On the Convergence of Maximum Variance Unfolding

Ery Arias-Castro   EARIASCA@MATH.UCSD.EDU
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093, USA

Bruno Pelletier   BRUNO.PELLETIER@UNIV-RENNES2.FR
Département de Mathématiques
IRMAR – UMR CNRS 6625
Université Rennes II, France

Abstract

Maximum Variance Unfolding is one of the main methods for (nonlinear) dimensionality reduction. We study its large sample limit, providing specific rates of convergence under standard assumptions. We find that it is consistent when the underlying submanifold is isometric to a convex subset, and we provide some simple examples where it fails to be consistent.

Keywords: maximum variance unfolding, isometric embedding, U-processes, empirical processes, proximity graphs.

1. Introduction

One of the basic tasks in unsupervised learning, aka multivariate statistics, is that of dimensionality reduction. While the celebrated Principal Components Analysis (PCA) and Multidimensional Scaling (MDS) assume that the data lie near an affine subspace, modern approaches postulate that the data are in the vicinity of a submanifold. Many such algorithms have been proposed in the past decade, for example, Isomap (Tenenbaum et al., 2000), Local Linear Embedding (LLE) (Roweis and Saul, 2000), Laplacian Eigenmaps (Belkin and Niyogi, 2003), Manifold Charting (Brand, 2003), Diffusion Maps (Coifman and Lafon, 2006), Hessian Eigenmaps (HLLE) (Donoho and Grimes, 2003), Local Tangent Space Alignment (LTSA) (Zhang and Zha, 2004), Maximum Variance Unfolding (Weinberger et al., 2004), and many others, some reviewed in Van der Maaten et al. (2008) and Saul et al. (2006).

Although some variants exist, the basic setting is that of a connected domain $D \subset \mathbb{R}^d$ isometrically embedded in Euclidean space as a submanifold $M \subset \mathbb{R}^p$, with $p > d$. We are provided with data points $x_1, \ldots, x_n \in \mathbb{R}^p$ sampled from (or near) M and our goal is to output $y_1, \ldots, y_n \in \mathbb{R}^d$ that can be isometrically mapped to (or close to) $x_1, \ldots, x_n$.

A number of consistency results exist in the literature. For example, Bernstein et al. (2000) show that, with proper tuning, geodesic distances may be approximated by neighborhood graph distances when the submanifold M is geodesically convex, implying that Isomap asymptotically recovers the isometry when D is convex. When D is not convex, it fails in general (Zha and Zhang, 2003). Very close in spirit to what we do here, Zha and Zhang (2007) introduce and study a continuum version of Isomap. In accordance with the discrete version, they show that their Continuum Isomap is able to

recover an isometry when the manifold is isometric to a convex domain in some lower-dimensional Euclidean space. To justify HLLE, Donoho and Grimes (2003) show that the null space of the (continuous) Hessian operator yields an isometric embedding. See also Ye and Zhi (2012) for related results in a discrete setting. Smith et al. (2008) prove that LTSA is able to recover the isometry, but only up to an affine transformation. We also mention other results in the literature which show that, as the sample size increases, the output the algorithm converges to is an explicit continuous embedding. For instance, a number of papers analyze how well the discrete graph Laplacian based on a sample approximates the continuous Laplace-Beltrami operator on a submanifold (Belkin and Niyogi, 2005; von Luxburg et al., 2008; Singer, 2006; Hein et al., 2005; Giné and Koltchinskii, 2006; Coifman and Lafon, 2006), which is intimately related to the Laplacian Eigenmaps. However, such convergence results do not guaranty that the algorithm is successful at recovering the isometry when one exists. In fact, as discussed in detail by Goldberg et al. (2008) and Perrault-Joncas and Meila (2012), many of them fail in very simple settings.

In this paper, we analyze Maximum Variance Unfolding (MVU) in the large-sample limit. We are only aware of a very recent work of Paprotny and Garcke (2012) that establishes that, under the assumption that D is convex, MVU recovers a distance matrix that approximates the geodesic distance matrix of the data. Our contribution is the following. In Section 2, we prove a convergence result, showing that the optimization problem that MVU solves converges (both in solution space and value) to a continuous version defined on the whole submanifold. The basic assumption here is that the submanifold M is compact. In Section 3, we derive quantitative convergence rates, with mild additional regularity assumptions. In Section 4, we consider the solutions to the continuum limit. When D is convex, we prove that MVU recovers an isometry. We also provide examples of non-convex D where MVU provably fails at recovering an isometry. We also prove that MVU is robust to noise, which Goldberg et al. (2008) show to be problematic for algorithms like LLE, HLLE and LTSA. Some concluding remarks are in Section 5.

2. From Discrete MVU to Continuum MVU

In this section we state and prove a qualitative convergence result for MVU. This result applies with only minimal assumptions and its proof is relatively transparent. What we show is that the (discrete) MVU optimization problem converges to an explicit continuous optimization problem when the sample size increases. Although this convergence does not imply a dimensionality reduction per se, the continuous optimization problem is amenable to scrutiny with tools from analysis and geometry, and that will enable us to better understand (in Section 4) when MVU succeeds, and when it fails, at recovering an isometry to a Euclidean domain when it exists.

Let us start by recalling the MVU algorithm (Weinberger and Saul, 2006; Weinberger et al., 2004, 2005). We are provided with data points $x_1, \ldots, x_n \in \mathbb{R}^p$. Let $\|\cdot\|$ denote the Euclidean norm. Let $\mathcal{Y}_{n,r}$ be the (random) set defined by

$$\mathcal{Y}_{n,r} = \left\{ y_1, \ldots, y_n \in \mathbb{R}^p : \|y_i - y_j\| \leq \|x_i - x_j\| \text{ when } \|x_i - x_j\| \leq r \right\}.$$

Having chosen a neighborhood radius $r > 0$, MVU solves the following optimization problem:

$$\text{Maximize} \quad \mathcal{E}(Y) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \|y_i - y_j\|^2, \quad \text{over } Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^{n \times p},$$

$$\text{subject to} \quad Y \in \mathcal{Y}_{n,r}. \tag{1}$$

When the data points are sampled from a distribution μ with support M, our main result in this section is to show that, when M is sufficiently regular and $r = r_n \to 0$ sufficiently slowly, the discrete optimization problem converges to the following continuous optimization problem:

CONTINUUM MVU

$$\text{Maximize} \quad \mathcal{E}(f) := \int_{M \times M} \|f(x) - f(x')\|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x'), \quad \text{over } f : M \to \mathbb{R}^p,$$

$$\text{subject to} \quad f \text{ is Lipschitz with } \|f\|_{\text{Lip}} \leq 1.$$

$\|f\|_{\text{Lip}}$ denotes the smallest Lipschitz constant of a function f. It is important to realize that the Lipschitz condition is with respect to the intrinsic metric on M (i.e., the metric inherited from the ambient space $\mathbb{R}^p$), defined as follows: for $x, x' \in M$, let

$$\delta_M(x, x') = \inf\{T : \exists \gamma : [0, T] \to M, \ 1\text{-Lipschitz, with } \gamma(0) = x \text{ and } \gamma(T) = x'\}.$$

When M is compact, the infimum is attained. In that case, $\delta_M(x, x')$ is the length of the shortest continuous path on M starting at x and ending at x', and (M, δ_M) is a complete metric space, also called a *length space* in the context of metric geometry (Burago et al., 2001). Then $f : M \to \mathbb{R}^p$ is Lipschitz with $\|f\|_{\text{Lip}} \leq L$ if

$$\|f(x) - f(x')\| \leq L \delta_M(x, x'), \ \forall x, x' \in M. \tag{2}$$

For any $L > 0$, denote by $\mathcal{F}_L$ the class of Lipschitz functions $f : M \to \mathbb{R}^p$ satisfying (2).

One of the central condition is that M is sufficiently regular that the intrinsic metric on M is locally close to the ambient Euclidean metric.

Regularity assumption. There is a non-decreasing function $c : [0, \infty) \to [0, \infty)$ such that $c(r) \to 0$ when $r \to 0$, such that, for all $x, x' \in M$,

$$\delta_M(x, x') \leq \big(1 + c(\|x - x'\|)\big)\|x - x'\|. \tag{3}$$

This assumption is also central to Isomap. Bernstein et al. (2000) prove that it holds when M is a compact, smooth and geodesically convex submanifold (e.g., without boundary). In Lemma 4, we extend this to compact, smooth submanifolds with smooth boundary, and to tubular neighborhoods of such sets. The latter allows us to study noisy settings.

Note that we always have

$$\|x - x'\| \leq \delta_M(x, x'). \tag{4}$$

Let $\mathcal{S}_1$ denote the set of functions that are solutions of Continuum MVU. We state the following qualitative result that makes minimal assumptions.

Theorem 1 *Let μ be a (Borel) probability distribution with support $M \subset \mathbb{R}^p$, which is connected, compact and satisfying (3), and assume that $x_1, \ldots, x_n$ are sampled independently from μ. Then, for $r_n \to 0$ sufficiently slowly, we have*

$$\sup\{\mathcal{E}(Y) : Y \in \mathcal{Y}_{n,r_n}\} \to \sup\{\mathcal{E}(f) : f \in \mathcal{F}_1\}, \tag{5}$$

and for any solution $\hat{Y}_n = (\hat{y}_1, \ldots, \hat{y}_n)$ of Discrete MVU,

$$\inf_{f \in \mathcal{S}_1} \max_{1 \le i \le n} \|\hat{y}_i - f(x_i)\| \to 0, \tag{6}$$

almost surely as $n \to \infty$.

Thus Discrete MVU converges to Continuum MVU in the large sample limit, if M satisfies the crucial regularity condition (3) and other mild assumptions. In Section 3, we provide explicit quantitative bounds for the convergence results (5) and (6) at the very end, under some additional (though natural) assumptions. In Section 4, we focus entirely on Continuum MVU, with the goal of better understanding the functions that are solutions to that optimization problem. Because of (6), we know that the output of Discrete MVU converges in a strong sense to one of these functions.

The rest of the section is dedicated to proving Theorem 1. We divide the proof into several parts which we discuss at length, and then assemble to prove the theorem.

2.1 Coverings and Graph Neighborhoods

For $r > 0$, let G_r denote the undirected graph with nodes $x_1, \ldots, x_n$ and an edge between x_i and x_j if $\|x_i - x_j\| \le r$. This is the r-neighborhood graph based on the data. Remember that we consider Discrete MVU with $r = r_n \to 0$ as the sample size $n \to \infty$. For the result to hold, it is essential that G_{r_n} be connected, for otherwise $\sup\{\mathcal{E}(Y) : Y \in \mathcal{Y}_{n,r_n}\} = \infty$, while $\sup\{\mathcal{E}(f) : f \in \mathcal{F}_1\}$ is finite. The latter comes from the fact that, for any $f \in \mathcal{F}_1$,

$$\mathcal{E}(f) \le \int_{M \times M} \delta_M(x, x')^2 \mu(dx) \mu(dx') \le \mathrm{diam}(M)^2,$$

where we used (2) in the first inequality, and $\mathrm{diam}(M)$ is the intrinsic diameter of M, that is,

$$\mathrm{diam}(M) := \sup_{x, x' \in M} \delta_M(x, x').$$

Recall that the only assumptions on M made in Theorem 1 are that M is compact, connected, and satisfies (3), and this implies that $\mathrm{diam}(M) < \infty$. Indeed, as a compact subset of $\mathbb{R}^p$, M is bounded, hence $\sup_{x, x' \in M} \|x - x'\| < \infty$. Reporting this in (3) immediately implies that $\mathrm{diam}(M) < \infty$.

That said, we ask more of (r_n) than simply having G_{r_n} connected. For $\eta > 0$, define

$$\Lambda(\eta) = \{\forall x \in M, \exists i = 1, \ldots, n : \|x - x_i\| \le \eta\},$$

which is the event that $x_1, \ldots, x_n$ forms an η-covering of M.

Connectivity requirement. $r_n \to 0$ in such a way that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\Lambda(\lambda_n r_n)^{\mathsf{c}}\right) < \infty, \text{ for some sequence } \lambda_n \to 0. \tag{7}$$

Since M is the support of μ, there is always a sequence (r_n) that satisfy the Connectivity requirement. For $\eta > 0$, we say that $z_1, \ldots, z_N \in M$ is η-packing (for the Euclidean metric) if it is a maximal collection of points such that $\|z_i - z_j\| > \eta$ for all $i \neq j$, meaning there is no other point $z \in M$ such that $\|z - z_i\| > \eta$ for all i. By definition then, an η-packing $z_1, \ldots, z_N \in M$ is also an η-covering, in the sense that, for all $z \in M$, there is $i \in \{1, \ldots, N\}$ such that $\|z - z_i\| \leq \eta$. (Note that not all η-coverings are η-packings.) Let N_η denote the maximum cardinality of an η-packing of M. By compactness of M, $N_\eta < \infty$ for any $\eta > 0$. Let $p_\eta = \min_j \mu(B(z_j, \eta))$, where $B(z, \eta)$ denotes the Euclidean ball centered at z and of radius $\eta > 0$. Since M is the support of μ, $\mu(B(z, \eta)) > 0$ for any $z \in M$ and any $\eta > 0$. Hence, $p_\eta > 0$ for any $\eta > 0$. We have

$$
\begin{aligned}
\mathbb{P}\left(\Lambda(2\eta)^c\right) &= \mathbb{P}\left(\text{there exists } x \in M : \forall i = 1, \ldots, n, \|x - x_i\| > 2\eta\right) \\
&\leq \mathbb{P}\left(\text{there is } j \text{ such that } B(z_j, \eta) \text{ is empty of data points}\right) \\
&\leq \sum_{j=1}^{N_\eta} \mathbb{P}(B(z_j, \eta) \text{ is empty of data points}) \\
&\leq N_\eta (1 - p_\eta)^n.
\end{aligned}
$$

Let $\eta_n = \inf\{\eta > 0 : N_\eta(1 - p_\eta)^n \leq 1/n^2\}$; the sequence $1/n^2$ is chosen here for the simplicity of the exposition, but more general sequence can be considered, as will become apparent at the end of the paragraph.

Since $p_\eta > 0$ for all $\eta > 0$, $\eta_n \to 0$. To see this, let $\eta^\star = \mathrm{diam}(M)$. Clearly, for all $\eta \geq \eta^\star$, $p_\eta = 1$, which implies that the set of $\eta > 0$ such that $N_\eta(1 - p_\eta)^n \leq 1/n^2$ is non-empty. In particular, for all $n \geq 1$, we have $\eta_n \leq \eta^\star$. Now, let $\varepsilon > 0$ be fixed. Since $p_\varepsilon > 0$, there exists an integer n_ε such that $N_\varepsilon(1 - p_\varepsilon)^n \leq 1/n^2$ for all $n \geq n_\varepsilon$, so that $\eta_n \leq \varepsilon$ for all $n \geq n_\varepsilon$. Since ε is arbitrary, this proves that the sequence (η_n) converges to 0 as n tends to infinity.

With such a choice of (η_n), we have $\sum_{n \geq 1} \mathbb{P}(\Lambda(2\eta_n)^c) \leq \sum_{n \geq 1} 1/n^2 < \infty$. Therefore, if we take $r_n = \sqrt{\eta_n}$, it satisfies the Connectivity requirement. In Section 3.2 we derive a quantitative bound on r_n that guaranty (7) under additional assumptions. Note that the sequence $(1/n^2)$ in the definition of η_n can be replaced by any summable decreasing sequence.

The rationale behind the requirement on (r_n) is the same as in Bernstein et al. (2000): it allows to approximate each curve on M with a path in G_{r_n} of nearly the same length. We use this in the following subsection.

2.2 Interpolation

Assuming that the sampling is dense enough that $\Lambda(\eta)$ holds, we interpolate a set of vectors $Y \in \mathcal{Y}_{n,r}$ with a Lipschitz function $f \in \mathcal{F}_{1+O(\eta/r)}$. Formally, we have the following.

Lemma 1 *Assume that $\Lambda(\eta)$ holds for some $\eta \leq r/4$. Then any vector $Y = (y_1, \ldots, y_n) \in \mathcal{Y}_{n,r}$ is of the form $Y = (f(x_1), \ldots, f(x_n))$ for some $f \in \mathcal{F}_{1+6\eta/r}$.*

We prove this result. The first step is to show that this is at all possible in the sense that

$$\|y_i - y_j\| \leq \left(1 + 6\eta/r\right) \delta_M(x_i, x_j), \ \forall i, j. \tag{8}$$

This shows that the map $g : \{x_1, \ldots, x_n\} \to \mathbb{R}^p$ defined by $g(x_i) = y_i$ for all i, is Lipschitz (for δ_M and the Euclidean metrics) with constant $L = 1 + 6\eta/r$. We apply a form of Kirszbraun's Extension—Lang and Schroeder (1997, Theorem B) or Brudnyi and Brudnyi (2012, Theorem 1.26)—to extend g to the whole M into $f \in \mathcal{F}_{1+6\eta/r}$.

Therefore, let's turn to proving (8). The arguments are very similar to those in Bernstein et al. (2000). If $\delta_M(x_i, x_j) \leq r$, then, by (4), $\|x_i - x_j\| \leq r$, which implies that

$$\|y_i - y_j\| \leq \|x_i - x_j\| \leq \delta_M(x_i, x_j).$$

Now suppose that $\delta_M(x_i, x_j) > r$. Let γ be a path in M connecting x_i to x_j of minimal length $l = \delta_M(x_i, x_j)$. Split γ into N arcs of lengths $l_1 = r/2$ plus one arc of length $l_{N+1} < l_1$, so that

$$\frac{l}{l_1} - 1 \leq N \leq \frac{l}{l_1}.$$

Denote by $x_i = x'_0, x'_1, \ldots, x'_N, x'_{N+1} = x_j$ the extremities of the arcs along γ.

For $k = 1, \ldots, N$, let $t_k \in \arg\min_t \|x'_k - x_t\|$. Assuming $\Lambda_n(\eta)$ holds, for all k we have $\delta_M(x'_k, x_{t_k}) \leq \eta$, so that

$$\|x_{t_k} - x_{t_{k-1}}\| \leq \delta_M(x_{t_k}, x_{t_{k-1}}) \leq \delta_M(x'_k, x'_{k-1}) + 2\eta \leq l_1 + 2\eta \leq r/2 + 2(r/4) = r.$$

Hence, because $Y = (y_1 \ldots, y_n) \in \mathcal{Y}_{n,r}$,

$$\|y_{t_k} - y_{t_{k-1}}\| \leq l_1 + 2\eta.$$

Similarly, for the last arc, recalling that $x_{t_{N+1}} = x_j$, we have $\delta_M(x_j, x_{t_N}) \leq l_{N+1} + \eta < l_1 + \eta < r$, and therefore

$$\|y_{t_{N+1}} - y_{t_N}\| \leq l_{N+1} + \eta.$$

Consequently,

$$\begin{aligned}\|y_i - y_j\| &\leq N(l_1 + 2\eta) + (l_{N+1} + \eta)\\ &= Nl_1 + l_{N+1} + (2N+1)\eta\\ &= l + (2N+1)\eta.\end{aligned}$$

We have

$$(2N+1)\eta \leq \left(2\frac{l}{l_1} + 1\right)\eta \leq l\frac{3\eta}{l_1} = l\frac{6\eta}{r},$$

and so (8) holds.

2.3 Bounds on The Energy

We call $\mathcal{E}$ the energy functional. For a function $f : \{x_1, \ldots, x_n\} \to \mathbb{R}^p$, let $Y_n(f) = (f(x_1), \ldots, f(x_n))^T \in \mathbb{R}^{n \times p}$. Assume that $\Lambda(\eta)$ holds $\eta \leq r/4$. Then Lemma 1 implies that any $Y \in \mathcal{Y}_{n,r}$ is equal to $Y_n(f)$ for some $f \in \mathcal{F}_{1+6\eta/r}$. Hence,

$$\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) \leq \sup_{f \in \mathcal{F}_{1+6\eta/r}} \mathcal{E}(Y_n(f)). \tag{9}$$

Recall the function $c(r)$ introduced in (3), and assume that $r > 0$ is small enough that $c(r) < 1$. For $f \in \mathcal{F}_{1-c(r)}$, and for any i, j such that $\|x_i - x_j\| \leq r$, we have

$$\|f(x_i) - f(x_j)\| \leq (1 - c(r))\delta_M(x_i, x_j) \leq (1 - c(r))(1 + c(\|x_i - x_j\|))\|x_i - x_j\|.$$

Since the function c is non-decreasing, $c(\|x_i - x_j\|) \leq c(r)$, and so

$$\|f(x_i) - f(x_j)\| \leq \left(1 - c(r)^2\right) \|x_i - x_j\| \leq \|x_i - x_j\|.$$

Consequently, $Y_n(f) \in \mathscr{Y}_{n,r}$, implying that

$$\sup_{Y \in \mathscr{Y}_{n,r}} \mathcal{E}(Y) \geq \sup_{f \in \mathscr{F}_{1-c(r)}} \mathcal{E}(Y_n(f)). \tag{10}$$

As a result of (9) and (10), we have

$$\left| \sup_{Y \in \mathscr{Y}_{n,r}} \mathcal{E}(Y) - \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \right| \leq \sup_{1-c(r) \leq L \leq 1+6\eta/r} \left| \sup_{f \in \mathscr{F}_L} \mathcal{E}(Y_n(f)) - \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \right|. \tag{11}$$

We have

$$\left| \sup_{f \in \mathscr{F}_L} \mathcal{E}(Y_n(f)) - \sup_{f \in \mathscr{F}_L} \mathcal{E}(f) \right| \leq \sup_{f \in \mathscr{F}_L} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right|,$$

and applying the triangle inequality, we arrive at

$$\left| \sup_{f \in \mathscr{F}_L} \mathcal{E}(Y_n(f)) - \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \right| \leq \sup_{f \in \mathscr{F}_L} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| + \left| \sup_{f \in \mathscr{F}_L} \mathcal{E}(f) - \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \right|.$$

Since $\mathscr{F}_L = L\mathscr{F}_1$ and $\mathcal{E}(Lf) = L^2 \mathcal{E}(f)$, we have

$$\left| \sup_{f \in \mathscr{F}_L} \mathcal{E}(f) - \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \right| \leq |L^2 - 1| \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \leq |L^2 - 1| \operatorname{diam}(M)^2,$$

and

$$\sup_{f \in \mathscr{F}_L} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| = L^2 \sup_{f \in \mathscr{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right|. \tag{12}$$

Consequently,

$$\left| \sup_{f \in \mathscr{F}_L} \mathcal{E}(Y_n(f)) - \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \right| \leq L^2 \sup_{f \in \mathscr{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| + |L^2 - 1| \operatorname{diam}(M)^2.$$

Reporting this inequality in (11) on the event $\Lambda(\eta)$ with $\eta \leq r/4$, we have

$$\left| \sup_{Y \in \mathscr{Y}_{n,r}} \mathcal{E}(Y) - \sup_{f \in \mathscr{F}_1} \mathcal{E}(f) \right| \leq (1 + 6\eta/r)^2 \sup_{f \in \mathscr{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| + \beta(r,\eta)\left(2 + \beta(r,\eta)\right) \operatorname{diam}(M)^2, \tag{13}$$

where $\beta(r,\eta) := \max(c(r), 6\eta/r)$.

Finally, we show that $\mathcal{E}$ is continuous (in fact Lipschitz) on $\mathscr{F}_1$ for the supnorm. For any f and g in $\mathscr{F}_1$, and any x and x' in M, we have:

$$
\begin{aligned}
& \left| \|f(x) - f(x')\|^2 - \|g(x) - g(x')\|^2 \right| \\
\leq\ & \|f(x) - f(x') - g(x) + g(x')\| \, \|f(x) - f(x') + g(x) - g(x')\| \\
\leq\ & \left[\|f(x) - g(x)\| + \|f(x') - g(x')\| \right] \left[\|f(x) - f(x')\| + \|g(x) - g(x')\| \right] \\
\leq\ & 4\|f - g\|_\infty \operatorname{diam}(M).
\end{aligned}
$$

The first inequality is that of Cauchy-Schwarz. Hence,

$$\left| \mathcal{E}(f) - \mathcal{E}(g) \right| \leq 4\|f - g\|_\infty \operatorname{diam}(M), \tag{14}$$

and

$$\left| \mathcal{E}(Y_n(f)) - \mathcal{E}(Y_n(g)) \right| \leq 4\|f - g\|_\infty \operatorname{diam}(M). \tag{15}$$

2.4 More Coverings and the Law of Large Numbers

The last step is to show that the supremum of the empirical process (12) converges to zero. For this, we use a packing (covering) to reduce the supremum over $\mathcal{F}_1$ to a maximum over a finite set of functions. We then apply the Law of Large Numbers to each difference in the maximization.

Fix $x_0 \in M$ and define

$$\mathcal{F}_1^0 = \{f \in \mathcal{F}_1 : f(x_0) = 0\}.$$

Note that $f \in \mathcal{F}_1$ if, and only if, $f - f(x_0) \in \mathcal{F}_1^0$, and by the fact that $\mathcal{E}(f + a) = \mathcal{E}(f)$ for any function or vector f and any constant $a \in \mathbb{R}^p$, we have

$$\sup_{f \in \mathcal{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| = \sup_{f \in \mathcal{F}_1^0} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right|.$$

The reason to use $\mathcal{F}_1^0$ is that it is bounded in supnorm. Indeed, for $f \in \mathcal{F}_1^0$, we have

$$\|f(x)\| = \|f(x) - f(x_0)\| \leq \delta_M(x, x_0) \leq \operatorname{diam}(M), \ \forall x \in M.$$

Let $\mathcal{N}_\infty(\mathcal{F}_1^0, \varepsilon)$ denote the covering number of $\mathcal{F}_1^0$ for the supremum norm, that is, the minimal number of balls that are necessary to cover $\mathcal{F}_1^0$, and let $f_1, \ldots, f_N \in \mathcal{F}_1$ be an ε-covering of $\mathcal{F}_1^0$ of minimal size $N := \mathcal{N}_\infty(\mathcal{F}_1^0, \varepsilon)$. Since $\mathcal{F}_1^0$ is equicontinuous and bounded, it is compact for the topology of the supremum norm by the Arzelà-Ascoli Theorem, so that $\mathcal{N}_\infty(\mathcal{F}_1^0, \varepsilon) < \infty$ for any $\varepsilon > 0$.

Fix $f \in \mathcal{F}_1^0$ and let k be such that $\|f - f_k\| \leq \varepsilon$. By (14) and (15), we have

$$
\begin{aligned}
\left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| &\leq \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(Y_n(f_k)) \right| + \left| \mathcal{E}(Y_n(f_k)) - \mathcal{E}(f_k) \right| + \left| \mathcal{E}(f_k) - \mathcal{E}(f) \right| \\
&\leq 8 \operatorname{diam}(M) \|f - f_k\|_\infty + \left| \mathcal{E}(Y_n(f_k)) - \mathcal{E}(f_k) \right| \\
&= 8 \operatorname{diam}(M)\varepsilon + \left| \mathcal{E}(Y_n(f_k)) - \mathcal{E}(f_k) \right|.
\end{aligned}
$$

Thus,

$$\sup_{f \in \mathcal{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| \leq 8 \operatorname{diam}(M)\varepsilon + \max\{ \left| \mathcal{E}(Y_n(f_k)) - \mathcal{E}(f_k) \right| : k = 1, \ldots, \mathcal{N}_\infty(\mathcal{F}_1^0, \varepsilon) \}. \quad (16)$$

The Law of Large Numbers (LLN) imply that, for any bounded f, $\mathcal{E}(Y_n(f)) \to \mathcal{E}(f)$, almost surely as $n \to \infty$. Indeed,

$$
\begin{aligned}
\mathcal{E}(Y_n(f)) &= \frac{n^2}{n(n-1)} \frac{1}{n^2} \sum_{i,j} \|f(x_i) - f(x_j)\|^2 \\
&= \frac{2n}{n-1} \left[\frac{1}{n} \sum_i \|f(x_i)\|^2 - \left\| \frac{1}{n} \sum_i f(x_i) \right\|^2 \right] \\
&\to 2\mathbb{E}\|f(x)\|^2 - 2\|\mathbb{E}f(x)\|^2 = \mathcal{E}(f), \quad \text{almost surely as } n \to \infty,
\end{aligned}
$$

by the LLN applied to each term. Therefore, when $\varepsilon > 0$ is fixed, the second term in (16) tends to zero almost surely, and since $\varepsilon > 0$ is arbitrary, we conclude that

$$\sup_{f \in \mathcal{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| \to 0, \text{in probability, as } n \to \infty. \quad (17)$$

2.5 Large Deviations of the Sample Energy

To show an almost sure convergence in (17), we need to refine the bound on the supremum of the empirical process (12). For this, we apply Hoeffding's Inequality for U-statistics (Hoeffding, 1963), which is a special case of (de la Peña and Giné, 1999, Theorem 4.1.8).

Lemma 2 (Hoeffding's Inequality for U-statistics) *Let* $\phi : M \times M \to \mathbb{R}$ *be a bounded measurable map, and let* $\{x_i : i \geq 1\}$ *be a sequence of i.i.d. random variables with values in* M*. Assume that* $\mathbb{E}[\phi(x_1, x_2)] = 0$ *and that* $b := \|\phi\|_\infty < \infty$*, and let* $\sigma^2 = \text{Var}(\phi(x_1, x_2))$*. Then, for all* $t > 0$*,*

$$\mathbb{P}\left[\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \phi(x_i, x_j) > t\right] \leq \exp\left(-\frac{nt^2}{5\sigma^2 + 3bt}\right).$$

Let $f \in \mathcal{F}_1$. To bound the deviations of $\mathcal{E}(Y_n(f))$, we apply this result with $\phi(x, x') = \|f(x) - f(x')\|^2 - \mathcal{E}(f)$. Then,

$$\mathcal{E}(Y_n(f)) - \mathcal{E}(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \phi(x_i, x_j).$$

By construction, $\mathbb{E}[\phi(x_1, x_2)] = 0$. Since f is Lipschitz with constant 1, for any x and x' in M, $\|f(x) - f(x')\|^2 \leq \text{diam}(M)^2$ and $\mathcal{E}(f) \leq \text{diam}(M)^2$. Hence $\|\phi\|_\infty \leq \text{diam}(M)^2$, and $\text{Var}(\phi(x_1, x_2)) \leq \|\phi\|_\infty^2 \leq \text{diam}(M)^4$. Applying Lemma 2 (twice), we deduce that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|\mathcal{E}(Y_n(f)) - \mathcal{E}(f)| > \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{5\text{diam}(M)^4 + 3\text{diam}(M)^2\varepsilon}\right). \tag{18}$$

Using (18) in (16), coupled with the union bound, we get that

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_1} |\mathcal{E}(Y_n(f)) - \mathcal{E}(f)| > 9\varepsilon\,\text{diam}(M)\right) \leq \mathcal{N}_\infty(\mathcal{F}_1^0, \varepsilon) \cdot 2\exp\left(-\frac{n\varepsilon^2}{5\text{diam}(M)^2 + 3\varepsilon}\right). \tag{19}$$

Clearly, the RHS is summable for every $\varepsilon > 0$ fixed, so the convergence in (17) happens in fact with probability one, that is,

$$\sup_{f \in \mathcal{F}_1} |\mathcal{E}(Y_n(f)) - \mathcal{E}(f)| \to 0, \text{ almost surely, as } n \to \infty.$$

2.6 Convergence in Value: Proof of (5)

Assume r_n satisfies the Connectivity requirement, and that n is large enough that we have $\max(c(r_n), 6\lambda_n) < 1$. When $\Lambda(\lambda_n r_n)$ holds, by (13), we have

$$\left|\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) - \sup_{f \in \mathcal{F}_1} \mathcal{E}(f)\right| \leq (1 + 6\lambda_n)^2 \sup_{f \in \mathcal{F}_1} |\mathcal{E}(Y_n(f)) - \mathcal{E}(f)| + 3\max\left(c(r_n), 6\lambda_n\right)\text{diam}(M)^2,$$

while when $\Lambda(\lambda_n r_n)$ does not hold, since the energies are bounded by $\text{diam}(M)^2$, we have

$$\left|\sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) - \sup_{f \in \mathcal{F}_1} \mathcal{E}(f)\right| \leq 2\text{diam}(M)^2.$$

Combining these inequalities, we deduce that

$$
\begin{aligned}
\left| \sup_{Y \in \mathcal{Y}_{n,r}} \mathcal{E}(Y) - \sup_{f \in \mathcal{F}_1} \mathcal{E}(f) \right| \leq{} & 3 \max \left(c(r_n), 6\lambda_n \right) \operatorname{diam}(M)^2 \, \mathbb{I}_{\Lambda(\lambda_n r_n)} \\
& + 2 \operatorname{diam}(M)^2 \, \mathbb{I}_{\Lambda(\lambda_n r_n)^{\mathsf{c}}} \\
& + (1 + 6\lambda_n)^2 \sup_{f \in \mathcal{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right|. \quad (20)
\end{aligned}
$$

Almost surely, the sum of the first two terms on the RHS tends to 0 by the fact that $c(r) \to 0$ when $r \to 0$, and (7) since r_n satisfies the Connectivity requirement. The third term tends to 0 by (17). Hence, (5) is established.

2.7 Convergence in Solution: Proof of (6)

Assume r_n satisfies the Connectivity requirement, and that n is large enough that $\lambda_n \leq 1/2$. Let $\hat{Y}_n$ denote any solution of Discrete MVU. When $\Lambda(\lambda_n r_n)$ holds, there is $\hat{f}_n \in \mathcal{F}_{1+6\lambda_n}$ such that $\hat{Y}_n = Y_n(\hat{f}_n)$. Note that the existence of the interpolating function $\hat{f}_n$ holds on $\Lambda(\lambda_n r_n)$ for each fixed n, and that this does not imply the existence of an interpolating sequence $(\hat{f}_n)_{n \geq 1}$. That said, for each ω in the event $\liminf_n \Lambda(\lambda_n r_n)$, there exists a sequence $\hat{f}_n(.; \omega)$ and an integer $n_0(\omega)$ such that $\hat{Y}_n = Y_n(\hat{f}_n)$ for all $n \geq n_0(\omega)$, that is, the sequence is interpolating a solution of Discrete MVU for all n large enough. In addition, when r_n satisfies the Connectivity requirement, then with probability one, $\Lambda(\lambda_n r_n)^{\mathsf{c}}$ holds for only finitely many n's by the Borel-Cantelli lemma, implying that, with probability one, $\Lambda(\lambda_n r_n)$ holds infinitely often.

In fact, without loss of generality, we may assume that $\hat{f}_n \in \mathcal{F}_{1+6\lambda_n}^0 \subset \mathcal{F}_4^0$. Since $\mathcal{F}_4^0$ is equicontinuous and bounded, it is compact for the topology of the supnorm by the Arzelà-Ascoli Theorem. Hence, any subsequence of $\hat{f}_n$ admits a subsequence that converges in supnorm. And since $\mathcal{F}_L^0$ increases with L and $\mathcal{F}_1^0 = \cap_{L>1} \mathcal{F}_L^0$, any accumulation point of $(\hat{f}_n)$ is in $\mathcal{F}_1^0$.

In fact, if we define $\mathcal{S}_1^0 = \mathcal{S}_1 \cap \mathcal{F}_1^0$, then all the accumulation points of $(\hat{f}_n)$ are in $\mathcal{S}_1^0$. Indeed, we have

$$
\mathcal{E}(\hat{f}_n) = \mathcal{E}(\hat{f}_n) - \mathcal{E}(Y_n(\hat{f}_n)) + \mathcal{E}(Y_n(\hat{f}_n)),
$$

with

$$
\left| \mathcal{E}(\hat{f}_n) - \mathcal{E}(Y_n(\hat{f}_n)) \right| \leq \sup_{f \in \mathcal{F}_1} \left| \mathcal{E}(Y_n(f)) - \mathcal{E}(f) \right| \to 0,
$$

by (17), and

$$
\mathcal{E}(Y_n(\hat{f}_n)) = \sup_{Y \in \mathcal{Y}_{n,r_n}} \mathcal{E}(Y) \to \sup_{f \in \mathcal{F}_1} \mathcal{E}(f),
$$

by (5), almost surely as $n \to \infty$. Hence, if $f_\infty = \lim_k \hat{f}_{n_k}$, by continuity of $\mathcal{E}$ on $\mathcal{F}_4^0$, we have

$$
\mathcal{E}(f_\infty) = \lim_k \mathcal{E}(\hat{f}_{n_k}) = \sup_{f \in \mathcal{F}_1} \mathcal{E}(f),
$$

and given that $f_\infty \in \mathcal{F}_1^0$, we have $f_\infty \in \mathcal{S}_1^0$ by definition.

The fact that $(\hat{f}_n)$ is compact with all accumulation points in $\mathcal{S}_1^0$ implies that

$$
\inf_{f \in \mathcal{S}_1^0} \| \hat{f}_n - f \|_\infty \to 0, \quad (21)
$$

and since we have $\max_{1 \leq i \leq n} \| \hat{y}_i - f(x_i) \| = \| \hat{f}_n(x_i) - f(x_i) \| \leq \| \hat{f}_n - f \|_\infty$, this immediately implies (6). The convergence in (21) is a consequence of the following simple result.

Lemma 3 *Let (a_n) be a sequence in a compact metric space with metric* δ, *that has all its accumulation points in a set A. Then*

$$\inf_{a \in A} \delta(a_n, a) \to 0.$$

Proof If this is not the case, then there is $\varepsilon > 0$ such that, $\inf_{a \in A} \delta(a_n, a) \geq \varepsilon$ for infinitely many n's, denoted $n_1 < n_2 < \cdots$. The space being compact, (a_{n_k}) has at least one accumulation point, which is in A by assumption. However, by construction, (a_{n_k}) cannot have an accumulation point in A. This is a contradiction. ∎

3. Quantitative Convergence Bounds

We obtained a general, qualitative convergence result for MVU in the preceding section and now specify some of the supporting arguments to obtain quantitative convergence speeds. This will require some (natural) additional assumptions on μ and M. While the proof of a result like Theorem 1 is necessarily complex, we endeavored in making it as transparent and simple as we could. The present section is more technical, and the reader might choose to first read Section 4 to learn about the solutions to Continuum MVU, which imply consistency (and inconsistencies) for MVU as a dimensionality-reduction algorithm.

We consider two specific types of sets M:

- *Thin sets.* M is a d-dimensional compact, connected, C^2 submanifold with C^2 boundary (if nonempty). In addition, $M \subset M_\star$, where $M_\star$ is a d-dimensional, geodesically convex C^2 submanifold.

- *Thick sets.* M is a compact, connected subset that is the closure of its interior and has a C^2 boundary.

The ambient space is $\mathbb{R}^p$. Note that our results are equally valid for piecewise smooth sets. Thin sets are a model for noiseless data, where that the data points are sampled from a submanifold. Note that they may have holes and boundaries. And thick sets are a model for noisy data, where that the data points are sampled from the vicinity of a submanifold.

An important example of thick sets are tubular neighborhoods of thin sets. For a set $A \subset \mathbb{R}^p$ and $\eta > 0$, the η-neighborhood of A is the set of points in $\mathbb{R}^p$ within Euclidean distance η of A, and is denoted $B(A, \eta)$. The reach of a set $A \subset \mathbb{R}^p$ is defined in Federer (1959) as the largest η such that, for any $x \in B(A, \eta)$ there is a unique point $a \in A$ closest to x. We denote by $\rho(A)$ the reach of A. Note that any thin set A has positive reach, which bounds its radius of curvature from below. While for any thick set A, ∂A is a thin set without boundary, for any $\eta < \rho(A)$, $\bar{B}(A, \eta)$ is a thick set, with boundary having reach $\geq \rho(A) - \eta$.

In what follows, C and C_k denote constants that depend only on p and d, which may change with each appearance.

3.1 The Regularity Condition

The first thing we do is specify the function c in (3). When M is a thin set, we define $r_M = \min\big(\rho(M_\star), \rho(\partial M)\big)$, where by convention $\rho(\emptyset) = \infty$. And when M is a thick set, we let $r_M = \rho(\partial M)$.

The following result seems valid when $r_M = \rho(M)$ in both cases, but the proof seems much more involved.

Lemma 4 *Whether M is a thin or a thick set,* (3) *is valid with*

$$c(r) = \frac{4r}{r_M} \mathbb{1}_{\{r < r_M/2\}} + \mathbb{1}_{\{r \geq r_M/2\}}.$$

Proof We borrow results from Niyogi et al. (2008). Let $x, x' \in M$ such that $\|x - x'\| \leq r_M/2$.

First, suppose that M is thick. Consider the (straight) line segment joining these two points. If this segment is included in M, then $\delta_M(x, x') = \|x - x'\|$. Otherwise, it intersects ∂M in at least two points; among these points, let z be the closest to x and z' the closest to x'. Since ∂M has no boundary, it is geodesically convex, so that there is a geodesic on ∂M, denoted ξ, joining z and z'. Niyogi et al. (2008, Prp. 6.3) applies since $\|z - z'\| \leq \|x - x'\| \leq r_M/2 \leq \rho(\partial M)/2$, and $\rho(\partial M)$ coincides with the condition number of ∂M as defined in Niyogi et al. (2008)—and denoted by τ there. Hence, if ℓ is the length of ξ, we have

$$\ell \leq \rho(\partial M) - \rho(\partial M)\sqrt{1 - \frac{2\|z - z'\|}{\rho(\partial M)}} \leq \|z - z'\| + 4\|z - z'\|^2/r_M, \tag{22}$$

using the fact that $\sqrt{1-t} \geq 1 - t/2 - t^2$ for all $t \in [0,1]$ and $r_M \leq \rho(\partial M)$. Let γ be the path made of ξ concatenated with the segments $[xz]$ and $[z'x']$. If L is the length of γ, we have

$$\begin{aligned} L &= \|x - z\| + \|z' - x'\| + \ell \\ &\leq \|x - z\| + \|z' - x'\| + \|z - z'\| + 4\|z - z'\|^2/r_M \\ &\leq \|x - x'\| + 4\|x - x'\|^2/r_M, \end{aligned}$$

using the fact that x, z, z', x' are in that order on the line segment joining x and x'. This concludes the proof when M is thick.

When M is thin, we distinguish two cases. Either there is a geodesic joining x and x', and Niyogi et al. (2008, Prp. 6.3) is directly applicable. Otherwise, M is not geodesically convex. Let $\gamma_\star$ be a geodesic on $M_\star$ joining x and x'. Necessarily, it hits the boundary ∂M in at least two points. Let z, z', ξ and ℓ be defined as before. We again have (22). Let $(xz)_\star$ and $(z'x')_\star$ denote the arcs along $\gamma_\star$ joining x and z, and z' and x', respectively. Applying Niyogi et al. (2008, Prp. 6.3) to each arc, which is possible since $r_M \leq \rho(M_\star)$, we also have

$$\text{length}((xz)_\star) \leq \|x - z\| + 4\|x - z\|^2/r_M, \qquad \text{length}((z'x')_\star) \leq \|z' - x'\| + 4\|z' - x'\|^2/r_M.$$

Let γ be the curve made of concatenating these two arcs and ξ, and let L denote its length. We have

$$\begin{aligned} L &= \text{length}((xz)_\star) + \text{length}((z'x')_\star) + \ell \\ &\leq \|x - z\| + \frac{4\|x - z\|^2}{r_M} + \|z' - x'\| + \frac{4\|z' - x'\|^2}{r_M} + \|z - z'\| + \frac{4\|z - z'\|^2}{r_M} \\ &\leq \|x - x'\| + \frac{4\|x - x'\|^2}{r_M}. \end{aligned}$$

This concludes the proof when M is thin. ∎

3.2 Covering Numbers and a Bound on the Neighborhood Radius

At what speed can we have $r_n \to 0$ and still have (7) hold? This question is of practical importance, since the neighborhood radius may affect the output of MVU in a substantial way. Computationally, it is preferable to have r_n small, so there are fewer constraints in (1). However, we already explained that r_n needs to be large enough that, at the very minimum, the resulting neighborhood graph is connected. In fact, we required the stronger condition (7).

To keep the exposition simple, we assume that μ is comparable to the uniform distribution on M, that is, we assume that there is a constant $\alpha > 0$ such that

$$\mu(B(x,\eta)) \geq \alpha \operatorname{vol}_d(B(x,\eta) \cap M), \quad \forall x \in M, \forall \eta > 0, \tag{23}$$

where vol_d denotes the d-dimensional Hausdorff measure and d denotes the Hausdorff dimension of M. We need the following result. Let ω_d be the volume of the d-dimensional unit ball.

Lemma 5 *Whether M is thin or thick, there is $C > 0$ such that, for any $\eta \leq r_M$ and any $x \in M$,*

$$\operatorname{vol}_d(B(x,\eta) \cap M) \geq C\eta^d.$$

Proof It suffices to prove the result for $x \in M \setminus \partial M$ and for η small enough.

Thick set. We first assume that M is thick. Take $x \in M$ and $\eta < r_M$. If $\operatorname{dist}(x, \partial M) \geq \eta$, then $B(x,\eta) \subset M$ and the result follows immediately. Otherwise, let u be the metric projection of x onto ∂M, and define $z = x + (\eta/4)(x-u)/\|x-u\|$. By the triangle inequality, $B(z,\eta/4) \subset B(x,\eta)$. Also, by Federer (1959, Theorem 4.8), u is also the metric projection of $z \in M$ onto ∂M, so that $\operatorname{dist}(z, \partial M) = \|z-u\| = \|x-u\| + \eta/4 > \eta/4$. And, necessarily, $z \in M$, for otherwise the line segment joining z to x would intersect ∂M, and any point on that intersection would be closer to z than u is, which cannot be. Therefore, $B(z,\eta/4) \subset B(x,\eta) \cap M$ and the result follows immediately.

Thin set. We now assume that M is thin. For $y \in M$, let T_y be the tangent subspace of M at y and let π_y denote the orthogonal projection onto T_y. Because M is a C^2 submanifold, for every $y \in M$, there is $\varepsilon_y > 0$ such that π_y is a C^2 diffeomorphism on $K_y := B(y, \varepsilon_y) \cap M$, with π_y^{-1} being 2-Lipschitz on $\pi_y(K_y)$—the latter comes from the fact that $D_y\pi_y$ is the identity map and $z \to D_z\pi_y$ is continuous. Since M is compact, there is $y_1, \dots, y_m \in M$, with $m < \infty$, such that $M \subset \cup_j B(y_j, \varepsilon_j/2)$. Let $\varepsilon = \min_j \varepsilon_{y_j}$, which is strictly positive. Let y be among the y_j's such that $x \in B(y, \varepsilon_j/2)$. Assuming that $\eta < \varepsilon/2$, we have that $B(x,\eta) \subset B(y, \varepsilon_j)$. Let $U := B(y, \varepsilon_j)$, $K = K_y$, $T = T_y$ and $\pi = \pi_y$ for short.

We first show that, if $\partial M \cap K \neq \emptyset$ and $W := \pi(\partial M \cap K)$, then $\rho(W) \geq \rho(\partial M)$. Indeed, for any $z, z' \in K$, we have

$$\operatorname{dist}(\pi(z') - \pi(z), \operatorname{Tan}(W, \pi(z))) \leq \operatorname{dist}(z'-z, \operatorname{Tan}(\partial M, z)) \leq \frac{1}{2\rho(\partial M)}\|z'-z\|^2,$$

where the first inequality follows from the facts that $\operatorname{Tan}(W, \pi(z)) = \pi(\operatorname{Tan}(\partial M, z))$ and that π is 1-Lipschitz, and the second inequality from Federer (1959, Theorem 4.18) applied to ∂M. In turn, Federer (1959, Theorem 4.17) applied to W implies that $\rho(W) \geq \rho(\partial M)$.

We can now reason as we did for thick sets, but with a twist. To be sure, let $a = \pi(x)$ and notice that $B(a,\eta) \cap T = \pi(B(x,\eta)) \subset \pi(U)$ since $B(x,\eta) \subset U$. If $\operatorname{dist}(a,W) \geq \eta/2$, $B(a,\eta/2) \cap T \subset \pi(K)$. If $\operatorname{dist}(a,W) < \eta/2$, let b be the metric projection of a onto W and define $c = a + (\eta/8)(a-b)/\|a-b\|$. Arguing exactly as we did for thick sets, we have that $B(c,\eta/8) \cap T \subset B(a,\eta/2) \cap \pi(K)$. Let

1759

$L = \pi^{-1}(B(c, \eta/8) \cap T)$. Note that $L \subset \pi^{-1}(B(a, \eta/2) \cap T) \cap K \subset B(x, \eta) \cap K \subset B(x, \eta) \cap M$, since π is injective on K and π^{-1} is 2-Lipschitz on $\pi(K)$. In addition, since π is 1-Lipschitz on K, we have $\text{vol}_d(L) \geq \text{vol}_d(\pi(L)) = \text{vol}_d(B(c, \eta/8) \cap T)$. This immediately implies the result. $\blacksquare$

When (23) is satisfied, and M is either thin or thick, we can provide sharp rates for r_n. Just as we did in Section 2.1, we work with coverings of M. Let $\mathcal{N}(M, \eta)$ denote the cardinality of a minimal η-covering of M for the Euclidean norm.

Lemma 6 *Suppose $\eta \leq r_M$. When M is thick,*

$$\mathcal{N}(M, \eta) \leq C \text{vol}_p(M) \eta^{-p};$$

and when M is thin and $0 \leq \sigma < \rho(M)$,

$$\mathcal{N}(B(M, \sigma), \eta) \leq C \text{vol}_d(M) \max(\sigma, \eta)^{p-d} \eta^{-p}.$$

The constant C depends only on p and d.

Proof Suppose M is thick and let $z_1, \ldots, z_{N_\eta}$ an η-packing of M of size $N_\eta := \mathcal{N}(M, \eta)$. Since $B(z_i, \eta/2) \cap B(z_j, \eta/2) = \emptyset$ when $i \neq j$, we have

$$\text{vol}_p(M) \geq \sum_j \text{vol}_p(B(z_j, \eta/2) \cap M) \geq N_\eta C_p \eta^p,$$

where C_p is the constant in Lemma 5. The bound on N_η follows.

Suppose M is thin. When $\sigma \leq \eta/4$, let $z_1, \ldots, z_{N_{\eta/4}}$ an $(\eta/4)$-packing of M. Then by the triangle inequality, $B(M, \sigma) \subset \cup_j B(z_j, \eta/2)$, and therefore $\mathcal{N}(B(M, \sigma), \eta) \leq N_{\eta/4}$. When $\sigma \geq \eta/4$, let $z_1, \ldots, z_N$ be an $(\eta/4)$-packing of $B(M, \sigma - \eta/4)$. Since $B(z_i, \eta/8) \cap B(z_j, \eta/8) = \emptyset$ when $i \neq j$, and $B(z_i, \eta/8) \subset B(M, \sigma)$, we have

$$\text{vol}_p(B(M, \sigma)) \geq \sum_j \text{vol}_p(B(z_j, \eta/8)) = N \omega_p (\eta/8)^p.$$

Hence, $N \leq \omega_p^{-1} (\eta/8)^{-p} \text{vol}_p(B(M, \sigma))$. By Weyl's volume formula for tubes (Weyl, 1939), we have $\text{vol}_p(B(M, \sigma)) \leq C_1 \text{vol}_d(M) \sigma^{p-d}$ for a constant C_1 depending on p and d. Then the result follows from the fact that, by the triangle inequality, $B(M, \sigma) \subset \cup_j B(z_j, \eta/2)$, so that $\mathcal{N}(B(M, \sigma), \eta) \leq N$. $\blacksquare$

We are now ready to take a closer look at (7). Let η_n be defined as in Section 2.1. By (23) and Lemma 5, we have $p_\eta \geq C_1 \alpha \eta^d$, and we have $\mathcal{N}(M, \eta) \leq C_2 \eta^{-d}$ by Lemma 6, where C_1 and C_2 depend only on M. Hence,

$$\mathcal{N}(M, \eta)(1 - p_\eta)^n \leq C_2 \eta^{-d} (1 - C_1 \alpha \eta^d)^n \leq C_2 \eta^{-d} e^{-nC_1 \alpha \eta^d} \leq \frac{1}{n^2},$$

when

$$\eta^d \geq (C_1 \alpha n)^{-1} \log (C_2 \eta^{-d} n^2).$$

We deduce that any $r_n \gg r_n^\ddagger := (\log(n)/n)^{1/d}$ satisfies (7) with any $\lambda_n \to 0$ such that $\lambda_n \gg r_n^\ddagger / r_n$.

3.3 Packing Numbers of Lipschitz Functions on M

It appears necessary to provide a bound for $\mathcal{N}_\infty(\mathcal{F}_1^0, \eta)$. For this, we follow the seminal work of Kolmogorov and Tikhomirov (1961) on entropy bounds for classical functions classes (including Lipschitz classes). We provide details for completeness.

Lemma 7 *For any M compact, connected subset of $\mathbb{R}^p$ satisfying* (3)*, there is a constant C such that*

$$\log \mathcal{N}_\infty(\mathcal{F}_1^0, \eta) \leq C\left(\log(1/\eta) + \mathcal{N}(M, \eta/C)\right),$$

for all $0 < \eta \leq 1$.

Proof Take $0 < \varepsilon \leq 1/\sqrt{p}$ and let $C_0 = 2\sqrt{p}(2 + c(2))$. For $j = (j_1, \ldots, j_p) \in \mathbb{Z}^p$, let $Q_j = \prod_{s=1}^p [j_s \varepsilon, (j_s + 1)\varepsilon)$. Let $J = \{j : Q_j \cap M \neq \emptyset\}$, which we see as a subgraph of the lattice for the 2^p-nearest neighbor topology.

Note that $|J| \leq C_1 \mathcal{N}(M, \varepsilon)$. Indeed, let $e_1, \ldots, e_{2^p}$ be the vertices of the unit hypercube of $\mathbb{R}^p$ and let $Z_s = e_s + (2\mathbb{Z})^p$. Also, let $Z_0 = (2\mathbb{Z})^p$. By construction, $Z_1, \ldots, Z_{2^p}$ is a partition of $\mathbb{Z}^p$. Therefore, there is s (say $s = 1$) such that $|J \cap Z_s| \geq |J|/2^p$. For each $j \in J \cap Z_1$, pick $x_j \in Q_j \cap M$. By construction, for any $j \neq j'$ both in $J \cap Z_1$, $\|x_j - x_{j'}\| > 2\varepsilon$, so $|J \cap Z_1|$ is smaller than the 2ε-packing number of M, which is smaller than the ε-covering number of M.

Note also that $\cup_j Q_j$ is connected because M is. Let $\pi_1, \ldots, \pi_\ell$ be a sequence covering J and such that Q_{π_s} and $Q_{\pi_{s-1}}$ are adjacent. A depth-first construction gives a sequence π of length at most $\ell \leq C_2 |J|$, since each Q_j has a constant number ($= 2^p$) of adjacent hypercubes.

Let $y_1, \ldots, y_m$ be an enumeration of the ε-grid $(\varepsilon\mathbb{Z} \cap [-\operatorname{diam}(M), \operatorname{diam}(M)])^p$. Note that $m \leq C_3 \varepsilon^{-p}$ and that, for each s there are at most C_4 indices t such that $\|y_s - y_t\| \leq C_0 \varepsilon$.

Consider the class $\mathcal{G}$ of piecewise-constant functions $g : M \to \mathbb{R}^p$ of the form $g(x) = y_{t_j}$ for all $x \in Q_j \cap M$ and such that $\|y_{t_j} - y_{t_k}\| \leq C_0 \varepsilon$ when Q_j and Q_k are adjacent. This is a subclass of the class of functions of the form $g(x) = y_{t_{\pi(j)}}$ for all $x \in Q_{\pi(j)}$ and such that $\|y_{t_{\pi(j)}} - y_{t_{\pi(j-1)}}\| \leq C_0 \varepsilon$. The cardinality of the larger class is at most $m C_4^{\ell-1}$, since there are m possible values for $y_{t_{\pi(1)}}$ and then, at each step along π, there at most C_4 choices. Therefore,

$$
\begin{aligned}
\log |\mathcal{G}| &\leq \log m + \ell \log C_4 \\
&\leq \log(C_3) + p \log(1/\varepsilon) + C_2 C_1 \mathcal{N}(M, \varepsilon) \log(C_4) \\
&\leq C_5(\log(1/\varepsilon) + \mathcal{N}(M, \varepsilon)).
\end{aligned}
$$

For each j, choose $z_j \in Q_j \cap M$. Take any $f \in \mathcal{F}_1^0$. For each j, let t_j be such that $\|f(z_j) - y_{t_j}\| \leq \sqrt{p}\varepsilon$ and let g be defined by $g(x) = y_{t_j}$ for all $x \in Q_j$. Suppose Q_j and Q_k are adjacent, so that $\|z_j - z_k\| \leq 2\sqrt{p}\varepsilon \leq 2$. By the triangle inequality, (2) and (3), we have

$$
\begin{aligned}
\|y_{t_j} - y_{t_k}\| &\leq \|f(z_j) - f(z_k)\| + \|y_{t_j} - f(z_j)\| + \|y_{t_k} - f(z_k)\| \\
&\leq (1 + c(\|z_j - z_k\|))\|z_j - z_k\| + \sqrt{p}\varepsilon + \sqrt{p}\varepsilon \\
&\leq (1 + c(2))2\sqrt{p}\varepsilon + 2\sqrt{p}\varepsilon \\
&= C_0\varepsilon.
\end{aligned}
$$

so that $g \in \mathcal{G}$. Moreover, for $x \in Q_j \cap M$,

$$\|g(x) - f(x)\| = \|y_{t_j} - f(z_j)\| + \|f(z_j) - f(x)\| \leq \sqrt{p}\varepsilon + (1 + c(\sqrt{p}\varepsilon))\sqrt{p}\varepsilon \leq (2 + c(1))\sqrt{p}\varepsilon.$$

The result follows from choosing $\varepsilon = \eta/((2+c(1))\sqrt{p})$. ∎

In particular, if M is thin or thick, we have

$$\log \mathcal{N}_\infty(\mathcal{F}_1^0, \eta) \leq C\eta^{-d},$$

by Lemma 6 (applied with $\sigma = 0$ with M is thin) and Lemma 7. Recall that d is the intrinsic dimension of M.

3.4 Quantitative Convergence Bound

From (19) and Lemma 7, there is a constant $C > 0$ such that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_1} \left|\mathcal{E}(Y_n(f)) - \mathcal{E}(f)\right| > Cn^{-1/(d+2)}\right) \leq \exp(-n^{-d/(d+2)}).$$

Using this fact in (20), together with Lemma 4 and the order of magnitude for r_n derived in Section 3.2, leads to a bound on the rate of convergence in (5) via the Borel-Cantelli Lemma.

Theorem 2 *Suppose that M is either thin or thick, of dimension d, and that (23) holds. Assume that $r_n \to 0$ such that $r_n \gg r_n^\dagger := (\log(n)/(\alpha n))^{1/d}$ and take any $a_n \to \infty$. Then, with probability one,*

$$\left|\sup\{\mathcal{E}(Y) : Y \in \mathcal{Y}_{n,r_n}\} - \sup\{\mathcal{E}(f) : f \in \mathcal{F}_1\}\right| \leq a_n\left(r_n + \frac{r_n^\dagger}{r_n} + n^{-1/(d+2)}\right),$$

for n large enough.

We speculate that this convergence rate is not sharp and that the first term in brackets can be replaced by r_n^2. Indeed, we believe the result (Niyogi et al., 2008, Prop. 6.3) for approximating geodesics distances is not rate-optimal, leading to a loose Lemma 4, while we anticipate that, in fact, $c(r) = O(r^2)$.

Unfortunately, we do not have a quantitative bound on the rate of convergence of the solutions in (6).

4. Continuum MVU

Now that we established the convergence of Discrete MVU to Continuum MVU, we study the latter, and in particular its solutions. We mostly focus on the case where M is isometric to a Euclidean domain.

Isometry assumption. We assume that M is isometric to a compact, connected domain $D \subset \mathbb{R}^d$. Specifically, there is a bijection $\psi : M \to D$ satisfying $\delta_D(\psi(x), \psi(x')) = \delta_M(x, x')$ for all $x, x' \in M$.

As a glimpse of the complexity of the notion of isometry, and also for further reference, consider a domain D as above. Then the canonical inclusion ι of D in $\mathbb{R}^d$ is not necessarily an isometry between the metric spaces (D, δ_D) and $(\mathbb{R}^d, \|\cdot\|)$. To see this, let x and x' be two points of D. Let γ be a shortest path connecting x to x' in D. Suppose that $\iota : (D, \delta_D) \to (\mathbb{R}^d, \|\cdot\|)$ is an isometry. Then, $L(\iota \circ \gamma) = L(\gamma) = \delta_D(x, x') = \|\iota(x) - \iota(x')\|$. So the image path $\iota \circ \gamma$ is a shortest path connecting $\iota(x)$ to $\iota(x')$, hence a segment. Since this segment lies in $\iota(D) = D$, and since this holds for any pair of

points x, x' in D, this implies that D is convex. Conversely, if D is convex, the canonical inclusion ι is an isometry.

We start by showing that, in the case where M is isometric to a convex domain, then MVU recovers this convex domain modulo a rigid transformation, so that MVU is consistent is that case. The last part of the section is dedicated to a perturbation analysis that shows two things. First, that Continuum MVU changes slowly with the amount of noise, up to a point. And second, that when M is isometric to a domain that is not convex, MVU may not recover this domain. We provide some illustrative examples of that.

In the following, we identify $\mathbb{R}^d$ with $\mathbb{R}^d \times \{0\}^{p-d} \subset \mathbb{R}^p$.

4.1 Consistency under the Convex Assumption

If we assume that D is convex, then MVU recovers D up to a rigid transformation, in the following sense. Recall that $\mathcal{S}_1$ is the solution space of Continuum MVU.

Theorem 3 *Suppose that M is isometric to a convex subset $D \subset \mathbb{R}^d$ with isometry mapping $\psi : M \to D$, and that* (23) *holds. Then*
$$\mathcal{S}_1 = \{\zeta \circ \psi : \zeta \in \text{Isom}(\mathbb{R}^p)\}.$$

Proof Note first that, since D is convex, its intrinsic distance coincides with the Euclidean distance of $\mathbb{R}^d$, that is, $\delta_D = \|\cdot\|$. For all f in $\mathcal{F}_1$, we have

$$
\begin{aligned}
\mathcal{E}(f) &= \int_{M \times M} \|f(x) - f(x')\|^2 \mu(dx)\mu(dx') \\
&\leq \int_{M \times M} \delta_M(x, x')^2 \mu(dx)\mu(dx') \\
&= \int_{M \times M} \delta_D(\psi(x), \psi(x'))^2 \mu(dx)\mu(dx') \\
&= \int_{M \times M} \|\psi(x) - \psi(x')\|^2 \mu(dx)\mu(dx') \\
&= \int_{D \times D} \|z - z'\|^2 (\mu \circ \psi^{-1})(dz)(\mu \circ \psi^{-1})(dz'),
\end{aligned}
$$

while
$$\mathcal{E}(\psi) = \int_{D \times D} \|z - z'\|^2 (\mu \circ \psi^{-1})(dz)(\mu \circ \psi^{-1})(dz').$$

So
$$\sup_{f \in \mathcal{F}_1} \mathcal{E}(f) = \mathcal{E}(\psi) = \int_{D \times D} \|z - z'\|^2 (\mu \circ \psi^{-1})(dz)(\mu \circ \psi^{-1})(dz').$$

Hence $\psi \in \mathcal{S}_1$, and since $\mathcal{E}(\zeta \circ \psi) = \mathcal{E}(\psi)$ for any isometry $\zeta : \mathbb{R}^p \to \mathbb{R}^p$,

$$\{\zeta \circ \psi : \zeta \in \text{Isom}(\mathbb{R}^p)\} \subset \mathcal{S}_1.$$

Now let $f : M \to \mathbb{R}^p$ be a function in $\mathcal{F}_1$ so that $\|f(x) - f(x')\| \leq \delta_M(x, x')$ for any points x and x' in M. Suppose that f is not an isometry. Then there exists two points x and x' in M such that

$$\|f(x) - f(x')\| < \delta_M(x, x').$$

By continuity of f, there exists a nonempty open subset U of $M \times M$ containing (x, x') such that $\|f(z) - f(z')\| < \delta_M(z, z')$ for all (z, z') in U. In addition, $\mu(U) > 0$ by (23). Consequently

$$
\begin{aligned}
\mathcal{E}(f) &= \int_{M \times M \setminus U} \|f(x) - f(x')\|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x') + \int_U \|f(x) - f(x')\|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&< \int_{M \times M} \delta_M(x, x')^2 \mu(\mathrm{d}x)\mu(\mathrm{d}x') \\
&= \sup_{f \in \mathcal{F}_1} \mathcal{E}(f).
\end{aligned}
$$

So any function f in $\mathcal{F}_1$ which is not an isometry onto its image does not belong to $\mathcal{S}_1$.

At last, since for any isometry f in $\mathcal{S}_1$, the map $f \circ \psi^{-1} : \mathbb{R}^p \to \mathbb{R}^p$ is an isometry, there exists some isometry $\zeta \in \mathrm{Isom}(\mathbb{R}^p)$ such that $f = \zeta \circ \psi$, and we conclude that

$$
\{\zeta \circ \psi : \zeta \in \mathrm{Isom}(\mathbb{R}^p)\} = \mathcal{S}_1.
$$

∎

In conclusion, MVU recovers the isometry when the domain D is convex. Note that this is also the case of Isomap.

4.2 Noisy Setting

When the setting is noisy, with noise level $\sigma \geq 0$, $x_1, \ldots, x_n$ are sampled from μ_σ, a (Borel) probability distribution on $\mathbb{R}^p$ with support $M_\sigma := \bar{B}(M, \sigma)$, that is, M_σ is composed of all the points of $\mathbb{R}^p$ that are at a distance at most σ from M. To speak of noise stability, we assume that μ_σ converges weakly when $\sigma \to 0$. Let $\mathcal{F}_{1,\sigma}$ denote the class of 1-Lipschitz functions on M_σ, and so on. Our simple perturbation analysis is plainly based on the fact that $\mathcal{E}$ is continuous with respect to the noise level, in the following sense. This immediately implies that MVU is tolerant to noise.

Lemma 8 *Let $M \subset \mathbb{R}^p$ be of positive reach $\rho(M) > 0$ and assume that $\mu_\sigma \to \mu_0$ weakly when $\sigma \to 0$. Then as $\sigma \to 0$, we have*

$$
\sup_{f \in \mathcal{F}_{1,\sigma}} \mathcal{E}_\sigma(f) \to \sup_{f \in \mathcal{F}_1} \mathcal{E}(f), \tag{24}
$$

and

$$
\sup_{f \in \mathcal{S}_{1,\sigma}} \inf_{g \in \mathcal{S}_1} \sup_{x \in M_\sigma} \inf_{z \in M} \|f(x) - g(z)\| \to 0. \tag{25}
$$

Proof The metric projection $\pi : B(M, \rho(M)) \to M$ with $\pi(x) = \arg\min\{\|x - x'\| : x' \in M\}$, is well-defined and 1-Lipschitz (Federer, 1959, Theorem 4.8).

Consider any sequence $\sigma_m \to 0$ with $\sigma_m < \rho(M)$ for all $m \geq 1$, and let $f_m \in \mathcal{S}_{1,\sigma_m}^0$. Let g_m denote the restriction of f_m to M. Since $(g_m) \subset \mathcal{F}_1^0$ and $\mathcal{F}_1^0$ is compact for the supnorm, it admits a convergent subsequence. Assume (g_m) itself is convergent, without loss of generality. Then $g_m \to g_\star$, with $g_\star \in \mathcal{F}_1^0$. For $x \in B(M, \rho(M))$, define $f_\star(x) = g_\star(\pi(x))$. Then for $x \in M_{\sigma_m}$, we have

$$
\begin{aligned}
\|f_\star(x) - f_m(x)\| &\leq \|g_\star(\pi(x)) - g_m(\pi(x))\| + \|f_m(\pi(x)) - f_m(x)\| \\
&\leq \|g_\star - g_m\|_\infty + \|\pi(x) - x\| \\
&\leq \|g_\star - g_m\|_\infty + \sigma_m,
\end{aligned}
$$

since $f_m \in \mathcal{F}_{1,\sigma_m}$ and the segment $[\pi(x), x] \subset M_{\sigma_m}$. The latter is due to $\|\pi(x) - x\| \leq \sigma_m$ and $B(\pi(x), \sigma_m) \subset M_{\sigma_m}$, both by definition. Hence, as functions on M_{σ_m}, we have $\|f_\star(x) - f_m(x)\|_\infty \to 0$, that is,

$$\sup_{x \in M_{\sigma_m}} \|f_\star(x) - f_m(x)\| \to 0.$$

By (14), again applied to functions on M_{σ_m} for a fixed m, we have

$$
\begin{aligned}
\left| \mathcal{E}_{\sigma_m}(f_m) - \mathcal{E}_{\sigma_m}(f_\star) \right| &\leq 4\|f_\star(x) - f_m(x)\|_\infty \mathrm{diam}(M_{\sigma_m}) \\
&\leq 4\|f_\star(x) - f_m(x)\|_\infty \mathrm{diam}(B(M, \rho(M))) \\
&\to 0,
\end{aligned}
$$

and since $f_\star$ does not depend on m and is bounded, we also have

$$\mathcal{E}_{\sigma_m}(f_\star) \to \mathcal{E}(f_\star) = \mathcal{E}(g_\star) \leq \sup_{\mathcal{F}_1} \mathcal{E}. \tag{26}$$

Hence

$$
\begin{aligned}
\sup_{\mathcal{F}_{1,\sigma_m}} \mathcal{E}_{\sigma_m} &= \mathcal{E}_{\sigma_m}(f_m) \\
&= \mathcal{E}(f_\star) + \mathcal{E}_{\sigma_m}(f_\star) - \mathcal{E}(f_\star) + \mathcal{E}_{\sigma_m}(f_m) - \mathcal{E}_{\sigma_m}(f_\star) \\
&\leq \sup_{\mathcal{F}_1} \mathcal{E} + \mathcal{E}_{\sigma_m}(f_\star) - \mathcal{E}(f_\star) + \mathcal{E}_{\sigma_m}(f_m) - \mathcal{E}_{\sigma_m}(f_\star),
\end{aligned}
$$

and we deduce that

$$\varlimsup_{m \to \infty} \sup_{\mathcal{F}_{1,\sigma_m}} \mathcal{E}_{\sigma_m} \leq \sup_{\mathcal{F}_1} \mathcal{E},$$

and since this is true for all sequences $\sigma_m \to 0$ (and m large enough), we have

$$\varlimsup_{\sigma \to 0} \sup_{\mathcal{F}_{1,\sigma}} \mathcal{E}_\sigma \leq \sup_{\mathcal{F}_1} \mathcal{E}.$$

For the reverse relation, choose $g \in \mathcal{S}_1$ and for $x \in B(M, \rho(M))$ define $f(x) = g(\pi(x))$. As above, let $\sigma_m \to 0$ with $\sigma_m \leq \rho(M)$. Then $f \in \mathcal{F}_{1,\sigma_m}$ by composition, so that

$$\mathcal{E}_{\sigma_m}(f) \leq \sup_{\mathcal{F}_{1,\sigma_m}} \mathcal{E}_{\sigma_m}.$$

On the other hand,

$$\mathcal{E}_{\sigma_m}(f) \to \mathcal{E}(f) = \mathcal{E}(g) = \sup_{\mathcal{F}_1} \mathcal{E}.$$

Hence,

$$\sup_{\mathcal{F}_1} \mathcal{E} \leq \varlimsup_{\sigma \to 0} \sup_{\mathcal{F}_{1,\sigma}} \mathcal{E}_\sigma.$$

This concludes the proof of (24).

Equation (25) is now proved based on (24) in the same way (6) is proved based on (5), by contradiction. To be sure, assume (25) is not true. Then it is also not true for $\mathcal{S}_{1,\sigma}^0$ and $\mathcal{S}_1^0$. Hence, there is $\varepsilon > 0$, a sequence $\sigma_m \to 0$ and $f_m \in \mathcal{S}_{1,\sigma_m}^0$ such that

$$\inf_{g \in \mathcal{S}_1^0} \sup_{x \in M_{\sigma_m}} \inf_{z \in M} \|f_m(x) - g(z)\| \geq \varepsilon,$$

for infinitely many m's. Without loss of generality, we assume this is true for all m. For each m, let g_m be the restriction of f_m to M. Then, taking a subsequence if needed, $g_m \to g_\star \in \mathcal{F}_1^0$ in supnorm. As before, define $f_\star(x) = g_\star(\pi(x))$ for $x \in B(M, \rho(M))$. Following the same arguments, we have

$$\sup_{x \in M_{\sigma_m}} \|f_\star(x) - f_m(x)\| \to 0.$$

We also see that, necessarily, $g_\star \in \mathcal{S}_1^0$, for otherwise the inequality in (26) would be strict and this would imply that (24) does not hold. Hence

$$\sup_{x \in M_{\sigma_m}} \|f_\star(x) - f_m(x)\| \geq \sup_{x \in M_{\sigma_m}} \inf_{z \in M} \|f_m(x) - g_\star(z)\| \geq \inf_{g \in \mathcal{S}_1^0} \sup_{x \in M_{\sigma_m}} \inf_{z \in M} \|f_m(x) - g(z)\|.$$

This leads to a contradiction. Hence the proof of (25) is complete. ∎

4.3 Inconsistencies

We provide two emblematic situations where MVU fails to recover D. They are both consequences of MVU's robustness to noise. In both cases, we consider the simplest situation where $M = D \subset \mathbb{R}^2$ and μ is the uniform distribution. Note that ψ is the identity function in this case, that is, $\psi(x) = x$, and the Isometry Assumption is clearly satisfied. We use the same notation as in Section 4.2 and let μ_σ denote the uniform distribution on M_σ. *Nonconvex without holes.* Suppose $M_0 \subset \mathbb{R}^2$ is a curve homeomorphic to a line segment, but different from a line segment, and for $\sigma > 0$, let M_σ be the (closed) σ-neighborhood of M_0. We show that there is a numeric constant $\sigma_0 > 0$ such that, when $\sigma < \sigma_0$, ψ does not maximize the energy $\mathcal{E}_\sigma$. To see this, we use Lemma 8 to assert that $\mathcal{S}_{1,\sigma} \to \mathcal{S}_{1,0}$ in the sense of (25), and that $\psi \notin \mathcal{S}_{1,0}$, because $\mathcal{S}_{1,0}$ is made of all the functions that map M to a line segment isometrically. So there is $\sigma_0 > 0$ such that $\psi \notin \mathcal{S}_{1,\sigma}$ for all $\sigma < \sigma_0$. This also implies that no rigid transformation of $\mathbb{R}^2$ is part of $\mathcal{S}_{1,\sigma}$. If we now let $D = M = M_\sigma$ for some $0 < \sigma < \sigma_0$, we see that we do not recover D up to a rigid transformation.

Convex boundary and convex hole. Let K_a denote the axis-aligned ellipse of $\mathbb{R}^2$ with semi-major axis length equal to a and perimeter equal to 2π. Note that, necessarily, $1 \leq a < \pi/2$, with the extreme cases being the unit circle ($a = 1$) and the interval $[-\pi/2, \pi/2]$ swept twice ($a = \pi/2$). Denote by $b = b(a)$ the semi-minor axis length of K_a, implicitly defined by

$$\int_0^{2\pi} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t} \, dt = 2\pi.$$

We have

$$F(a) := \int_{K_a} \|x\|^2 dx = \int_0^{2\pi} \left(a^2 \cos^2 t + b^2 \sin^2 t\right) \sqrt{a^2 \sin^2 t + b^2 \cos^2 t} \, dt.$$

This daunting expression is much simplified when $a = 1$, in which case it is equal to 2π, and when $a = \pi/2$, in which case it is equal to $\pi^2/12$. Since the former is larger than the latter, and F is continuous in a, there is $a_\star$ such that, for $a > a_\star$, $F(a) < F(1)$. (We actually believe that $a_\star = 1$.)

Fix $a \in (a_\star, \pi/2)$ and let $M_0 = K_a = \phi^{-1}(K_1)$, where $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ sends $x = (x_1, x_2)$ to $\phi(x) = (x_1/a, x_2/b)$. Note that K_1 is the unit circle. By the previous calculations and our choice for a, the identity function ψ is not part of $\mathcal{S}_{1,0}$, since

$$\mathcal{E}_0(\psi) = \frac{1}{\pi} \int_{M_0} \|x\|^2 dx = \frac{1}{\pi} F(a) < \frac{1}{\pi} F(1) = 2 = \frac{1}{\pi} \int_{M_0} \|\phi(x)\|^2 dx = \mathcal{E}_0(\phi).$$

As before, let M_σ be the (closed) σ-neighborhood of M_0. Again, there is a numeric constant $\sigma_0 > 0$ such that, when $\sigma < \sigma_0$, ψ does not maximize the energy $\mathcal{E}_\sigma$, and we conclude again that if $D = M = M_\sigma$, MVU does not recover D up to a rigid transformation.

5. Discussion

We leave behind a few interesting problems.

- *Convergence rate for the solution(s).* We obtained a convergence rate for the energy in Theorem 2, but no corresponding result for the solution(s). Such a result necessitates a fine examination of the speed at which the energy decreases near the space of maximizing functions.

- *Flattening property of MVU.* Assume that M satisfies the Isometry Assumption. Though we showed that MVU is not always consistent in the sense that it may not recover the domain D up to a rigid transformation, we believe that MVU always flattens the manifold M in this case, meaning that it returns a set S which is a subset of some d-dimensional affine subspace. If this were true, it would make MVU consistent in terms of dimensionality reduction!

- *Solution space in general.* As pointed out by Paprotny and Garcke (2012), and as we showed in Theorem 1, characterizing the solutions to Continuum MVU is crucial to understanding the behavior of Discrete MVU. In Theorem 3, we worked out the case where M is isometric to a convex set. What can we say when M is isometric to a sphere? Is MVU able to recover this isometry? This question is non-trivial even when M is isometric to a circle. In fact, showing that the energy over ellipses (of same perimeter) is maximized for a circle is not straightforward, as seen in Section 4.3.

We speculate that a similar analysis would show that (Discrete) Isomap (Tenenbaum et al., 2000) converges to Continuum Isomap (Zha and Zhang, 2007). We are curious about the correspondence between Continuum Isomap and Continuum MVU.

Acknowledgments

References

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(16):1373–1396, 2003.

M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. In Peter Auer and Ron Meir, editors, *Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 835–851. Springer Berlin / Heidelberg, 2005. ISBN 978-3-540-26556-6.

M. Bernstein, V. De Silva, J.C. Langford, and J.B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Technical report, Department of Psychology, Stanford University, 2000.

M. Brand. Charting a manifold. *Advances In Neural Information Processing Systems*, pages 985–992, 2003.

A. Brudnyi and Y. Brudnyi. *Methods Of Geometric Analysis In Extension And Trace Problems. Volume 1*, volume 102 of *Monographs in Mathematics*. Birkhäuser/Springer Basel AG, Basel, 2012. ISBN 978-3-0348-0208-6.

D. Burago, Y. Burago, and S. Ivanov. *A Course In Metric Geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001. ISBN 0-8218-2129-6.

R.R. Coifman and S. Lafon. Diffusion maps. *Applied And Computational Harmonic Analysis*, 21 (1):5–30, 2006.

V. H. de la Peña and E. Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. ISBN 0-387-98616-2. From dependence to independence, Randomly stopped processes. *U*-statistics and processes. Martingales and beyond.

D.L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *P. Natl. Acad. Sci. USA*, 100(10):5591–5596, 2003.

H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959. ISSN 0002-9947.

E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High Dimensional Probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006. doi: 10.1214/074921706000000888. URL http://dx.doi.org/10.1214/074921706000000888.

Y. Goldberg, A. Zakai, D. Kushnir, and Y. Ritov. Manifold learning: the price of normalization. *J. Mach. Learn. Res.*, 9:1909–1939, 2008. ISSN 1532-4435.

M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds – weak and strong pointwise consistency of graph laplacians. In Peter Auer and Ron Meir, editors, *Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 470–485. Springer Berlin / Heidelberg, 2005. ISBN 978-3-540-26556-6.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. ISSN 0162-1459.

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961. ISSN 0065-9290.

U. Lang and V. Schroeder. Kirszbraun's theorem and metric spaces of bounded curvature. *Geom. Funct. Anal.*, 7(3):535–560, 1997. ISSN 1016-443X. doi: 10.1007/s000390050018. URL http://dx.doi.org/10.1007/s000390050018.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1):419–441, 2008. ISSN 0179-5376. doi: http://dx.doi.org/10.1007/s00454-008-9053-2.

A. Paprotny and J. Garcke. On a connection between maximum variance unfolding, shortest path problems and isomap. In *Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012.

D. Perrault-Joncas and M. Meila. Metric learning and manifolds: Preserving the intrinsic geometry. Technical report, Department of Statistics, University of Washington, 2012.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

L.K. Saul, K.Q. Weinberger, J.H. Ham, F. Sha, and D.D. Lee. Semisupervised learning. In B. Schoelkopf, O. Chapelle, and A. Zien, editors, *Spectral Methods For Dimensionality Reduction*, pages 293–308. MIT Press, 2006.

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

A.K. Smith, X. Huo, and H. Zha. Convergence and rate of convergence of a manifold-based dimension reduction algorithm. In *Proceedings of Neural Information Processing Systems*, pages 1529–1536. Citeseer, 2008.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

L.J.P. Van der Maaten, E.O. Postma, and H.J. Van den Herik. Dimensionality reduction: A comparative review. Technical report, Tilburg University, 2008.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.

K. Q. Weinberger and Lawrence K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *National Conference on Artificial Intelligence (AAAI)*, volume 2, pages 1683–1686, 2006.

K.Q. Weinberger, F. Sha, and L.K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *International Conference on Machine Learning (ICML)*, page 106, 2004.

K.Q. Weinberger, B.D. Packer, and L.K. Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *International Workshop on Artificial Intelligence and Statistics*, pages 381–388, 2005.

H. Weyl. On the volume of tubes. *Amer. J. Math.*, 61(2):461–472, 1939. ISSN 0002-9327.

Q. Ye and W. Zhi. Discrete hessian eigenmaps method for dimensionality reduction. Technical report, Department of Mathematics, University of Kentucky, 2012.

H. Zha and Z. Zhang. Continuum isomap for manifold learnings. *Computational Statistics & Data Analysis*, 52(1):184–200, 2007.

H.Z. Zha and Z. Zhang. Isometric embedding and continuum isomap. In *In Proceedings of the Twentieth International Conference on Machine Learning*. Citeseer, 2003.

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via tangent space alignment. *SIAM J. Sci. Comput.*, 26(1):313–338, 2004.