denoising effect of local regularization is apparent. Furthermore, we observe that the semicircles for $\bar{\mathcal{Y}}_n$ are "shorter" than those of $\mathcal{X}_n$. In other words, points near the ends are pulled away from the boundaries. Moreover, if one looks carefully at the plots for $\bar{\mathcal{Y}}_n$, points are denser near the top and bottom. This illustrates that local regularization not only reduces noise, but also moves points to regions of high probability. We refer to Chen et al. (2016); Fukunaga and Hostetler (1975) and the references therein for some discussion on mean-shift and mode-seeking type algorithms.

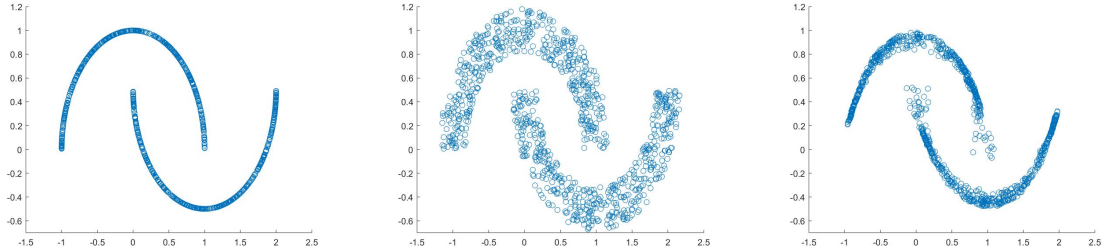
Remark 22 *The two moons dataset is sampled from a manifold with boundaries, and so our theory does not directly apply. However, the numerical results seem to suggest that our theory continues to hold in this setting.*



(a) $\sigma = 0.1$.



(b) $\sigma = 0.5$.



(c) $\sigma = 0.9$.

Figure 3: Visualization of the point clouds $\mathcal{X}_n$, $\mathcal{Y}_n$, and $\bar{\mathcal{Y}}_n$. Each row contains scatter plots of the first two coordinates of the points in the datasets $\mathcal{X}_n$, $\mathcal{Y}_n$, and $\bar{\mathcal{Y}}_n$.

Remark 23 *Additional numerical experiments not shown here suggest that applying local regularization within unsupervised spectral clustering gives qualitatively similar results to those shown in Figure 2 for a semi-supervised setting.*

4.2.2 MNIST

In this subsection we apply local regularization on the MNIST data-set of hand-written digits (LeCun (1998)). Each digit is described by a 784-dimensional vector, but the number of degrees of freedom of the data-generating mechanism is much smaller. For instance, in Hein and Audibert (2005) the authors estimate the intrinsic dimension of the digits 1 from MNIST to be 8. However, unlike in the previous examples, here there is no explicitly avail-

able underlying manifold from which the digits are sampled. Instead of adding additional noise to the dataset, we directly apply local regularization to the digits and show that doing so improves their binary classification. Since the level of noise is unknown choosing the localization parameter r cannot be guided by the theory and in our experiments, we tune it by performing 2-fold cross validation on the label sets. When there are few labels, we repeatedly generate holdout sets and compare the overall error. Due to this practical difficulty of tuning r, we propose two variants of $\Gamma_{\bar{\mathcal{Y}}_n}$ that can serve as alternatives in practice.

We study the classification performance of $\Gamma_{\bar{\mathcal{Y}}_n}$ for different pairs of digits. We consider a semi-supervised learning problem with $n = 1000$ images and $K = 20$ for the K-nearest neighbor variant. Table 3 shows the classification error percentage for four different pairs of digits when 4% of the digits are labeled. Table 4 shows the decrease on the classification error on the pair 4&9 as the percentage of labeled digits is increased.

Fully	3&8	5&8	4&9	7&9	K-NN	3&8	5&8	4&9	7&9
$\Gamma_{\mathcal{Y}_n}$	27.7%	48.0%	48.0%	48.0%	$\Gamma_{\mathcal{Y}_n}$	7.6%	5.5%	13.3%	7.3%
$\Gamma_{\bar{\mathcal{Y}}_n}$	13.4%	17.4%	30.0%	15.3%	$\Gamma_{\bar{\mathcal{Y}}_n}$	6.0%	3.6%	9.6%	5.4%

Table 3: Classification error for different pairs of digits 3&8, 5&8, 4&9, and 7&9.

Fully	4%	8%	12%	16%	K-NN	4%	8%	12%	16%
$\Gamma_{\mathcal{Y}_n}$	48.0%	42.7%	38.8%	29.4%	$\Gamma_{\mathcal{Y}_n}$	13.3%	10.9%	7.6%	5.1%
$\Gamma_{\bar{\mathcal{Y}}_n}$	30.0%	26.1%	21.9%	18.2%	$\Gamma_{\bar{\mathcal{Y}}_n}$	9.6%	6.4%	6.0%	4.5%

Table 4: Classification error for 4&9 with different number of labels.

(a) Threes in MNIST.

(b) Eights in MNIST.

Figure 4: Visualization of the regularization effects. The second row is the regularized version of the corresponding image in the first row. While arguably more blurred, the digits in the second row are more homogeneous within each group, making classification easier.

Again the K-NN variant performs much better than the fully-connected graph. As in Table 3, we see that except for the pair 3&8, the classification error for the other three pairs with $\Gamma_{\mathcal{Y}_n}$ is 480: after respecting the 40 labels, the other 960 images are classified as part of the same group. However, after regularization, the classification error is greatly reduced with $\Gamma_{\bar{\mathcal{Y}}_n}$. The same is true when we use the K-NN variant, but the improvement is smaller. Similarly as in Table 4, the improvement for local regularization becomes less dramatic as we go from the fully-connect graph to its K-NN variant and as the number of labels increases. This implies that there is certainly a limit for the improvement that local regularization can provide. Moreover, such improvement is most effective when label information is limited and one has to extract information from the geometry. Our theory and our experiments show that local regularization improves the recovery of geometric information and thereby

boosts the classification performance in that scenario. We present a visualization of the effect of local regularization in Figure 3. The two rows represent the image before and after local regularization respectively. We can see that especially for the eights, many of the images get "fixed" after regularization. Moreover, at a high level, images within each group in the second row look more similar among themselves than those in the first row. Because of this we expect the classification to be better.

Remark 24 *The four chosen pairs of digits are the hardest pairs to classify but local regularization can improve the performance for other pairs too. For unsupervised spectral clustering, local regularization still gives improvement, but using cross validation to choose r is no longer possible.*

4.3 Future Directions

As mentioned above, the practical choice of r can be challenging. We propose two alternatives that may be easier to work with and investigate their competence on the MNIST dataset.

4.3.1 k-NN regularization

This is a natural variant of $\Gamma_{\bar{\mathcal{Y}}_n}$ based on k-nearest neighbor regularization. Instead of specifying a neighborhood of y_i of radius r, we simply regress the data by averaging over its k nearest neighbors. Here k is not necessarily the same as K (the number of neighbors used to construct a similarity graph). Conceptually, choosing k amounts to setting different values of r at different points in such a way that the resulting neighborhoods contain roughly the same number of points. This construction is easier to work with since k is in general easier to tune than r.

4.3.2 Self-tuning regularization

This is a global regularization variant that does not require hyper parameters. Instead of averaging over a neighborhood of radius r, we take a global weighted average of the whole point cloud, where the weights are proportional to the similarities between the y_i. More specifically, we define a new distance in terms of the points $\hat{y}_i$, where

$$\hat{y}_i = \sum_{j=1}^{n} W(i,j) y_j,$$

and $W(i,j)$ is the defined as in (31). We see that points far from y_i have small contribution in the definition of $\hat{y}_i$ and so essentially one ends up summing over points in a neighborhood that is implicitly specified by the similarities. For points close to y_i, the weights are roughly on the same order. Hence $\hat{y}_i$ can be seen approximately as $\bar{y}_i$ plus a small contribution from points that are far from y_i. We expect this construction to behave a little worse than the $\Gamma_{\bar{\mathcal{Y}}_n}$ with optimal r. However, the fact that this construction does not require the tuning of any hyper-parameter makes it an appealing choice. Table 6 compares the classification performance of all graphs mentioned above (with the four different choices of distance function, and the two alternatives to build similarity graphs).

Fully	3&8	5&8	4&9	7&9	K-NN	3&8	5&8	4&9	7&9
$\Gamma_{\mathcal{Y}_n}$	27.7%	48.0%	48.0%	48.0%	$\Gamma_{\mathcal{Y}_n}$	7.6%	5.5%	12.8%	7.3%
$\Gamma_{\bar{\mathcal{Y}}_n}$	13.4%	17.4%	36.9%	15.3%	$\Gamma_{\bar{\mathcal{Y}}_n}$	6.9%	3.6%	9.7%	5.4%
k-NN	11.5%	7.4%	43.1%	18.3%	k-NN	5.3%	5.9%	9.6%	6.1%
self-tuning	16.1%	13.9%	33.4%	26.3%	self-tuning	7.6%	3.1%	8.8%	5.6%

Table 5: Comparison of classification errors with 4% labeled data.

Remark 25 *The idea of using labels to learn r (or k) can be understood as a specific instance of a more general idea: to use labels to better inform the learning of the underlying geometry of a dataset. What is more, one can try to simultaneously learn the geometry of the input space with the learning of the labeling function, instead of looking at these two problems in sequential form. This will be the topic of future research.*

Acknowledgments

The work of NGT and DSA was supported by the NSF Grant DMS-1912818/1912802. The authors are thankful to Facundo Mémoli for enlightening discussions, and two anonymous referees for their valuable feedback. Part of this manuscript was completed when the first author visited the Department of Statistics at the University of Chicago; NGT would like to thank the department for their hospitality.

Appendix A. Estimating r_- and r_+

A.1 Estimating r_-

We want to find values of $t > \frac{r}{2}$ for which for all $v \in T_{x_i}\mathcal{M}$ with $|v| \leq t$, and for all $\eta \in T_{\exp_{x_i}(v)}\mathcal{M}^\perp$ with $|\eta| \leq \sigma$ we have

$$|\exp_{x_i}(v) + \eta - y_i| < r.$$

We will later take the maximum value of t for which this holds and set r_- to be this maximum value.

Let $x = \exp_{x_i}(v)$. First, with the parallel transport map used in the proof of the geometric bias estimates (as in (25)) we can associate a vector $\tilde{\eta} \in T_{x_i}\mathcal{M}^\perp$ to a vector $\eta \in T_x\mathcal{M}^\perp$ with norm less than σ, for which

$$|\eta - \tilde{\eta}| \leq \frac{m}{R}\sigma t.$$

Now,

$$\begin{aligned}
|x + \eta - y_i| &\leq |x - x_i + \hat{\eta} - z_i| + |\eta - \hat{\eta}| \\
&= \left(|x - x_i|^2 + 2\langle x - x_i, \hat{\eta} - z_i\rangle + |\hat{\eta} - z_i|^2|\right)^{1/2} + |\eta - \hat{\eta}| \\
&\leq \left(|x - x_i|^2 + 2\langle x - x_i, \hat{\eta} - z_i\rangle + |\hat{\eta} - z_i|^2|\right)^{1/2} + \frac{m}{R}\sigma t.
\end{aligned}$$

We have

$$|x - x_i| \leq d_{\mathcal{M}}(x, x_i) = |v| \leq t,$$

and also

$$\langle x - x_i, \hat{\eta} - z_i \rangle = \langle x - (x_i + v), \hat{\eta} - z_i \rangle,$$

as it follows from the fact that $\eta, z_i \in T_{x_i}\mathcal{M}^{\perp}$ and $v \in T_{x_i}\mathcal{M}$. Using this, Cauchy-Schwartz, and (20) we conclude that

$$|\langle x - (x_i + v), \hat{\eta} - z_i \rangle| \leq 2\sigma |x - (x_i + v)| \leq 2\sigma \frac{|v|^2}{R},$$

and hence

$$\begin{aligned} |x + \eta - y_i| &\leq \left(t^2 + \frac{4}{R}\sigma t^2 + 4\sigma^2 \right)^{1/2} + \frac{m\sigma t}{R} \\ &= t \left(\sqrt{1 + \frac{4\sigma}{R} + \frac{4\sigma^2}{t^2}} + \frac{m\sigma}{R} \right) \\ &\leq t \left(\sqrt{1 + \frac{4\sigma}{R} + \frac{16\sigma^2}{r^2}} + \frac{m\sigma}{R} \right). \end{aligned}$$

From the above it follows that r_- defined as

$$r_- := r \left(\sqrt{1 + \frac{4\sigma}{R} + \frac{16\sigma^2}{r^2}} + \frac{m\sigma}{R} \right)^{-1}, \tag{32}$$

satisfies the desired properties and moreover

$$r - r_- \leq r \left(1 - \left(\sqrt{1 + \frac{4}{R}\sigma + \frac{16\sigma^2}{r^2}} + \frac{m\sigma}{R} \right)^{-1} \right). \tag{33}$$

A.2 Estimating r_+

To estimate r_+, we need the following lemma proved in García Trillos et al. (2018).

Lemma 26 *Suppose $x, \tilde{x} \in \mathcal{M}$ are such that $|x - \tilde{x}| \leq R/2$. Then*

$$|x - \tilde{x}| \leq d_{\mathcal{M}}(x, \tilde{x}) \leq |x - \tilde{x}| + \frac{8}{R}|x - \tilde{x}|^3. \tag{34}$$

To construct r_+ we find values of t with $2r \geq r + \sigma > t > 0$ such that if $|v| > t$ then

$$|\exp_{x_i}(v) + \sigma\eta - y_i| \geq r$$

for all $\eta \in T_{\exp_{x_i}(v)}\mathcal{M}^{\perp}$ of norm no larger than σ.

As in the construction of r_- we let $x := \exp_{x_i}(v)$. Similar computations give

$$\begin{aligned}
|x+\eta-y_i| &\geq |x-x_i+\hat{\eta}-z_i| - |\eta-\hat{\eta}| \\
&= \left(|x-x_i|^2 + 2\langle x-x_i, \hat{\eta}-z_i\rangle + |\hat{\eta}-z_i|^2|\right)^{1/2} - |\eta-\hat{\eta}| \\
&\geq \left(|x-x_i|^2 + 2\langle x-x_i, \hat{\eta}-z_i\rangle + |\hat{\eta}-z_i|^2|\right)^{1/2} - \frac{m}{R}\sigma|v| \\
&\geq \left(\left(|v| - \frac{1}{R}|v|^3\right)^2 - \frac{4}{R}\sigma|v|^2\right)^{1/2} - \frac{m}{R}\sigma|v| \\
&\geq |v|\left(\sqrt{1 - \frac{2|v|^2}{R} - \frac{4\sigma}{R}} - \frac{m\sigma}{R}\right) \\
&\geq |v|\left(\sqrt{1 - \frac{8r^2}{R} - \frac{4\sigma}{R}} - \frac{m\sigma}{R}\right) \\
&\geq t\left(\sqrt{1 - \frac{8r^2}{R} - \frac{4\sigma}{R}} - \frac{m\sigma}{R}\right),
\end{aligned}$$

where in the third inequality we have used (34) to conclude that

$$|\exp_{x_i}(v) - x_i| \geq d_{\mathcal{M}}(x, x_i) - C(d_{\mathcal{M}}(x, x_i))^3 = |v| - C|v|^3.$$

We can then take t to be such that the right hand side of (33) is equal to r. That is, we can take

$$r_+ := r\left(\sqrt{1 - \frac{8r^2}{R} - \frac{4\sigma}{R}} - \frac{m\sigma}{R}\right)^{-1}.$$

From these estimates we see that

$$r_+ - r_- \leq c\left(r^3 + r\sigma + \frac{\sigma^2}{r}\right).$$

References

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

E. Aamari, J. Kim, F Chazal, B Michel, A Rinaldo, L Wasserman, et al. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019.

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

A. L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):568–595, 2018.

T. Bühler and M. Hein. Spectral clustering based on the graph p-Laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 81–88. ACM, 2009.

D. Burago, S. Ivanov, and Y. Kurylev. A graph discretization of the Laplace-Beltrami operator. *J. Spectr. Theory*, 4:675–714, 2014.

Y.-C. Chen, C. R. Genovese, L. Wasserman, et al. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

M. P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. ISBN 0-8176-3490-8. doi: 10.1007/978-1-4757-2201-7. URL `https://doi.org/10.1007/978-1-4757-2201-7`. Translated from the second Portuguese edition by Francis Flaherty.

H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93 (3):418–491, 1959.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1): 32–40, 1975.

N. Garcia Trillos and D. Sanz-Alonso. Continuum limits of posteriors in graph bayesian inverse problems. *SIAM Journal on Mathematical Analysis*, 50(4):4020–4040, 2018.

N. García Trillos and D. Slepčev. Continuum limit of total variation on point clouds. *Archive for rational mechanics and analysis*, 220(1):193–241, 2016.

N. Garcia Trillos, Z. Kaplan, T. Samakhoana, and D. Sanz-Alonso. On the consistency of graph-based Bayesian learning and the scalability of sampling algorithms. *arXiv preprint arXiv:1710.07702*, 2017.

N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev. Spectral convergence of empirical graph Laplacians. *Preprinr*, 2018.

C. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *Journal of machine learning research*, 13(May):1263–1291, 2012.

A. Haddad, D. Kushnir, and R. R. Coifman. Texture separation via a reference set. *Applied and Computational Harmonic Analysis*, 36(2):335–347, 2014.

M. Hein and J. Audibert. Intrinsic dimensionality estimation of submanifolds in r d. In *Proceedings of the 22nd international conference on Machine learning*, pages 289–296. ACM, 2005.

M. Hein and M. Maier. Manifold denoising. In *Advances in Neural Information Processing Systems 19*, pages 561–568, Cambridge, MA, USA, September 2007. Max-Planck-Gesellschaft, MIT Press.

Y. LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

E. H. Lieb, M. Loss, et al. Graduate studies in mathematics. *Analysis*, 14, 2001.

A. V. Little and L. Maggioni, M.and Rosasco. Multiscale geometric methods for data sets i: Multiscale svd, noise and curvature. *Applied and Computational Harmonic Analysis*, 43(3):504–567, 2017.

F. Mémoli, Z. T. Smith, and Z. Wan. The wasserstein transform. *CoRR*, abs/1810.07793, 2018. URL `http://arxiv.org/abs/1810.07793`.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

P. J. Olver. *Introduction to Partial Differential Equations*. Springer.

I. Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. In *Probability in Banach spaces, 8 (Brunswick, ME, 1991)*, volume 30 of *Progr. Probab.*, pages 128–134. Birkhäuser Boston, Boston, MA, 1992.

J. Shi and J. Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.

A. Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

D. A Spielman and S.-H. Teng. Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra and its Applications*, 421(2-3):284–305, 2007.

J. W. Tukey and P. A. Tukey. Computer graphics and exploratory data analysis: An introduction. *The Collected Works of John W. Tukey: Graphics: 1965-1985*, 5:419, 1988.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

U. Von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

W. Wang and M. A. Carreira-Perpinán. Manifold blurring mean shift algorithms for manifold denoising. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1759–1766. IEEE, 2010.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2005.

D. Zhou and B. Schölkopf. Regularization on discrete spaces. In *Joint Pattern Recognition Symposium*, pages 361–368. Springer, 2005.

X. Zhu. Semi-supervised learning literature survey. 2005.