Dimensionality Reduction has Quantifiable Imperfections: Two Geometric Bounds

Kry Yik Chau Lui
Borealis AI
Canada
yikchau.y.lui@borealisai.com

Gavin Weiguang Ding
Borealis AI
Canada
gavin.ding@borealisai.com

Ruitong Huang
Borealis AI
Canada
ruitong.huang@borealisai.com

Robert J. McCann
Department of Mathematics
University of Toronto
Canada
mccann@math.toronto.edu

Abstract

In this paper, we investigate Dimensionality reduction (DR) maps in an information retrieval setting from a quantitative topology point of view. In particular, we show that no DR maps can achieve perfect precision and perfect recall simultaneously. Thus a continuous DR map must have imperfect precision. We further prove an upper bound on the precision of Lipschitz continuous DR maps. While precision is a natural measure in an information retrieval setting, it does not measure 'how' wrong the retrieved data is. We therefore propose a new measure based on Wasserstein distance that comes with similar theoretical guarantee. A key technical step in our proofs is a particular optimization problem of the L_2-Wasserstein distance over a constrained set of distributions. We provide a complete solution to this optimization problem, which can be of independent interest on the technical side.

1 Introduction

Dimensionality reduction (DR) serves as a core problem in machine learning tasks including information compression, clustering, manifold learning, feature extraction, logits and other modules in a neural network and data visualization [15, 8, 33, 18, 24]. In many machine learning applications, the data manifold is reduced to a dimension lower than its intrinsic dimension (e.g. for data visualizations, output dimension is reduced to 2 or 3; for classifications, it is the number of classes). In such cases, it is not possible to have a continuous bijective DR map (i.e. classic algebraic topology result on invariance of dimension [25]). With different motivations, many nonlinear DR maps have been proposed in the literature, such as Isomap, kernel PCA, and t-SNE, just to name a few [30, 32, 21]. A common way to compare the performances of different DR maps is to use a down stream supervised learning task as the ground truth performance measure. However, when such down stream task is unavailable, e.g. in an unsupervised learning setting as above, one would have to design a performance measure based on the particular context. In this paper, we focus on the information retrieval setting, which falls into this case. An information retrieval system extracts the features $f(x)$ from the raw data x for future queries. When a new query $y_0 = f(x_0)$ is submitted, the system returns the most relevant data with similar features, i.e. all the x such that $f(x)$ is close to y_0. For computational efficiency and storage, f is usually a DR map, retaining only the most informative features. Assume that the ground truth relevant data of x_0 is defined as a neighbourhood U of x that is a ball with radius r_U centered at

x [1], and the system retrieves the data based on relevance in the feature space, i.e. the inverse image, $f^{-1}(V)$, of a retrieval neighbourhood $V \ni f(x_0)$. Here V is the ball centered at $y_0 = f(x_0)$ with radius r_V that is determined by the system. It is natural to measure the system's performance based on the discrepancy between U and $f^{-1}(V)$. Many empirical measures of this discrepancy have been proposed in the literature, among which precision and recall are arguably the most popular ones [31, 22, 19, 33]. However, theoretical understandings of these measures are still very limited.

In this paper, we start with analyzing the theoretical properties of precision and recall in the information retrieval setting. Naively computing precision and recall in the discrete settings gives undesirable properties, e.g. precision always equals recall when computed by using k nearest neighbors. How to measure them properly is unclear in the literature (Section 3.2). On the other hand, numerous experiments have suggested that there exists a tradeoff between the two when dimensionality reduction happens [33], yet this tradeoff still remains a conceptual mystery in theory. To theoretically understand this tradeoff, we look for continuous analogues of precision and recall, and exploit the geometric and function analytic tools that study dimensionality reduction maps [14]. The first question we ask is what property a DR map should have, so that the information retrieval system can attain zero false positive error (or false negative error) when the relevant neighbourhood U and the retrieved neighbourhood V are properly selected. Our analyses show the equivalence between the achievability of perfect recall (i.e. zero false negative) and the continuity of the DR map. We further prove that no DR map can achieve both perfect precision and perfect recall simultaneously. Although it may seem intuitive, to our best knowledge, this is the first theoretical guarantee in the literature of the necessity of the tradeoff between precision and recall in a dimension reduction setting.

Our main results are developed for the class of (Lipschitz) continuous DR maps. The first main result of this paper is an upper bound for the precision of a continuous DR map. We show that given a continuous DR map, its precision decays exponentially fast with respect to the number of (intrinsic) dimensions reduced. To our best knowledge, this is the first theoretical result in the literature for the decay rate of the precision of a dimensionality reduction map. The second main result is an alternative measure for the performance of a continuous DR map, called W_2 measure, based on L_2-Wasserstein distance. This new measure is more desirable as it can also detect the distance distortion between U and $f^{-1}(V)$. Moreover, we show that our measure also enjoys a theoretical lower bound for continuous DR maps. Several other distance-based measures have been proposed in the literature [31, 22, 19, 33], yet all are proposed heuristically with meagre theoretical understanding. Simulation results suggest optimizing the Wasserstein measure lower bound corresponds to optimizing a weighted f-1 score (i.e. f-β score). Thus we can optimize precision and recall without dealing with their computational difficulties in the discrete settings.

Finally, let us make some comments on the technical parts of the paper. The first key step is the Waist Inequality from the field of quantitative algebraic topology. At a high level, we need to analyse $f^{-1}(V)$, inverse image of an open ball for an arbitrary continuous map f. The waist inequality guarantees the existence of a 'large' fiber, which allows us to analyse $f^{-1}(V)$ and prove our first main result. We further show that in a common setting, a significant proportion of fibers are actually 'large'. For our second main result, a key step in the proof is a complete solution to the following iterated optimization problem:

$$\inf_{W \colon \operatorname{Vol}_n(W)=M} W_2(\mathbb{P}_{B_r}, \mathbb{P}_W) = \inf_{W \colon \operatorname{Vol}_n(W)=M} \inf_{\gamma \in \Gamma(\mathbb{P}_{B_r}, \mathbb{P}_W)} \mathbb{E}_{(a,b)\sim\gamma}[\|a-b\|_2^2]^{1/2},$$

where B_r is a ball with radius r, $\mathbb{P}_{B_r}$ ($\mathbb{P}_W$, respectively) is a uniform distribution over B_r (W, respectively), and W_2 is the L_2-Wasserstein distance. Unlike a typical optimal transport problem where the transport function between source and target distributions is optimized, in the above problem the source distribution is also being optimized at the outer level. This becomes a difficult constrained iterated optimization problem. To address it, we borrow tools from optimal partial transport theory [9, 11]. Our proof techniques leverage the uniqueness of the solution to the optimal partial transport problem and the rotational symmetry of B_r to deduce W.

---

[1]The value of r_U is unknown, and it depends on the user and the input data x_0. However, we can assume r_U is small compared to the input domain size. For example, the number of relevant items to a particular user is much fewer than the number of total items.

1.1 Notations

We collect our notations in this section. Let m be the embedding dimension, $\mathcal{M}$ be an n dimensional data manifold[2] embedded in $\mathbb{R}^N$, where N is the ambient dimension. $\mathcal{M}$ is typically modelled as a Riemannian manifold, so it is a metric space with a volume form. Let $m < n < N$ and $f : \mathcal{M} \subset \mathbb{R}^N \to \mathbb{R}^m$ be a DR map. The pair (x, y) will be the points of interest, where $y = f(x)$. The inverse image of y under the map f is called fiber, denoted $f^{-1}(y)$. We say f is continuous at point x iff $\mathrm{osc}^f(x) = 0$, where $\mathrm{osc}^f(x) = \inf_{U; U \text{open}}\{\mathrm{diam}(f(U)); x \in U\}$ is the oscillation for f at $x \in \mathcal{M}$. We say f is *one-to-one* or *injective* when its fiber, $f^{-1}(y)$ is the singleton set $\{x\}$.

We let $A \oplus \epsilon := \{x \in \mathbb{R}^N | \mathrm{d}(x, A) < \epsilon\}$ denote the ϵ-neighborhood of the nonempty set A. In $\mathbb{R}^N$, we note the ϵ-neighborhood of the nonempty set A is the Minkowski sum of A with $B_\epsilon^N(x)$, where the Minkowski sum between two sets A and B is: $A \oplus B = \{a + b | a \in A, b \in B\}$. For example, an n dimension open ball with radius r, centered at a point x can be expressed as: $B_r^n(x) = x \oplus B_r^n(0) = x \oplus r$, where the last expression is used to simplify notation. If not specified, the dimension of the ball is n. We also use B_r to denote the ball with radius r when its center is irrelevant. Similarly, S_r^n denotes n-dimensional sphere in $\mathbb{R}^{n+1}$ with radius r. Let Vol_n denote n-dimensional volume.[3] When the intrinsic dimension of A is greater than n, we set $\mathrm{Vol}_n(A) = \infty$. Through the rest of the paper, we use U to denote $B_{r_U}(x)$ a ball with radius r_U centered at x and $V = B_{r_V}(y)$ a ball with radius r_V centered at y. These are metric balls in a metric space. For example, they are geodesic balls in a Riemannian manifold, whenever they are well defined. In Euclidean spaces, U is a Euclidean ball with L_2 norm. By $T_\#(\mu) = \nu$, we mean a map T pushes forward a measure μ to ν, i.e. $\nu(B) = \mu(T^{-1}(B))$ for any Borel set B. We say a measure μ is dominated by another measure ν, if for every measurable set A, $\mu(A) \leq \nu(A)$.

2 Precision and recall

We present the definitions of precision and recall in a continuous setting in this section. We then prove the equivalence between perfect recall and the continuity, followed by a theorem on the necessary tradeoff between the perfect recall and the perfect precision for a dimension reduction information retrieval system. The main result of this section is a theoretical upper bound for the precision of a continuous DR map.

2.1 Precision and recall

While precision and recall are commonly defined based on finite counts in practice, when analysing DR maps between spaces, it is natural to extend their definitions in a continuous setting as follows.

Definition 1 (Precision and Recall)**.** *Let f be a continuous DR map. Fix $(x, y = f(x))$, $r_U > 0$ and $r_V > 0$, let $U = B_{r_U}(x) \subset \mathbb{R}^N$ and $V = B_{r_V}^m(y) \subset \mathbb{R}^m$ be the balls with radius r_U and r_V respectively. The **precision** and **recall** of f at U and V are defined as:*

$$Precision^f(U, V) = \frac{Vol_n(f^{-1}(V) \cap U)}{Vol_n(f^{-1}(V))}; \qquad Recall^f(U, V) = \frac{Vol_n(f^{-1}(V) \cap U)}{Vol_n(U)}.$$

*We say f achieves **perfect precision** at x if for every r_U, there exists r_V such that $Precision^f(U, V) = 1$. Also, f achieves **perfect recall** at x if for every r_V, there exists r_U such that $Recall^f(U, V) = 1$. Finally, we say f achieves **perfect precision** (**perfect recall**, respectively) in an open set W, if f achieves perfect precision (perfect recall, respectively) at w for any $w \in W$.*

Note that perfect precision requires $f^{-1}(V) \subset U$ except a measure zero set. Similarly, perfect recall requires $U \subset f^{-1}(V)$ except a measure zero set. Figure 1 illustrates the precision and recall defined above. To measure the performance of the information retrieval system, we would like to understand how different $f^{-1}(V)$ is from the ideal response $U = B_{r_U}(x)$. Precision and recall provides two meaningful measures for this difference based on their volumes. Note that f achieves

---

[2]There is empirical and theoretical evidence that data distribution lies on low dimensional submanifold in the ambient space [26].

[3] Let A be a set. In Euclidean space, $\mathrm{Vol}_n(A) = \mathcal{L}^n(A)$ is the Lebesgue measure. For a general n-rectifiable set, $\mathrm{Vol}_n(A) = \mathcal{H}^n(A)$ is the Hausdorff measure. When A is not rectifiable, $\mathrm{Vol}_n(A) = \mathcal{M}_*^n(A)$ is the lower Minkowski content.
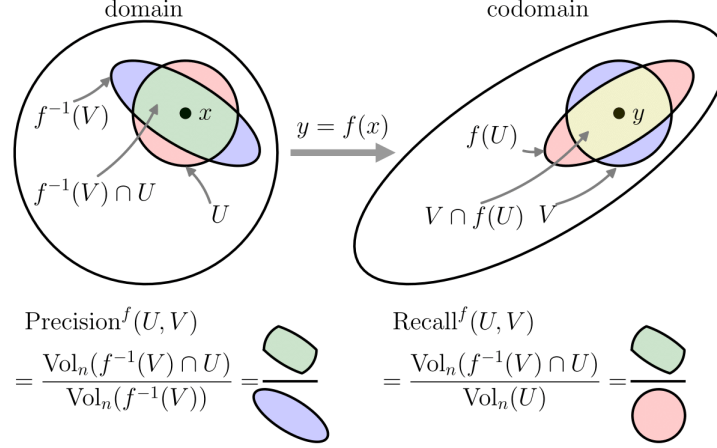
Figure 1: Illustration of precision and recall.

perfect precision at x implies that no matter how small the relevant radius r_U is for the image, the system would be able to achieve 0 false positive by picking proper r_V. Similarly perfect recall at x implies no matter how small r_V is, the system would not miss the most relevant images around x.

In fact, the definitions of perfect precision and perfect recall are closely related to continuity and injectivity of a function f. Here we only present an informal statement. Rigorous statements are given in the Appendix B.

Proposition 1. *Perfect recall is equivalent to continuity. If f is continuous, then perfect precision is equivalent to injectivity.*

The next result shows that no DR map f, continuous or not, can achieve perfect recall and perfect precision simultaneously - a widely observed but unproved phenomenon in practice. In other words, it rigorously justifies the intuition that perfectly maintaining the local neighbourhood structure is impossible for a DR map.

Theorem 1 (Precision and Recall Tradeoff). *Let $n > m$, $\mathcal{M} \subset \mathbb{R}^N$ be a Riemannian n-dimensional submanifold. Then for any (dimensionality reduction) map $f : \mathcal{M} \to \mathbb{R}^m$ and any open set $W \subset \mathcal{M}$, f cannot achieve both perfect precision and perfect recall on W.*

2.2 Upper bound for the precision of a continuous DR map

In this section, we provide a quantitative analysis for the imperfection of f. In particular, we prove an upper bound for the precision of a continuous DR map f (thus f achieves perfect recall). For simplicity, we assume the domain of f is an n-ball with radius R embedded in $\mathbb{R}^N$, denoted by B_R^n. Our main tool is the Waist Inequality [28, 1] in quantitative topology. See Appendix A for an exact statement.

Intuitively, the Waist Inequality guarantees the existence of $y \in \mathbb{R}^m$ such that $f^{-1}(y)$ is a 'large' fiber. If f is also L-Lipschitz, then for p in a small neighbourhood V of y, $f^{-1}(p)$ is also a 'large' fiber, thus $f^{-1}(V)$ has a positive volume in $\mathcal{M}$. Exploiting the lower bound for $\mathrm{Vol}_n\left(f^{-1}(V)\right)$ leads to our upper bound in Theorem 2 on the precision of f, $\mathrm{Precision}^f(U, V)$. A rigorous proof is given in the appendix Appendix C.

Theorem 2 (Precision Upper Bound, Worst Case). *Assume $n > m$, and that $f : B_R^n \to \mathbb{R}^m$ is a continuous map with Lipschitz constant L. Let r_U and $r_V > 0$ be fixed. Denote*

$$D(n, m) = \frac{\Gamma(\frac{n-m}{2} + 1)\Gamma(\frac{m}{2} + 1)}{\Gamma(\frac{n}{2} + 1)} . \tag{1}$$

Then there exists $y \in \mathbb{R}^m$ such that for any $x \in f^{-1}(y)$, we have:

$$\mathrm{Precision}^f(U, V) \leq D(n, m) \left(\frac{r_U}{R}\right)^{n-m} \frac{r_U^m}{p^m(r_V/L)} \tag{2}$$

4

where $p^m(r)$ is $r^m(1 + o(1))$, i.e. $\lim_{r \to 0} \frac{p^m(r)}{r^m} = 1$.

Remark 1. *Key to the bound is the waist inequality. As such, upper bounds on precision for other spaces (i.e. cube, see Klartag [16]) can be established, provided there is a waist inequality for the space. The Euclidean norm setting can also be extended to arbitrary norms, exploiting convex geometry (i.e. Akopyan and Karasev [2]). Rigorous proofs are given in the appendix C.*

Remark 2. *With m fixed as a constant, note that $D(n, m)$ decays asymptotically at a rate of $(1/n)^{m/2}$. Also note that $r_U < R$ implies $\left(\frac{r_U}{R}\right)^{n-m}$ decays exponentially. Typically, L can grow at a rate of $\sqrt{n}$. Moreover, while $p^m(r)$'s behaviour is given asymptotically, it is independent of n. Thus the upper bound decay is dominated by the exponential rate of $n - m$. For fixed n, m, this upper bound can be trivial when $r_U \gg r_V$. However, this rarely happens in practice in the information retrieval setting. Note that the number of relevant items, which is indexed by r_U, is often smaller than the number of retrieved items, that depends on r_V, while they are both much smaller than number of total items, indexed by R.*

We note however that this bound depends on the intrinsic dimension n. When $n \ll N$ and the ambient dimension N is used in place, the upper bound could be misleading in practice as it is much smaller than it should be. To estimate this bound in practice, a good estimate on intrinsic dimension [12] is needed, which is an active topic in the field and beyond the scope of this paper.

Theorem 2 guarantees the existence of a particular point $y \in \mathbb{R}^m$ where the precision of f on its neighbourhood is small. It is natural to ask if this is also true in an average sense for every y. In other words, we know a information retrieval system based on DR maps always has a blindspot, but is this blindspot behaviour a typical case? In general, when $m > 1$, this is false, due to a recent counter-example constructed by Alpert and Guth [3]. However, our next result shows that for a large number of continuous DR maps in the field, such upper bound still holds with high probability.

Theorem 3 (Precision Upper Bound, Average Case). *Assume $n > m$ and B_R^n is equiped with uniform probability distribution. Consider the following cases:*

- **case 1:** $m = 1$ *and* $f : B_R^n \to \mathbb{R}^m$ *is L Lipschitz continuous, or*

- **case 2:** $f : B_R^n \to \mathbb{R}^m$ *is a k-layer feedforward neural network map with Lipschitz constant L, with surjective linear maps in each layer.*

Let $0 < \delta^2 < R^2 - r_U^2$, $r_U, r_V > 0$ be fixed, then with probability at least q_1 for case 1 or q_2 for case 2, it holds that

$$Precision^f(U, V) \leq D(n, m) \left(\frac{r_U}{\sqrt{r_U^2 + \delta^2}} \right)^{n-m} \frac{r_U^m}{p^m(r_V/L)}, \tag{3}$$

where

$$q_1 = \frac{\frac{1}{2\pi R} \int_{B_{\Re}^m} Vol_{n-m+1} Proj_1^{-1}(t) dt}{Vol_n(B_R^n)}, \quad q_2 = \frac{\int_{B_{\Re}^m} Vol_{n-m} Proj_2^{-1}(t) dt}{Vol_n(B_R^n)},$$

$\Re = \sqrt{R^2 - r_U^2 - \delta^2}$, $Proj_1 : S_R^{n+1} \to \mathbb{R}^m$ and $Proj_2 : B_R^n \to \mathbb{R}^m$ are arbitrary surjective linear maps. Furthermore,

$$\lim_{\frac{r_U^2 + \delta^2}{R^2} \to 0} q_1 = 1 \quad \lim_{\frac{r_U^2 + \delta^2}{R^2} \to 0} q_2 = 1.$$

See Appendix Appendix D for an explicit characterization of $Proj_1^{-1}(t)$ and $Proj_2^{-1}(t)$. Theorem 2 and Theorem 3 together suggest that practioners should be cautious in applying and interpreting DR maps. One important application of DR maps is in data visualization. Among the many algorithms, t-SNE's empirical success made it the de facto standard. While [5] shows t-SNE can recover inter-cluster structure in some provable settings, the resulted intra-cluster embedding will very likely be subject to the constraints given in our work [4]. For example, recall within a cluster will be good, but the intra-cluster precision won't be. In more general cases and/or when perplexity is too small,

---

[4]Technically speaking, the DR maps induced by t-SNE may not be continuous, and hence our theorems do not apply directly. However, since we can measure how closely parametric t-SNE (which is continuous) behaves as t-SNE and there is empirical evidence to their similarity [20], our theorems may apply again.

t-SNE can create artificial clusters, separating neighboring datapoints. The resulted visualization embedding may enjoy higher precision, but its recall suffers. The interested readers are referred to Appendix G.1 for more experimental illustrations. Our work thus sheds light on the inherent tradeoffs in any visualization embedding. It also suggests the companion of a reliability measure to any data visualization for exploratory data analysis, which measures how a low dimensional visualization represents the true underlying high dimensional neighborhood structure.[5]

3   Wasserstein measure

Intuitively we would like to measure how different the original neighbourhood U of x is from the retrieved neighbourhood $f^{-1}(V)$ when using the neighbourhood of $f(x)$ in $\mathbb{R}^m$. Precision and Recall in Section 2.1 provide a semantically meaningful way for this purpose and we gave a non-trivial upper bound for precision when the feature extraction is a continuous DR map. However, precision and recall are purely volume-based measures. It would be more desirable if the measure could also reflect the information about the distance distortions between U and $f^{-1}(V)$. In this section, we propose an alternative measure to reflect such information based on the L_2-Wasserstein distance. Efficient algorithms for computing the empirical Wasserstein distance exists in the literature [4]. Unlike the measure proposed in Venna et al. [33], our measure also enjoys a theoretical guarantee similar to Theorem 2, which provides a non-trivial characterization for the imperfection of dimension reduction information retrieval.

Let $\mathbb{P}_U$ ($\mathbb{P}_{f^{-1}(V)}$, respectively) denote the uniform probability distribution over U ($f^{-1}(V)$, respectively), and $\Xi(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)})$ be the set of all the joint distribution over $B_R^n \times B_R^n$, whose marginal distributions are $\mathbb{P}_U$ over the first B_R^n and $\mathbb{P}_{f^{-1}(V)}$ over the second B_R^n. We propose to measure the difference between U and $f^{-1}(V)$ by the L_2-Wasserstein distance between $\mathbb{P}_U$ and $\mathbb{P}_{f^{-1}(V)}$:

$$W_2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)}) = \inf_{\xi \in \Xi(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)})} \mathbb{E}_{(a,b) \sim \xi}[\|a - b\|_2^2]^{1/2}.$$

In practice, it is reasonable to assume that $\mathrm{Vol}_n(U)$ is small in most retrieval systems. In such cases, low $W_2(P_U, P_{f^{-1}(V)})$ cost is closely related to high precision retrieval. To see that, when $\mathrm{Vol}_n(U)$ is small, achieving high precision retrieval requires small $\mathrm{Vol}_n(f^{-1}(V))$, which is a precise quantitative way of saying f being roughly injective. Moreover, as seen in Section 2.1, f being roughly injective $\approx f$ giving high precision retrieval. As a result, we can expect high precision retrieval performance when optimizing $W_2(P_U, P_{f^{-1}(V)})$ measure. Such relation is also empirically confirmed in the simulation in Section 3.2.

Besides its computational benefits, for a continuous DR map f, the following theorem provides a lower bound on $W_2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)})$ with a similar flavour to the precision upper bound in Theorem 1.

Theorem 4 (Wasserstein Measure Lower Bound). *Let $n > m$, $f : B_R^n \to \mathbb{R}^m$ be a L-Lipschitz continuous map, where R is the radius of the ball B_R^n. There exists $y \in \mathbb{R}^m$ such that for any $x \in f^{-1}(y)$, r_U and $r_V > 0$ such that $r \geq r_U$,*

$$W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)}) \geq \frac{n}{n+2} (r - r_U)^2$$

where $r = \left(\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n-m}{2}+1)\Gamma(\frac{m}{2}+1)} \right)^{\frac{1}{n}} R^{\frac{n-m}{n}} (p^m(r_V/L))^{\frac{1}{n}}$. In particular, as $n \to \infty$,

$$W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)}) = \Omega\left((R - r_U)^2 \right).$$

We sketch the proof here. A complete proof can be found in Appendix E. The proof starts with a lower bound of $\mathrm{Vol}_n\left(f^{-1}(V)\right)$ by the topologically flavoured waist inequality (Equation (6)). Heuristically $\mathrm{Vol}_n(f^{-1}(V))$ is much larger than $\mathrm{Vol}_n(U)$ when $n \gg m$ and $R \gg r_U$. The main component of the proof is to establish an explicit lower bound for $W_2(\mathbb{P}_U, \mathbb{P}_W)$ over all possible W of a fixed volume $\mathcal{V}$,[6] where U is a ball with radius r_U, as shown in Theorem 5. In particular, we prove that the shape

---

[5]Such attempts existed in literature on visualization of dimensionality reduction (e.g. [33]). However, since these works are based on heuristics, it is less clear what they measure, nor do they enjoy theoretical guarantee.

[6]An antecedent of this problem was studied in Section 2.3 of [23], where the authors optimize over the more restricted class of ellipses with fixed area. For our purpose, the minimization is over bounded measurable sets.

of optimal W^* must be rotationally invariant, thus W^* must be a union of spheres. This is achieved by levering the uniqueness of the solution to the optimal partial transport problem [9, 11]. We then prove that the optimal solution for W is the ball that has a common center with U.

Theorem 5. *Let $U = B_{r_U}$ and $\mathcal{V} \geq Vol(U)$. Then*

$$\inf_{W:\,Vol_n(W)\geq\mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = \inf_{W:\,Vol_n(W)=\mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = W_2(\mathbb{P}_U, \mathbb{P}_{B_{r_\mathcal{V}}}),$$

where $B_{r_\mathcal{V}}$ is an $r_\mathcal{V}$ ball with the same center with U such that $Vol_n(B_{r_\mathcal{V}}) = \mathcal{V}$. Moreover, $T(x) = \frac{r_U}{r_\mathcal{V}} x$, for $x \in B_{r_\mathcal{V}}$ is the optimal transport map (up to a measure zero set), so that

$$W_2(\mathbb{P}_U, \mathbb{P}_{B_{r_\mathcal{V}}}) = \int_{B_{r_\mathcal{V}}} |x - T(x)|^2 \, d\mathbb{P}_{B_{r_\mathcal{V}}}(x).$$

Complementarily, when $0 < \mathcal{V} < Vol_n(U)$, the infimum $\inf_{W:\,Vol_n(W)=\mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = 0$, is not attained by any set. On the other hand, $\inf_{W:\,Vol_n(W)\geq\mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = 0$ by taking $W = U$.

Remark 3. *Our lower bound in Theorem 4 is (asymptotically) tight. Note that by Theorem 4, $W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)})$ has a (maximum) lower bound of scale $(R - r_U)^2$. On the other hand, by Theorem 5, $W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)}) \leq W_2^2(\mathbb{P}_U, \mathbb{P}_{B_R^n}) = \Omega((R - r_U)^2)$, where the equality is by standard algebraic calculations.*

3.1 Iso-Wasserstein inequality

We believe Theorem 5 is of independent interest itself, as it has the same flavor as the isoperimetric inequality (See Appendix A for an exact statement.) which arguably is the most important inequality in metric geometry. In fact, the first statement of Theorem 5 can be restated as the following inequality:

Theorem 6 (Iso-Wasserstein Inequality). *Let $B_{r_1}, B_{r_2} \subset B_R^n$ be two concentric n balls with radii $r_1 \leq r_2$ centered at the origin. For all measurable $A \subset B_R^n$ with $Vol_n(A) = Vol_n(B_{r_2})$, we have*

$$W_2(\mathbb{P}(A), \mathbb{P}(B_{r_1})) \geq W_2(\mathbb{P}(B_{r_2}), \mathbb{P}(B_{r_1}))$$

where $\mathbb{P}(S)$ denotes a uniform probability distribution on S, i.e. $\mathbb{P}(S)$ has density $\frac{1}{Vol_n(S)}$.

Recall that an isoperimetric inequality in Euclidean space roughly says balls have the least perimeter among all equal volume sets. Theorem 6 acts as a transportation cousin of the isoperimetric inequality. While the isoperimetric inequality compares $n - 1$ volume between two sets, the iso-Wasserstein inequality compares their Wasserstein distances to a small ball. The extrema in both inequalities are attained by Euclidean balls.

3.2 Simulations

In this section, we demonstrate on a synthetic dataset that our lower bound in Theorem 4 can be a reasonable guidance for selecting the retrieval neighborhood radius r_V, which emphasizes on high precision. The simulation environment is to compute the optimal r_V by minimizing the lower bound in Theorem 4, with a given relevant neighborhood radius r_U and embedding dimension m. Note that minimizing its lower bound instead of the exact cost itself is beneficial as it avoids the direct computation of the cost. Recall the lower bound of $W_2(P_U, P_{f^{-1}(V)})$ is (asymptotically) tight (Remark 3) and matches the its upper bound when $n - m \gg 0$. If the lower bound behaves roughly like $W_2(P_U, P_{f^{-1}(V)})$, our simulation result also serves as an empirical evidence that $W_2(P_U, P_{f^{-1}(V)})$ weighs more on high precision.

Specifically, we generate 10000 uniformly distributed samples in a 10-dimensional unit ℓ_2-ball. We choose r_U such that on average each data point has 500 neighbors inside B_{r_U}. We then linearly project these 10 dimensional points into lower dimensional spaces with embedding dimension m from 1 to 9. For each m, a different r_V is used to calculate discrete precision and recall. This simulates how optimal r_V according to Wasserstein measure changes with respect to m. The result is shown in on the left in Figure 2. Similarly, we can fix $m = 5$ and track optimal r_V's behavior when r_U changes. This is shown on the right in Figure 2.

We evelute our measures based on traditional information retrieval metrics such as f-score. To compute it, we need the discrete/sample-based precision and recall. As discussed in the introduction,
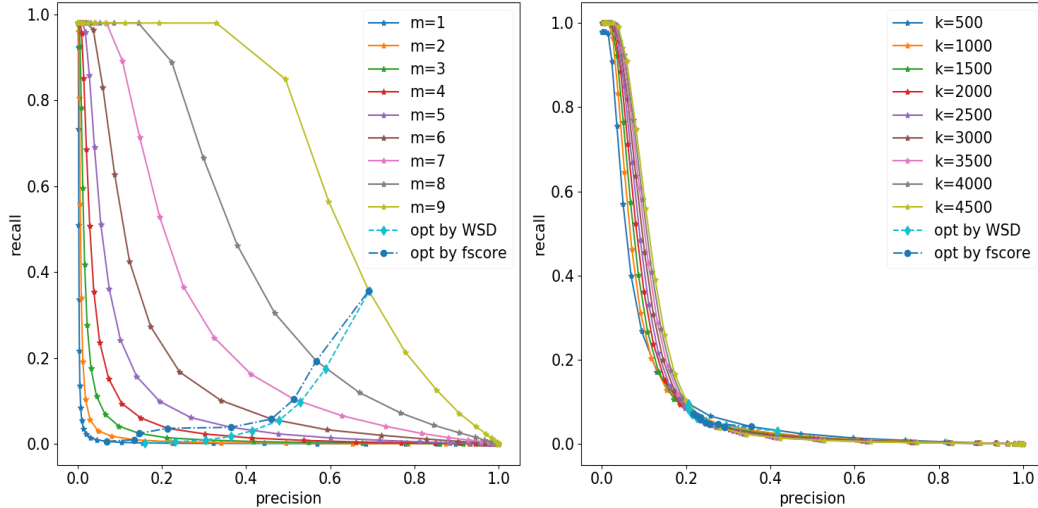
Figure 2: Precision and recall results on uniform samples in a 10 dimensional unit ball. The left figure contains precision-recall curves for a fixed r_U and the optimal r_V is chosen according to $m = 1, \cdots, 9$. The right figure plots the curves for $m = 5$ and the optimal r_V's is chosen for different r_U, where r_U is indexed by k, the average number of neighbors across all points.

a naive sample based calculations of precision and recall makes $Precision = Recall$ at all times. We compute them alternatively by discretizing Definition 1, by fixing radii r_U and r_V. So each U and $f^{-1}(V)$ contain different numbers of neighbors.

$$Precision = \frac{\#(\text{points within } r_U \text{ from } x \text{ and within } r_V \text{ from } y)}{\#(\text{points within } r_V \text{ from } y)} \tag{4}$$

$$Recall = \frac{\#(\text{points within } r_U \text{ from } x \text{ and within } r_V \text{ from } y)}{\#(\text{points within } r_U \text{ from } x)} \tag{5}$$

The optimal r_V according to the lower bound in Theorem 4 (the blue circle-dash-dotted line) aligns closely with the optimal f-score with $\beta = 0.3$ where β weighted f-score, also known as f-βscore, is:

$$(1 + \beta^2)\frac{Precision * Recall}{\beta^2 * Precision + recall}.$$

Note that f-score with $\beta < 1$ indeed emphasizes on high precision.

In this provable setting, we have demonstrated our bound's utility. This shows W_2 measures' potential for evaluating dimension reduction. In general cases, we won't have such tight lower bounds and it is natural to optimize according to the sample based W_2 measures instead. We performed some preliminary experiments on this heuristic, shown in Appendix G.

4 Relation to metric space embedding and manifold learning

We lastly situate our work in the lines of research on metric space embedding and manifold learning. One obvious difference between our work and the literature of metric space embedding and manifold learning is that our work mainly focuses on intrinsic dimensionality reduction maps, i.e. $n \gg m$, while in metric space embedding and manifold learning, having $n \le m < N$ is common.

Our work also differs from the literature of metric space embedding and manifold learning in its learning objective. Learning in these fields aims to preserve the metric structure of the data. Our work attempts to preserve precision and recall, a weaker structure in the sense of embedding dimension (Proposition 2). While they typically look for lowest embedding dimension subject to certain loss (e.g. smoothness, local or global isometry), in contrast, our learning goal is to minimize the loss (precision and recall etc.) subject to a fixed embedding dimension constraint. In these cases, desired structures will break (Theorem 3) because we cannot choose the embedding dimension m (e.g. for visualizations $m = 2$; for classifications $m = $ number of classes).

We now discuss the technical relations with metric space embedding and manifold learning. Many datasets can be modelled as a finite metric space $\mathcal{M}_k$ with k points. A natural unsupervised learning task is to learn an embedding that approximately preserves pairwise distances. The Bourgain embedding [7] guarantees the metric structure can be preserved with distortion $O(\log k)$ in $l_p^{O(\log^2 k)}$. When the samples are collected in Euclidean spaces, i.e. $\mathcal{M}_k \subset l_2$, the Johnson-Lindenstrauss lemma [10] improves the distortion to $(1 + \epsilon)$ in $l_2^{O(\log(k/\epsilon^2))}$. These embeddings approximately preserve all pairwise distances - global metric structure of $\mathcal{M}_k$ is compatible to the ambient vector space norms. Coming back to our work, it is natural to mimic this approach for precision and recall in $\mathcal{M}_k$. The first problem is that the naive sample based precision and recall are always equal (Section 3.2). A second problem is discrete precision and recall is a non-differentiable objective. In fact, the difficulty of analyzing discrete precision and recall motivates us to look for continuous analogues.

Roughly, our approach is somewhat similar to manifold learning where researchers postulate that the data $\mathcal{M}_k$ are sampled from a continuous manifold $\mathcal{M}$, typically a smooth or Riemannian manifold $\mathcal{M}$ with intrinsic dimension n. In this setting, one is interested in embedding $\mathcal{M}$ into l_2 locally isometrically. Then one designs learning algorithms that can combine the local information to learn some global structure of $\mathcal{M}$. By relaxing to the continuous cases just like our setting, manifold learning researchers gain access to vast literature in geometry. By the Whitney embedding [24], $\mathcal{M}$ can be smoothly embedded into $\mathbb{R}^{2n}$. By the Nash embedding [34], a compact Riemannian manifold $\mathcal{M}$ can be isometrically embedded into $\mathbb{R}^{p(n)}$, where $p(n)$ is a quadratic polynomial. Hence the task in manifold learning is wellposed: one seeks an embedding $f : \mathcal{M} \subset \mathbb{R}^N \to \mathbb{R}^m$ with $m \leq 2n \ll N$ in the smooth category or $m \leq p(n) \ll N$ in the Riemannian category. Note that the embedded manifold metrics (e.g. the Riemannian geodesic distances) are not guaranteed to be compatible to the ambient vector space's norm structure with a fixed distortion factor, unlike the Bourgain embedding or the Johnson-Lindenstrauss lemma in the discrete setting. A continuous analogue of the norm compatible discrete metric space embeddings is the Kuratowski embedding, which embeds global-isometrically (preserving pairwise distance) any metric space to an infinite dimensional Banach space L^∞. With ϵ distortion relaxation, it is possible to embed a compact Riemannian manifold to a finite dimensional normed space. But this appears to be very hard, in that the embedding dimension may grow faster than exponentially in n [29].

Like DR in manifold learning and unlike DR in discrete metric space embedding, rather than global structure we want to preserve local notions such as precision and recall. Unlike DR in manifold learning, since precision and recall are almost equivalent to continuity and injectivity (Theorem 1), we are interested in embeddings in the topological category, instead of the smooth or the Riemannian category. Thus, our work can be considered as manifold learning from the perspective of information retrieval, which leads to the following result.

Proposition 2. *If $m \geq 2n$, where n is the dimension of the data manifold $\mathcal{M}$ in domain and m is the dimension of codomain $\mathbb{R}^m$, then there exists a continuous map $f : \mathcal{M} \to \mathbb{R}^m$ such that f achieves perfect precision and recall for every point $x \in \mathcal{M}$.*

Note that the dimension reduction rate is actually much stronger than the case of Riemannian isometric embedding where the lowest embedding dimension grows polynomially [34]. This is because preserving precision and recall is weaker than isometric embedding. A practical implication is that, we can reduce many more dimensions if we only care about precision and recall.

5   Conclusions

We characterized the imperfection of dimensionality reduction mappings from a quantitative topology perspective. We showed that perfect precision and perfect recall cannot be both achieved by any DR map. We then proved a non-trivial upper bound for precision for Lipschitz continuous DR maps. To further quantify the distortion, we proposed a new measure based on L_2-Wasserstein distances, and also proved its lower bound for Lipschitz continuous DR maps. It is also interesting to analyse the relation between the recall of a continuous DR map and its modulus of continuity. However, the generality and complexity of the fibers (inverse images) of these maps so far defy our effort and this problem remains open. Furthermore, it is interesting to develop a corresponding theory in the discrete setting.

6 Acknowledgement

References

[1] Arseniy Akopyan and Roman Karasev. A tight estimate for the waist of the ball. *Bulletin of the London Mathematical Society*, 49(4):690–693, 2017.

[2] Arseniy Akopyan and Roman Karasev. Waist of balls in hyperbolic and spherical spaces. *International Mathematics Research Notices*, page rny037, 2018. doi: 10.1093/imrn/rny037. URL http://dx.doi.org/10.1093/imrn/rny037.

[3] Hannah Alpert and Larry Guth. A family of maps with many small fibers. *Journal of Topology and Analysis*, 7(01):73–79, 2015.

[4] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1961–1971, 2017.

[5] Sanjeev Arora, Wei Hu, and Pravesh K. Kothari. An analysis of the t-sne algorithm for data visualization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1455–1462. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/arora18a.html.

[6] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. In *ACM Transactions on Graphics (TOG)*, volume 30, page 158. ACM, 2011.

[7] Jean Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52(1):46–52, 1985.

[8] Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In *NIPS*, pages 298–306, 2010.

[9] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and Monge-Ampére obstacle problems. *Annals of mathematics*, 171:673–730, 2010.

[10] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[11] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

[12] Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports*, 6, 2016.

[13] Victor Guillemin and Alan Pollack. *Differential topology*, volume 370. American Mathematical Soc., 2010.

[14] LARRY Guth. The waist inequality in gromov's work. *The Abel Prize 2008*, pages 181–195, 2012.

[15] Gísli R. Hjaltason and Hanan Samet. Properties of embedding methods for similarity searching in metric spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(5): 530–549, May 2003. ISSN 0162-8828. doi: 10.1109/TPAMI.2003.1195989. URL https://doi.org/10.1109/TPAMI.2003.1195989.

[16] Bo'az Klartag. Convex geometry and waist inequalities. *Geometric and Functional Analysis*, 27(1):130–164, 2017.

[17] Jonathan Korman and Robert J McCann. Insights into capacity-constrained optimal transport. *Proceedings of the National Academy of Sciences*, 110(25):10064–10067, 2013.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[19] Sylvain Lespinats and Michaël Aupetit. Checkviz: Sanity check and topological clues for linear and non-linear mappings. In *Computer Graphics Forum*, volume 30, pages 113–125. Wiley Online Library, 2011.

[20] Laurens Maaten. Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics*, pages 384–391, 2009.

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[22] Rafael Messias Martins, Danilo Barbosa Coimbra, Rosane Minghim, and Alexandru C Telea. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics*, 41:26–42, 2014.

[23] Robert J McCann and Adam M Oberman. Exact semi-geostrophic flows in an elliptical ocean basin. *Nonlinearity*, 17(5):1891, 2004.

[24] James McQueen, Marina Meila, and Dominique Joncas. Nearly isometric embedding by relaxation. In *NIPS*, pages 2631–2639, 2016.

[25] Michael Müger. A remark on the invariance of dimension. *Mathematische Semesterberichte*, 62(1):59–68, 2015.

[26] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *NIPS*, pages 1786–1794, 2010.

[27] Lawrence E Payne. Isoperimetric inequalities and their applications. *SIAM review*, 9(3): 453–488, 1967.

[28] P Rayón and M Gromov. Isoperimetry of waists and concentration of maps. *Geometric & Functional Analysis GAFA*, 13(1):178–215, 2003.

[29] Malte Roeer. On the finite dimensional approximation of the Kuratowski-embedding for compact manifolds. *arXiv preprint arXiv:1305.1529*, 2013.

[30] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

[31] Tobias Schreck, Tatiana Von Landesberger, and Sebastian Bremm. Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9(3):181–193, 2010.

[32] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[33] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(Feb):451–490, 2010.

[34] Nakul Verma. Distance preserving embeddings for general n-dimensional manifolds. *Journal of Machine Learning Research*, 14(1):2415–2448, 2013.

[35] Xianfu Wang. Volumes of generalized unit balls. *Mathematics Magazine*, 78(5):390–395, 2005.

A Waist Inequality and Isoperimetric Inequality

Theorem 7 (Waist Inequality, Akopyan and Karasev [1]). *Let $m \leq n$ and f be a continuous map from the ball B_R^n of radius R to $\mathbb{R}^m$. Then there exists some $y \in \mathbb{R}^m$ such that*

$$Vol_{n-m}\left(f^{-1}(y)\right) \geq Vol_{n-m}\left(B_R^{n-m}\right).{}^7$$

Moreover, for all $\epsilon > 0$:

$$Vol_n\left(f^{-1}(y) \oplus \epsilon\right) \geq \frac{1}{2\pi R} Vol_{n-m+1}\left(S_R^{n-m+1}\right) Vol_m\left(B_1^m\right) p^m(\epsilon), \tag{6}$$

where $p^m(\epsilon)$ is $\epsilon^m\left(1 + o(1)\right)$, i.e. $\lim_{\epsilon \to 0} \frac{p^m(\epsilon)}{\epsilon^m} = 1$, and $f^{-1}(y) \oplus \epsilon$ denotes the set of points $x \in B_R^n$ such that $d(x, f^{-1}(y)) < \epsilon$, S_R^{n-m+1} is the (n+m-1)-dimensional sphere of radius R, and B_1^m is the unit m ball.

Remark 4. *When $m = 1$, Waist Inequality generalizes classic concentration of measure on B_R^n, which says most volume of a high dimensional ball concentrates around its equator slab, as $n \to \infty$. When $m > 1$, we can roughly interpret the theorem as $f^{-1}(y) \oplus \epsilon$ is big in $n - m$ dimensions in the sense of volume, thus it generalizes concentration of measure when $m > 1$.*

Intuitively the Waist inequality states that a higher dimensional space is too big in the sense of **volume** that we cannot hope to squeeze it **continuously** into lower dimensional spaces, without collapsing in some direction(s). In other words, if an input domain is higher dimensional and thus in some sense large, then it must be large in at least one direction. Waist inequality is a precise quantitative version of the topological invariance of dimension, which states balls of different dimensions cannot be homeomorphically mapped to each other. It is this mis-match between high and low dimensional nature of volumes that motivates us to formulate and prove the imperfection between precision and recall. A recent survey of the inequality can be found in [14].

Theorem 8 (Isoperimetric Inequality). *Suppose $U \subset \mathbb{R}^n$ is a bounded (Hausdorff) measurable set, with (Hausdorff) $n - 1$ measurable boundary, denoted as $Vol_{n-1}\partial U$. Then:*

$$Vol_n(U) = Vol_n(B_1^n) \implies Vol_{n-1}(\partial U) \geq Vol_{n-1}(\partial B_1^n)$$

Stated differently,

$$Vol_n(U) \leq \frac{1}{n^{\frac{n}{n-1}} Vol_n(B_1)^{\frac{1}{n-1}}} Vol_{n-1}(\partial U)^{\frac{n}{n-1}}$$

The first way of looking at the isoperimetric inequality is from an optimization viewpoint. It states that Euclidean balls are optimal sets in terms of minimizing the $n - 1$ hypersurface volume, with a constraint on their n volume. The second (equivalent) inequality is from an inequality angle. It allows us to control the n volume of a set in terms of its boundary's $n - 1$ volume. For more information about this fundamental inequality, we refer the reader to [27].

Among all equal volume sets on the plane, the isoperimetric inequality says that the disc has the least perimeter. This statement compares all domains to balls. The waist inequality is its close cousin with perhaps stronger topological flavor. This is a statement about all continuous maps $f : B_R^n \to \mathbb{R}^m$: we can find $f^{-1}(y)$ such that $Vol_{n-m}(f^{-1}(y)) \geq Vol_{n-m}B_R^{n-m}$. This compares all continuous maps's volume-maximal fiber to balls. See Fig. 3 for an illustration in 3D.

B Precision, Recall, One-To-One, and Continuity

We extend the definitions of continuity and injectivity to allow exceptions on a measure zero set. For a dimensionality reduction map $f : \mathbb{R}^n \to \mathbb{R}^m$, we say it is essentially one-to-one if its 'injectivity' is essentially no more than the reduction part. The manifold setting $f : \mathcal{M}^n \to \mathbb{R}^m$ is handled naturally by using coordinates and parametrization by open sets in $\mathbb{R}^n$, as in classical differential topology and differential geometry.

---

[7] It is natural to consider $n - m$ dimensional volume for $f^{-1}(y)$, due to Sard's theorem [13] and implicit function theorem: since almost every $y \in f(B^n)$ is a regular value, $f^{-1}(y)$ is an $n - m$ dimensional submanifold, for such regular y. For an arbitrary continuous function, $Vol_{n-m} = \mathcal{M}_*^{n-m}$ is the lower Minkowski content, where the Waist Inequality is established [2]. For $n - m$ rectifiable sets, $Vol_{n-m} = \mathcal{M}_*^{n-m} = \mathcal{H}^{n-m}$.
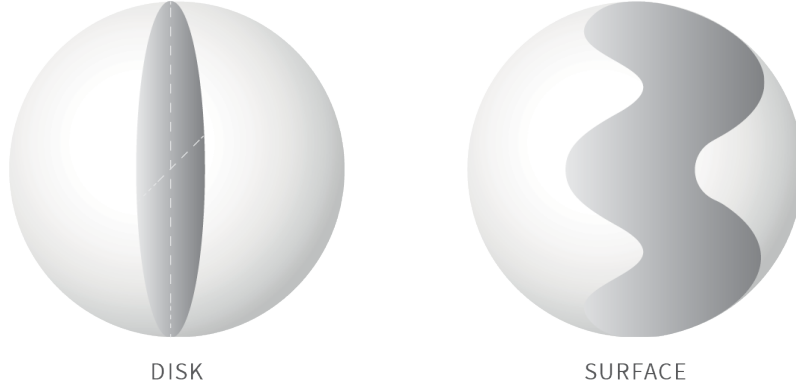
DISK   SURFACE

Figure 3: The above pictorial illustration compares $f^{-1}(y)$ - the pancake surface living in a 3-ball to a disc living in the 3-ball. We see that $f^{-1}(y)$ has bigger or equal area than the corresponding disc.

Definition 2 (Essential Continuity). *f is essentially continuous at x, if for any $\epsilon > 0$, there exists $r > 0$, such that for all the neighbourhood $U \ni x$ satisfying $diam(U) \leq r$,*

$$Vol_n \left(\{u \in U \, : \, |f(u) - f(x)| > \epsilon\} \right) = 0.$$

We say f is essentially continuous on a set W if f is essentially continuous at every $w \in W$.

Definition 3 (Essential Injectivity). *f is essentially one-to-one or essentially injective at x, if for $y = f(x) \in \mathbb{R}^m$, $Vol_{n-m}\left(f^{-1}(y)\right) = 0$[8]. f is essentially one-to-one on a set W if f is essentially one-to-one at every $w \in W$.*

Note that the definition of essential continuity (one-to-one, respectively) strictly generalizes the definition of continuity (one-to-one, respectively). In other words, every continuous function is essentially continuous, and there exists discontinuous functions that are essentially continuous. The following lemma shows that if f is essentially continuous on an open set W, then f is continuous on W.

Lemma 1 (Essential continuity in a neighborhood). *Essential continuity in a neighborhood and continuity in a neighborhood are equivalent.*

Proof. It is sufficient to prove that if f is essentially continuous on an open set W, then f is continuous on W. Assume that f is not continuous on W, i.e., there exists $\eta > 0$, $w \in W$ and a sequence $\{w_1, \ldots, w_n, \ldots\}$ such that $\lim_{n \to \infty} w_n = w$, but $|f(w_n) - f(w)| \geq \eta$. Since f is essentially continuous on W, there exists a neighbourhood of w, $U \subset W$, such that $Vol_n(E_U) = 0$, where $E_U = \{u \in U \, : \, |f(u) - f(w)| > \eta/3\}$. Note that for large enough M, $w_M \in E_U$. Moreover, since f is also essentially continuous at w_M, for a small neighbourhood V of w_M, $Vol_n(\{v \in V \, : \, |f(v) - f(w_M)| \leq \eta/3\}) = Vol_n(V) > 0$. However, note that this positive measure set $\{v \in V \, : \, |f(v) - f(w_M)| \leq \eta/3\}$ is a subset of E_U by the definition of E_U, contradicting $Vol_n(E_U) = 0$. $\qquad\square$

We next prove the equivalence between perfect recall and essential continuity.

Proposition 3. *For any map $f : \mathcal{M} \subset \mathbb{R}^N \to \mathbb{R}^m$, f achieves perfect recall in an open set W, if and only if f is essentially continuous on W.*

Proof. **(Perfect Recall $\Rightarrow$ Essential Continuity)** For any $x \in W$, any $\epsilon > 0$, let $V = \{f(v) \in \mathbb{R}^m \, : \, |f(v) - f(x)| \leq \epsilon\}$. Since f achieves perfect recall at x, there exists $r > 0$, such that $Vol_n(f^{-1}(V) \cap B_r(x)) = Vol_n(B_r(x))$. Therefore, for any U such that $U \subset B_r(x)$,

$$Vol_n \left(\{u \in U \, : \, |f(u) - f(x)| > \epsilon\} \right) \leq Vol_n \left(\{u \in U \, : \, u \notin f^{-1}(V) \cap B_r(x)\} \right) = 0.$$

---

[8]If the dimension of $f^{-1}(y)$ is greater than $n - m$, we define its volume to be ∞

Thus f is essentially continuous at x.

(Essential Continuity $\Rightarrow$ Perfect Recall) By Lemma 1, f is continuous on W. For any $x \in W$, assume $f(x) = y$. For any $r_V > 0$, $f^{-1}(B_{r_V}(y))$ is an open set in $\mathcal{M}$. Therefore, there exists small enough r_U such that $B_{r_U}(x) \subset f^{-1}(V)$, thus $\text{Recall}^f(B_{r_U}(x), B_{r_V}(y)) = 1$. $\qquad\square$

Based on this proposition, we can further prove that if f is (essentially) continuous on W, then f has neither perfect precision nor essential injectivity property on W.

Proposition 4. *Let $f : \mathcal{M}^n \subset \mathbb{R}^N \to \mathbb{R}^m$, with $m < n$. If f is (essentially) continuous with approximate differential well defined on an open set W almost everywhere,* [9]*, then f possesses neither perfect precision nor essential injectivity on W.*

Proof. **(Continuous in neighborhood $\Rightarrow$ Not Essentially Injective)** We first prove that if f is continuous on $W \subset \mathbb{R}^n$, then f is not essentially one-to-one on W. To prove that f does not have perfect precision, it is sufficient to prove that the perfect precision of f implies f being essentially one-to-one. We handle the manifold case at the end of the proof, by coordination: $\phi : U \subset \mathcal{M}^n \to V \subset \mathbb{R}^n$, and parametrization $\phi^{-1} : V \subset \mathbb{R}^n \to U \subset \mathcal{M}^n$.

Assume f is essentially one-to-one on W, thus for any $y \in f(W) \subset \mathbb{R}^m$,

$$\text{Vol}_{n-m}(f^{-1}(y)) = \int_{f^{-1}(y)} d\text{Vol}_{n-m}(p) = 0.$$

Since $W \subset \mathbb{R}^n$ is open, there is an open ball $B_\tau^n \subset W$ such that we can consider the restriction of f onto B_τ^n. Now Theorem 7 guarantees the existence of $y_\tau \in f(B_\tau^n)$ such that

$$\text{Vol}_{n-m}(f^{-1}(y_\tau)) \geq \text{Vol}_{n-m}(B_\tau^n) > 0.$$

This contradiction completes the proof in the Euclidean case.

Now, for a map $f : W \subset \mathcal{M}^n \to \mathbb{R}^m$. We consider the restriction of f on $U \subset W$ where U is homeomorphic to $\mathbb{R}^n$. Then the composite map: $f \circ \phi^{-1} \to \mathbb{R}^m$ is again a map between Euclidean spaces. The argument above applies and we complete this part of the proof.

(Perfect Precision $\Rightarrow$ Essential One-to-one) Assume that f is not essentially one-to-one on W, thus f is not one-to-one on W. Therefore, there exist y, z_1, and z_2 such that $f(z_1) = f(z_2) = y$. Without loss of generality, assume $d(z_1, z_2) = 1$. Since f has perfect precision, picking $U = B_{0.4}^m(z_1)$, there exists $r_{V,1}$, such that $\text{Vol}_n\left(f^{-1}(B_r^m(y)) \cap B_{0.4}^m(z_1)\right) = \text{Vol}_n\left(f^{-1}(B_r^m(y))\right)$ for $r \leq r_{V,1}$. Similarly, there exists $r_{V,2}$, such that $\text{Vol}_n\left(f^{-1}(B_r^m(y)) \cap B_{0.4}^m(z_2)\right) = \text{Vol}_n\left(f^{-1}(B_r^m(y))\right)$ for $r \leq r_{V,2}$. Further note that $B_{0.4}^m(z_1) \cap B_{0.4}^m(z_2) = \emptyset$. For $r \leq \min\{r_{V,1}, r_{V,2}\}$, then

$$\text{Vol}_n(f^{-1}(B_r^m(y))) \geq \text{Vol}_n(f^{-1}(B_r^m(y)) \cap B_{0.4}^m(z_1)) + \text{Vol}_n(f^{-1}(B_r^m(y)) \cap B_{0.4}^m(z_2))$$
$$= 2 * \text{Vol}_n(f^{-1}(B_r^m(y))).$$

Therefore, $\text{Vol}_n\left(f^{-1}(B_r^m(y))\right) = 0$. Now since f is continuous, $f^{-1}(B_r^m(y))$ is an open set in $\mathcal{M}$, thus $\text{Vol}_n\left(f^{-1}(B_r^m(y))\right)$ cannot be 0, a contradiction. $\qquad\square$

Based on Propositions 3 and 4, the proof of Theorem 1 is straightforward.

Proof of Theorem 1. It is sufficient to prove that if f achieves perfection recall at W, then f cannot achieve perfect precision at W. Since f achieves perfect recall at W, by Proposition 3 f is continuous, thus by Proposition 4 f cannot achieve perfect precision at W. $\qquad\square$

C Proof of Theorem 2

We present the proof of Theorem 2 in this section. The following proposition develops a lower bound for the volume of the inverse image of f on a particular small open set.

---

[9]This is a weaker condition than Lipschitz, including functions of bounded variation. A Lipschitz function is differentiable almost everywhere.

Proposition 5. *If f is a continuous function with Lipschitz constant L, then for any $y \in \mathbb{R}^m$ and $\epsilon > 0$,*

$$Vol_n \left(f^{-1}(B_\epsilon^m(y)) \right) \geq Vol_n \left(f^{-1}(y) \oplus \frac{\epsilon}{L} \right).$$

Proof. Since f is Lipschitz, for any x such that $d(x, f^{-1}(y)) \leq \frac{\epsilon}{L}$, $|f(x) - f(y)| \leq \epsilon$. Thus

$$f^{-1}(y) \oplus \frac{\epsilon}{L} = \{x \in \mathcal{M} : d(x, f^{-1}(y)) \leq \frac{\epsilon}{L}\} \subset \{x \in \mathcal{M} : |f(x) - f(y)| \leq \frac{\epsilon}{L}\} = f^{-1}(B_\epsilon^m(y)).$$

Therefore,

$$\text{Vol}_n \left(f^{-1}(y \oplus \epsilon) \right) \geq \text{Vol}_n \left(f^{-1}(y) \oplus \frac{\epsilon}{L} \right).$$

$\square$

Proof of Theorem 2. By Theorem 7, there exists $y \in \mathbb{R}^m$ such that

$$\text{Vol}_n \left(f^{-1}(y) \oplus \epsilon \right) \geq \frac{1}{2\pi R} \text{Vol}_{n-m+1} \left(S_R^{n-m+1} \right) \text{Vol}_m \left(B_1^m \right) \epsilon^m \left(1 + o(1) \right).$$

For any $x \in f^{-1}(y)$, $r_U, r_V > 0$, recall that $\text{Precision}^f(U, V) = \frac{\text{Vol}_n(f^{-1}(V) \cap U)}{\text{Vol}_n(f^{-1}(V))} \leq \frac{\text{Vol}_n(U)}{\text{Vol}_n(f^{-1}(V))}$, thus a lower bound of $\text{Vol}_n(f^{-1}(V))$ leads to an upper bound for $\text{Precision}^f(U, V)$. Further note that

$$\begin{aligned}
\text{Vol}_n(f^{-1}(V)) &= \text{Vol}_n \left(f^{-1}(y \oplus r_V) \right) \\
&\geq \text{Vol}_n \left(f^{-1}(y) \oplus (r_V/L) \right) \\
&\geq \frac{1}{2\pi R} \text{Vol}_{n-m+1}(S^{n-m+1}) \text{Vol}_m(B_1^m) R^{n-m+1} p^m(r_V/L) \\
&= \frac{\pi^{(n-m)/2}}{\Gamma(\frac{n-m}{2} + 1)} \frac{\pi^{m/2}}{\Gamma(\frac{m}{2} + 1)} R^{n-m} p^m(r_V/L), \quad (7)
\end{aligned}$$

where the first inequality is due to Proposition 5, the second inequality is due to the Waist Inequality Equation (6), and $p^m(x) = x^m(1 + o(1))$. Combining the volume calculation on U,

$$\begin{aligned}
\text{Precision}^f(U, V) &\leq \frac{\frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} r_U^n}{\frac{\pi^{n-m/2}}{\Gamma(\frac{n-m}{2} + 1)} \frac{\pi^{m/2}}{\Gamma(\frac{m}{2} + 1)} R^{n-m} p^m(r_V/L)} \\
&\leq \frac{\Gamma(\frac{n-m}{2} + 1)\Gamma(\frac{m}{2} + 1)}{\Gamma(\frac{n}{2} + 1)} (\frac{r_U}{R})^{n-m} \frac{r_U^m}{p^m(r_V/L)}.
\end{aligned}$$

$\square$

Theorem 2 generalizes as long as there is a corresponding waist theorem for that space. And roughly the condition of having a waist theorem is that a space is 'truly' n dimensional. We therefore conjecture that Theorem 2 holds in various settings in machine learning where we are dealing with truly n dimensional data. In the rest of this section, we are going to prove analogues of Theorem 2 under the non-Euclidean norm.

We define the necessary concepts first. In the non-Eucldiean case, the generalized unit ball is a convex body.

Definition 4 (Generalized Unit Ball, e.g. Wang [35]). *Let $p_1, p_2, \ldots, p_n \geq 1$. A generalized unit n ball is defined as the following convex body:*

$$B_{p_1, p_2, \ldots, p_n} = \{(x_1, x_2, \ldots, x_n) : |x_1|^{p_1} + \ldots + |x_n|^{p_n} \leq 1\} \quad (8)$$

Theorem 9 (Volume of Generalized Ball, Wang [35]).

$$Vol_n B_{p_1, p_2, \ldots, p_n} = 2^n \frac{\Gamma(1 + 1/p_1) \ldots \Gamma(1 + 1/p_n)}{\Gamma(1 + 1/p_1 + \ldots + 1/p_n)} \quad (9)$$

Definition 5 (Log-Concave Measure). *A Borel measure μ on $\mathbb{R}^n$ is log-concave if for any compacts sets $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^n$, and for any $0 < \lambda < 1$:*

$$\mu(\lambda A \oplus (1-\lambda)B) \geq \mu(A)^{\lambda}\mu(B)^{1-\lambda} \tag{10}$$

Theorem 10 (Brunn-Minkowski Inequality). *Let Vol_n denote Lebesgue measure on $\mathbb{R}^n$. Let A and B be two nonempty compact subsets of $\mathbb{R}^n$. Then:*

$$[Vol_n(A \oplus B)]^{1/n} \geq [Vol_n(A)]^{1/n} + [Vol_n(B)]^{1/n} \tag{11}$$

The following lemma is well known in concentration of measure and convex geometry. We prove it here for completeness.

Lemma 2 (Lebesgue Measure on Convex Sets is Log-Concave). *Let Vol_n denote Lebesgue measure on $\mathbb{R}^n$. The (induced) restricted measure, Vol_n, by restricting Vol_n to any convex sets is log-concave.*

Proof. Plugging λA and $(1-\lambda)B$ to theorem 10, we have:

$$\mathrm{Vol}_n^{1/n}(\lambda A \oplus (1-\lambda)B) \geq \mathrm{Vol}_n^{1/n}(\lambda A) + \mathrm{Vol}_n^{1/n}((1-\lambda)B) \tag{12}$$

$$= \lambda \mathrm{Vol}_n^{1/n}(A) + (1-\lambda)\mathrm{Vol}_n^{1/n}(B) \tag{13}$$

$$\geq \mathrm{Vol}_n^{\lambda/n}(A)\mathrm{Vol}_n^{(1-\lambda)/n}(B) \tag{14}$$

where the first equality follows because the λ (or $1-\lambda$ respectively) is scaled be a factor or λ^n and taking nth root gives the equality, and the last inequality follows from the weighted arithmetic-geometric mean inequality. Raising to the nth power, we get:

$$\mathrm{Vol}_n(\lambda A \oplus (1-\lambda)B) \geq \mathrm{Vol}_n^{\lambda}(A)\mathrm{Vol}_n^{(1-\lambda)}(B) \tag{15}$$

To finish the proof, we note that for any A and B as nonempty compact subsets of a convex set $K \subset \mathbb{R}^n$ in the Euclidean space, the Lebesgue measures restricted on K, $\mathrm{Vol}_n(A)$ and $\mathrm{Vol}_n(B)$ can be written as Lebegues measures on A and B. Convexity of K ensures $\lambda A \oplus (1-\lambda)B$ is still in the set K. $\qquad\square$

To deduce an analogue of Theorem 2, we need the following waist inequality for log-concave measures.

Theorem 11 (Waists of Arbitrary Norms, Theorem 5.4 of Akopyan and Karasev [2]). *Suppose $K \subset \mathbb{R}^n$ is a convex body, μ a finite log-concave measure supported on K, and $f : K \longrightarrow \mathbb{R}^m$ is continuous. Then for any $\epsilon \in [0,1]$ there exists $y \in \mathbb{R}^m$ such that:*

$$\mu(f^{-1}(y) \oplus \epsilon K) \geq \epsilon^m \mu(K) \tag{16}$$

Proposition 6 (Precision on Arbitrarilly Normed Balls). *Let $m < n$. Let $f : B_{R;p_1,p_2,\dots,p_n} \longrightarrow \mathbb{R}^m$ be a L-Lipschtiz continuous map defined on a generalized n ball with radius R from Definition 4. Let r_U and r_V be radii of two generalized balls, with dimensions n and m respectively. Then there exists y depending on r_V such that:*

$$Prec^f(U,V) \leq (\frac{r_U}{R})^{n-m}(\frac{r_U}{r_V/L})^m \tag{17}$$

Proof. We would like to apply theorem 11. Since $B_{R;p_1,p_2,\dots,p_n}$ is a convex body, Lebesgue meaure Vol_n on $B_{R;p_1,p_2,\dots,p_n}$ is log-concave by Lemma 2. Then by Theorem 11, for r_V/L, there exists $y \in \mathbb{R}^m$ such that:

$$\mathrm{Vol}_n(f^{-1}(y) \oplus \frac{r_V}{L}K) \geq (\frac{r_V}{L})^m \mathrm{Vol}_n(K) \tag{18}$$

where $K = B_{R;p_1,p_2,\dots,p_n}$. Now by Proposition 5,

$$\mathrm{Vol}_n(f^{-1}(V)) = \mathrm{Vol}_n(f^{-1}(B_{r_V;p_1,p_2,\dots,p_n})) \geq (\frac{r_V}{L})^m \mathrm{Vol}_n(K) \tag{19}$$

16

Therefore:

$$Prec^f(U, V) \leq \frac{\text{Vol}_n(U)}{\text{Vol}_n(f^{-1}(V))} \tag{20}$$

$$\leq \frac{\text{Vol}_n(B_{r_U;p_1,p_2,\ldots,p_n})}{\text{Vol}_n(B_{R;p_1,p_2,\ldots,p_n})(\frac{r_V}{L})^m} \tag{21}$$

$$= \frac{2^n \frac{\Gamma(1+1/p_1)\ldots\Gamma(1+1/p_n)}{\Gamma(1+1/p_1+\ldots+1/p_n)} r_U^{n-m} r_U^m}{2^n \frac{\Gamma(1+1/p_1)\ldots\Gamma(1+1/p_n)}{\Gamma(1+1/p_1+\ldots+1/p_n)} R^{n-m}(\frac{r_V}{L})^m} \tag{22}$$

$$= (\frac{r_U}{R})^{n-m}(\frac{r_U}{r_V/L})^m \tag{23}$$

$$\square$$

D Proof of Theorem 3

The proof of Theorem 3 is based on the idea that the fibers of certain type of continuous DR maps are mostly 'large'. A map f has a large fiber at y if $f^{-1}(y)$'s volume is lower bounded by that of a linear map. This concept of 'large' fiber is actually an essential concept in the proof of the waist inequality. The intuition we try to capture is that fibers of f are considered big if their $n - m$ volumes are comparable to that of a surjective linear map.

The next two theorems show that for either of the following cases:

- $m = 1$; or
- $f : B_R^n \to \mathbb{R}^m$ be a k-layer neural network map with Lipschitz constant L, whose linear layers are surjective.

the fibers of f are mostly 'large'.

Theorem 12 (Average Waist Inequality for Balls, m = 1). *Let f be a continuous map from B_R^n to $\mathbb{R}$, and $\tau = Vol_{n+1}\left(Proj^{-1}(y) \oplus \epsilon\right)$ for an arbitrary $y \in Proj(S_R^{n+1})$, then for all $\epsilon > 0$*

$$Vol_n\left(\left\{z \in B_R^n : Vol_n\left(f^{-1}(f(z)) \oplus \epsilon\right) \geq \frac{1}{2\pi R}\tau\right\}\right)$$

$$\geq \frac{1}{2\pi R} Vol_{n+1}\left(\left\{x \in S_R^{n+1} : Vol_{n+1}\left(Proj^{-1}(Proj(x)) \oplus \epsilon\right) \geq \tau\right\}\right).$$

Proposition 7. *Let f be a k layer neural network with nonlinear activations (ReLu, LeakyReLu, tanh, etc.) from B_R^n to $(0,1)^m$ and Proj be an arbitrary linear projection on B_R^n. Then for any τ the following inequality holds,*

$$Vol_n\left(\left\{x \in B_R^n : Vol_n\left(f^{-1}(f(x)) \oplus \epsilon\right) \geq \tau\right\}\right)$$

$$\geq Vol_n\left(\left\{x \in B_R^n : Vol_n\left(Proj^{-1}(Proj(x)) \oplus \epsilon\right) \geq \tau\right\}\right).$$

The proof of Theorem 12 is postponed to Appendix D.1, while the proof of Proposition 7 is postponed to Appendix D.2. We are now ready to derive a bound on DR maps' **average-case performance** over the domain based on Theorem 12 and Proposition 7.

Proof of Theorem 3. We only present the proof when $f : B_R^n \to \mathbb{R}^m$ is a k-layer neural network map with Lipschitz constant L by Proposition 7. The other case can be proved similarly by Theorem 12.

Given any $y \in \text{Proj}(B_R^n)$, pick $\tau = \text{Vol}_n\left(\text{Proj}^{-1}(y) \oplus \epsilon\right)$. By Proposition 7 for all $\epsilon > 0$,

$$\text{Vol}_n\left(\left\{x \in B_R^n : \text{Vol}_n\left(f^{-1}(f(x)) \oplus \epsilon\right) \geq \text{Vol}_n\left(\text{Proj}^{-1}(y) \oplus \epsilon\right)\right\}\right) \tag{24}$$

$$\geq \text{Vol}_n\left(\left\{x \in B_R^n : \text{Vol}_n\left(\text{Proj}^{-1}(\text{Proj}(x)) \oplus \epsilon\right) \geq \text{Vol}_n\left(\text{Proj}^{-1}(y) \oplus \epsilon\right)\right\}\right).$$

Since Proj is a linear map, we have

$$\text{Vol}_n\left(\left\{x \in B_R^n : \text{Vol}_n\left(\text{Proj}^{-1}(\text{Proj}(x)) \oplus \epsilon\right) \geq \text{Vol}_n\left(\text{Proj}^{-1}(y) \oplus \epsilon\right)\right\}\right)$$

$$= \text{Vol}_n\left(\left\{x \in B_R^n : \text{Vol}_{n-m}\left(\text{Proj}^{-1}(\text{Proj}(x))\right) \geq \text{Vol}_{n-m}\left(\text{Proj}^{-1}(y)\right)\right\}\right).$$

Further note that $\text{Proj}^{-1}(y)$ is an $n-m$ ball with radius $r(y) = \sqrt{R^2 - \|y\|^2}$. Thus,

$$\text{Vol}_n\left(\left\{x \in B_R^n : \text{Vol}_{n-m}\left(\text{Proj}^{-1}\left(\text{Proj}(x)\right)\right) \geq \text{Vol}_{n-m}\left(\text{Proj}^{-1}(y)\right)\right\}\right)$$
$$= \int_{B_{\|y\|}^m} \text{Vol}_{n-m}(\text{Proj}^{-1}(t))\mathrm{d}t.$$

Therefore,

$$\text{Vol}_n\left(\left\{x \in B_R^n : \text{Vol}_n\left(f^{-1}\left(f(x)\right) \oplus \epsilon\right) \geq \text{Vol}_n\left(\text{Proj}^{-1}(y) \oplus \epsilon\right)\right\}\right) \geq \int_{B_{\|y\|}^m} \text{Vol}_{n-m}(\text{Proj}^{-1}(t))\mathrm{d}t.$$

$$(25)$$

Lastly, pick y such that $\|y\| = \sqrt{R^2 - r_U^2 - \delta^2}$, so $\text{Proj}^{-1}(y)$ has radius $\sqrt{r_U^2 + \delta^2}$. Let $\mathcal{E}$ denote the event $\text{Vol}_n\left(f^{-1}\left(f(x)\right) \oplus \epsilon\right) \geq \text{Vol}_n\left(B_{\sqrt{r_U^2+\delta^2}}^{n-m} \oplus \epsilon\right)$, thus

$$\mathbb{P}(\mathcal{E}) \geq \frac{\int_{B_{\sqrt{R^2-u^2-\delta^2}}^m} \text{Vol}_{n-m}\text{Proj}^{-1}(t)\mathrm{d}t}{\text{Vol}_n(B_R^n)}.$$

The remaining proof is almost identical to the proof of Theorem 2. Under the event $\mathcal{E}$,

$$\text{Vol}_n(f^{-1}(V)) = \text{Vol}_n\left(f^{-1}(y \oplus r_V)\right)$$
$$\geq \text{Vol}_n\left(f^{-1}(y) \oplus (r_V/L)\right)$$
$$\geq \text{Vol}_{n-m}(B_1^{n-m})\text{Vol}_m(B_1^m)(\sqrt{r_U^2 + \delta^2})^{n-m}(r_V/L)^m$$
$$= \frac{\pi^{(n-m)/2}}{\Gamma(\frac{n-m}{2} + 1)}\frac{\pi^{m/2}}{\Gamma(\frac{m}{2} + 1)}\sqrt{r_U^2 + \delta^2}^{n-m}(r_V/L)^m,$$

$$(26)$$

where the first inequality is due to Proposition 5, the second inequality is due to the event $\mathcal{E}$. Combining the volume calculation on U,

$$\text{Precision}^f(U, V) \leq \frac{\frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}r_U^n}{\frac{\pi^{n-m/2}}{\Gamma(\frac{n-m}{2}+1)}\frac{\pi^{m/2}}{\Gamma(\frac{m}{2}+1)}\sqrt{r_U^2 + \delta^2}^{n-m}(r_V/L)^m}$$
$$\leq \frac{\Gamma(\frac{n-m}{2} + 1)\Gamma(\frac{m}{2} + 1)}{\Gamma(\frac{n}{2} + 1)}(\frac{r_U}{\sqrt{r_U^2 + \delta^2}})^{n-m}\frac{r_U^m}{(r_V/L)^m}.$$

$\square$

D.1 Proof of Theorem 12

The proof uses the following average waist inequality for spheres. Let $P : S_R^{n+1} \longrightarrow B_R^n$ be the orthogonal projection, σ_R and ν_R denote the corresponding Hausdorff measures on S_R^{n+1} and B_R^n. Further, let $\text{Proj} : S_R^{n+1} \to \mathbb{R}$ be the restriction to S_R^{n+1} of a surjective linear map $\widehat{\text{Proj}} : \mathbb{R}^{n+2} \to \mathbb{R}$.

Theorem 13 (Average Waist Inequality for Spheres [3]). *Let f be a continuous map from S_R^{n+1} to $\mathbb{R}$, then for all $y \in Proj(S_R^{n+1})$, we have:*

$$Vol_{n+1}\{x \in S_R^{n+1} : Vol_{n+1}(f^{-1}(f(x)) \oplus \epsilon) \geq Vol_{n+1}(Proj^{-1}(y) \oplus \epsilon)\}$$
$$\geq$$
$$Vol_{n+1}\{x \in S_R^{n+1} : Vol_{n+1}(Proj^{-1}(Proj(x)) \oplus \epsilon) \geq Vol_{n+1}(Proj^{-1}(y) \oplus \epsilon)\},$$

where

$$Vol_{n+1}\left(Proj^{-1}(y) \oplus \epsilon\right) = 2\pi Vol_n\left(S_{R_{Proj^{-1}(y)}}^n\right) Vol_1\left(B_1^1\right)\left(p^1(\epsilon)\right),$$

$p^1(\epsilon)$ is $\epsilon\,(1 + o(1))$, i.e. $\lim_{\epsilon \to 0} \frac{p^1(\epsilon)}{\epsilon} = 1$, and $f^{-1}(y) \oplus \epsilon$ denotes the set of points $x \in S_R^{n+1}$ such that $d(x, f^{-1}(y)) < \epsilon$, S_R^n is the n-dimensional sphere of radius R, and $S_{R_{Proj^{-1}(y)}}^n$ is the sphere with radius $R_{Proj^{-1}(y)}$ depending on where y is taken in $f(S_R^{n+1})$, i.e. $R_{Proj^{-1}(y)}^2 = R^2 - y^2$.

18

We are going to adapt the proof technique of theorem 1 from [1], by replacing the existential waist inequality (7) with its average version - theorem 13. We need the following lemma:

Lemma 3 (Orthogonal Projection e.g. Akopyan and Karasev [1]). *Let $P : S_R^{n+1} \longrightarrow B_R^n$ be the orthogonal projection. Then P is 1 - Lipschitz and $P_\# \sigma_R = 2\pi R \nu_R$. In other words, P sends the uniform Hausdorff measure σ in S_R^{n+1} to the uniform Lebesgue measure ν_n in B_R^n up to constant $2\pi R$.*

Proof of Theorem 12. Given a map $f : B_R^n \longrightarrow \mathbb{R}$, consider $\hat{f} = f \circ P : S_R^{n+1} \to \mathbb{R}$, where P is the orthogonal projection. By Lemma 3, P is 1-Lipschitz, thus for any $y \in \mathbb{R}$,

$$P^{-1}\left(f^{-1}(y)\right) \oplus \epsilon \subset P^{-1}\left(f^{-1}(y) \oplus \epsilon\right) \Rightarrow \mathrm{Vol}_{n+1}\left(\hat{f}^{-1}(y) \oplus \epsilon\right) \leq \mathrm{Vol}_{n+1}\left(P^{-1}\left(f^{-1}(y) \oplus \epsilon\right)\right). \tag{27}$$

Further, since $P_\# \sigma_R = 2\pi R \nu_R$,

$$\mathrm{Vol}_{n+1}\left(P^{-1}\left(f^{-1}(y) \oplus \epsilon\right)\right) = 2\pi R \mathrm{Vol}_n\left(f^{-1}(y) \oplus \epsilon\right). \tag{28}$$

Combining Equations (27) and (28), for $\tau \in \mathbb{R}$,

$$\left\{x \in S_R^{n+1} \ : \ \hat{f}(x) = y, \mathrm{Vol}_{n+1}\left(\hat{f}^{-1}(y) \oplus \epsilon\right) \geq \tau\right\}$$
$$\subset \left\{x \in S_R^{n+1} \ : \ \hat{f}(x) = y, \mathrm{Vol}_n\left(f^{-1}(y) \oplus \epsilon\right) \geq \frac{\tau}{2\pi R}\right\}. \tag{29}$$

Similarly, by $P_\# \sigma_R = 2\pi R \nu_R$,

$$\mathrm{Vol}_{n+1}\left(\left\{x \in S_R^{n+1} \ : \ \hat{f}(x) = y, \mathrm{Vol}_n\left(f^{-1}(y) \oplus \epsilon\right) \geq \frac{\tau}{2\pi R}\right\}\right)$$
$$= 2\pi R \mathrm{Vol}_n\left(\left\{z \in B_R^n \ : \ f(z) = y, \mathrm{Vol}_n\left(f^{-1}(y) \oplus \epsilon\right) \geq \frac{\tau}{2\pi R}\right\}\right). \tag{30}$$

Thus by combining Equations (29) and (30), we have

$$\mathrm{Vol}_{n+1}\left(\left\{x \in S_R^{n+1} \ : \ \hat{f}(x) = y, \mathrm{Vol}_{n+1}\left(\hat{f}^{-1}(y) \oplus \epsilon\right) \geq \tau\right\}\right)$$
$$\leq 2\pi R \mathrm{Vol}_n\left(\left\{z \in B_R^n \ : \ f(z) = y, \mathrm{Vol}_n\left(f^{-1}(y) \oplus \epsilon\right) \geq \frac{\tau}{2\pi R}\right\}\right)$$

Finally, note that $\hat{f}$ meets the condition in theorem 13. Thus for all $y \in \mathrm{Proj}(S_R^{n+1})$:

$$\mathrm{Vol}_{n+1}\left(\left\{x \in S_R^{n+1} : \mathrm{Vol}_{n+1}\left(\mathrm{Proj}^{-1}\left(\mathrm{Proj}(x)\right) \oplus \epsilon\right) \geq \mathrm{Vol}_{n+1}\left(\mathrm{Proj}^{-1}(y) \oplus \epsilon\right)\right\}\right)$$
$$\leq \mathrm{Vol}_{n+1}\left(\left\{x \in S_R^{n+1} : \mathrm{Vol}_{n+1}\left(\hat{f}^{-1}\left(\hat{f}(x)\right) \oplus \epsilon\right) \geq \mathrm{Vol}_{n+1}\left(\mathrm{Proj}^{-1}(y) \oplus \epsilon\right)\right\}\right)$$
$$\leq 2\pi R \mathrm{Vol}_n\left(\left\{z \in B_R^n \ : \ \mathrm{Vol}_n\left(f^{-1}\left(f(z)\right) \oplus \epsilon\right) \geq \frac{1}{2\pi R}\mathrm{Vol}_{n+1}\left(\mathrm{Proj}^{-1}(y) \oplus \epsilon\right)\right\}\right).$$

$\square$

D.2 Proof of Proposition 7

We first prove that Proposition 7 holds for any surjective linear map.

Proposition 8. *Let f be any surjective linear map (PCA, linear neural networks) from B_R^n to $\mathbb{R}^m$, and Proj be an arbitrary surjective linear projection from B_R^n to $\mathbb{R}^m$. Then for any τ the following inequality holds,*

$$Vol_n\left(\left\{x \in B_R^n \ : \ Vol_n\left(f^{-1}\left(f(x)\right) \oplus \epsilon\right) \geq \tau\right\}\right)$$
$$\geq Vol_n\left(\left\{x \in B_R^n : Vol_n\left(Proj^{-1}\left(Proj(x)\right) \oplus \epsilon\right) \geq \tau\right\}\right).$$

Proof. By the singular value decomposition, any linear dimension reduction map f can be decomposed as a composition or unitary operators (U_m and V_n), signed dialation of full rank (Σ), and

19

projection operator of rank m ($\widehat{\mathrm{Proj}}$), where $\widehat{\mathrm{Proj}}$ linearly projects from $\mathbb{R}^n$ to $\mathbb{R}^m$ (or more commonly $\Sigma \circ \widehat{\mathrm{Proj}}$ is called rectangular diagonal matrix map): $f = U_m \circ \Sigma \circ \widehat{\mathrm{Proj}} \circ V_n^*$. The set

$$
\begin{aligned}
&\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left(f^{-1}\left(f(x)\right) \oplus \epsilon\right) \geq \tau\right\}\\
=&\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left((U_m \circ \Sigma \circ \widehat{\mathrm{Proj}} \circ V_n^*)^{-1}\left(U_m \circ \Sigma \circ \widehat{\mathrm{Proj}} \circ V_n^*(x)\right) \oplus \epsilon\right) \geq \tau\right\}\\
=&\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left((V_n^*)^{-1} \circ \widehat{\mathrm{Proj}}^{-1} \circ \Sigma^{-1} \circ U_m^{-1} \circ U_m \circ \Sigma \left(\widehat{\mathrm{Proj}} \circ V_n^*(x)\right) \oplus \epsilon\right) \geq \tau\right\}\\
=&\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left((V_n^*)^{-1} \circ \widehat{\mathrm{Proj}}^{-1} \circ \left(\widehat{\mathrm{Proj}} \circ V_n^*(x)\right) \oplus \epsilon\right) \geq \tau\right\}\\
=&\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left(V_n \circ \widehat{\mathrm{Proj}}^{-1} \circ \left(\widehat{\mathrm{Proj}} \circ V_n^*(x)\right) \oplus \epsilon\right) \geq \tau\right\}\\
=&\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left(V_n \circ \widehat{\mathrm{Proj}}^{-1} \circ \left(\widehat{\mathrm{Proj}}(x)\right) \oplus \epsilon\right) \geq \tau\right\}\\
=&\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left(\widehat{\mathrm{Proj}}^{-1} \circ \widehat{\mathrm{Proj}}(x) \oplus \epsilon\right) \geq \tau\right\},
\end{aligned}
$$

where the last two equalities follow because unitary operator V_n^* and V_n don't affect volumes because they are linear isometries. We note this shows the distribution of fiber volume is the same for any surjective linear map. Finally, note that by symmetry,

$$
\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left(\widehat{\mathrm{Proj}}^{-1} \circ \widehat{\mathrm{Proj}}(x) \oplus \epsilon\right) \geq \tau\right\} = \left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left(\mathrm{Proj}^{-1} \circ \mathrm{Proj}(x) \oplus \epsilon\right) \geq \tau\right\}.
$$

$\square$

Lemma 4 (Monotonicity of Fiber Volume under Compositions). *Let $f : B_R^n \longrightarrow X$ and $g : X \longrightarrow \mathbb{R}^m$ be any maps for some set X. Then for any τ we have the following inequality:*

$$
\begin{aligned}
&\mathit{Vol}_n\left(\left\{x \in B_R^n \;:\; \mathit{Vol}_n\left((f \circ g)^{-1}\left(f \circ g(x)\right) \oplus \epsilon\right) \geq \tau\right\}\right)\\
\geq&\mathit{Vol}_n\left(\left\{x \in B_R^n : \mathit{Vol}_n\left(f^{-1}\left(f(x)\right) \oplus \epsilon\right) \geq \tau\right\}\right).
\end{aligned}
$$

Proof. Consider: $a \in \left\{x \in B_R^n : \mathrm{Vol}_n\left(f^{-1}\left(f(x)\right) \oplus \epsilon\right) \geq \tau\right\}$ and we let $b = f(a)$. We obviously have $b \in g^{-1} \circ g(b)$. Therefore $a \in f^{-1}(b) \subset f^{-1} \circ g^{-1} \circ g(f(a))$. Thus,

$$
\left\{x \in B_R^n : \mathrm{Vol}_n\left(f^{-1}\left(f(x)\right) \oplus \epsilon\right) \geq \tau\right\} \subset \quad \left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left((f \circ g)^{-1}\left(f \circ g(x)\right) \oplus \epsilon\right) \geq \tau\right\}.
$$

$\square$

Proof of Proposition 7. We proceed by induction on k. When $k = 1$, it is given by lemma 4, by noting a one layer net is a composition of any activation with a surjective linear map, L_1. Assume this is true for a k layer neural net, f_k, with k layers such that $k \geq 1$. So we have:

$$
\begin{aligned}
&\mathrm{Vol}_n\left(\left\{x \in B_R^n \;:\; \mathrm{Vol}_n\left(f_k^{-1}\left(f_k(x)\right) \oplus \epsilon\right) \geq \tau\right\}\right)\\
\geq&\mathrm{Vol}_n\left(\left\{x \in B_R^n : \mathrm{Vol}_n\left(\mathrm{Proj}^{-1}\left(\mathrm{Proj}(x)\right) \oplus \epsilon\right) \geq \tau\right\}\right).
\end{aligned}
$$

We need to check a neural net f_{k+1} with $k + 1$ layers: $f_{k+1} = \tanh \circ L_{k+1} \circ f_k$. But this is again a composition between functions and we can apply Lemma 4. This completes the proof. $\square$

In light of Proposition 8, we can characterize $\mathrm{Proj}_1^{-1}(t)$ and $\mathrm{Proj}_2^{-1}(t)$ explicitly. Since the bound holds for any surjective linear map, we can choose in particular $\mathrm{Proj}_1^{-1}(t)$ and $\mathrm{Proj}_2^{-1}(t)$ to be the coordinate projection from $\mathbb{R}^n$ to $\mathbb{R}^m$ (with all eigenvalues equal to 1). Then $t = (t_1, \cdots, t_m) \in B_R^m$, $\mathrm{Proj}_1^{-1}(t) = S_{\Re_1}^{n-m+1}$ and $\mathrm{Proj}_2^{-1}(t) = B_{\Re_2}^{n-m}$, where $\Re_1 = \Re_2 = \sqrt{R^2 - \sum_{i=1}^m t_i^2}$.

E Proofs for Section 3

This section is devoted to the proofs for Section 3. We first present the proof of Theorem 4.

Proof of Theorem 4. By Equation (7),

$$\text{Vol}_n(f^{-1}(V)) \geq \frac{\pi^{n/2}}{\Gamma(\frac{n-m}{2}+1)\Gamma(\frac{m}{2}+1)} R^{n-m} p^m (r_V/C).$$

Let $B_{r^\#}$ be the ball with the same volume as $\text{Vol}_n(f^{-1}(V))$ and a common center with U. Thus

$$r^\# \geq r = \left(\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n-m}{2}+1)\Gamma(\frac{m}{2}+1)}\right)^{\frac{1}{n}} R^{\frac{n-m}{n}} (p^m (r_V/C))^{\frac{1}{n}}. \tag{31}$$

By Theorem 5,

$$W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)}) \geq W_2^2(\mathbb{P}_U, \mathbb{P}_{B_{r^\#}}) = \int_{B_r(u)} |x - T(x)|^2 \, d\mathbb{P}_{B_{r^\#}}(x),$$

thus it is sufficient to lower bound the last term. Under the condition that $\text{Vol}_n(f^{-1}(V)) \geq \text{Vol}_n(U)$,

$$\int_{B_{r^\#}} |x - T(x)|^2 \, d\mathbb{P}_{B_{r^\#}}(x) = \int_{B_{r^\#}} |x - \frac{r_U}{r^\#} x|^2 \, d\mathbb{P}_{B_{r^\#}}(x)$$

$$= \left(1 - \frac{r_U}{r^\#}\right)^2 \int_{B_{r^\#}} |x|^2 \, d\mathbb{P}_{B_{r^\#}}(x).$$

Further,

$$\int_{B_{r^\#}} |x|^2 \, d\mathbb{P}_{B_{r^\#}}(x) = \int_0^{r^\#} r^2 \frac{1}{\text{Vol}_n(f^{-1}(V))} dS^{n-1}(r) dr$$

$$= \frac{1}{\text{Vol}_n(f^{-1}(V))} \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})} \int_0^{r^\#} r^{n+1} dr$$

$$= \frac{n}{n+2} (r^\#)^2.$$

Therefore,

$$W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)}) \geq \left(1 - \frac{r_U}{r^\#}\right)^2 \frac{n}{n+2} (r^\#)^2 = \frac{n}{n+2} (r^\# - r_U)^2.$$

Note that the above lower bound is monotonically increasing with respect to $r^\#$ for $r^\# > r_U$. Therefore from Equation (31), when $r > r_U$, replacing $r^\#$ by r gives a lower bound for $W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)})$.

Further, note that as $n \to \infty$, $r \to R$, we have:

$$W_2^2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)}) = \Omega\left((R - r_U)^2\right).$$

$\square$

The rest of this section is to prove Theorem 5. The key step is to show the following lemma.

Lemma 5 (Reduction to Optimal Partial Transport). *Given $f(x) = 1/\mathcal{V} \leq 1/Vol(B_r)$, the optimal distribution f_M for the optimal transport problem*

$$\min_{\mathbb{P} : \mathbb{P} \text{ is dominated by } f} W_2(\mathbb{P}, \mathbb{P}_{B_r}) \tag{32}$$

is the uniform distribution over $B_{r^\#}$ where $r^\#$ is the radius such that $Vol(B_{r^\#}) = \mathcal{V}$.

By Lemma 5, let $f(x) = 1/\mathcal{V}$, the optimal solution for the problem

$$\inf_{W : \text{Vol}_n(W) = \mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = W_2(\mathbb{P}_U, \mathbb{P}_{B_r})$$

is the same as support of the optimizer of Equation (32), thus proving the first statement of Theorem 5.

The proof of Lemma 5 is based on the uniqueness of the optimal transport map for the optimal partial transport problem [9, 11]. We summarize the statements in [11][10] as a theorem here for completeness.

---

[10]The Brenier theorem is not stated in the paper, but it holds under standard derivation.

Theorem 14 (Figalli [11]). *Let $f, g \in L^1(B_R^n)$ be two nonnegative functions, and denote by $\Xi_\le(f, g)$ the set of nonnegative finite Borel measures on $B_R^n \times B_R^n$ whose first and second marginals are dominated by f and g respectively, i.e. $\xi(A \times B_R^n) \le \int_A f(x) dx$ and $\xi(B_R^n \times A) \le \int_A g(y) dy$, for all Borel $A \subset B_R^n$. Denote $\mathcal{M}(\xi) := \int_{B_R^n \times B_R^n} d\xi$ and fix $M \in [\|\min(f(x), g(x))\|_{L_1}, \min(\|f\|_{L_1}, \|g\|_{L_1})]$. Then there exists a unique optimizer ξ_M[11] to the following optimal partial transport problem:*

$$\inf_{\xi \in \Xi_\le(f,g); \mathcal{M}(\xi)=M} C(\xi) = \inf_{\xi \in \Xi_\le(f,g); \mathcal{M}(\xi)=M} \int_{B_R^n \times B_R^n} |x-y|^2 d\xi(x, y)$$

Moreover, there exist Borel sets $A_1, A_2 \subset B_R^n$ such that ξ_M has left and right marginals whose densities $f_M = 1_{A_1} f$ and $g_M = 1_{A_2} g$ are given by the restrictions of f and g to A_1 and A_2 respectively, where 1_A denotes characteristic function on the set A.

Finally, there exists a unique optimal transport map T[12], such that

$$\min_{\xi \in \Xi_\le(f,g); \mathcal{M}(\xi)=M} C(\xi) = \int_{B_R^n} |T(x) - x|^2 df_M(x),$$

where f_M is the marginal of ξ_M over the first B_R^n.

We will prove Lemma 5 in two different ways. The first is based on calculus and reducing the problem to one dimensional optimal transport. The second one utilizes the extreme points property that characterizes the densities $f_M = 1_{A_1} f$ and $g_M = 1_{A_2} g$ (Proposition 3.2 and Theorem 3.3 in [17]). [13]

Proof of Lemma 5, first approach. Let $\mathcal{V} = \mathrm{Vol}(W)$, define $f(x) = 1/\mathcal{V}$ be a constant function on B_R^n and $g(x) = \frac{1}{\mathrm{Vol}(B_r)}$ if $x \in B_r$ and 0 otherwise. Also, let $M = 1$. solving the problem

$$\min_{\mathbb{P} : \mathbb{P} \text{ is dominated by } f} W_2(\mathbb{P}, \mathbb{P}_{B_r}) \tag{33}$$

is equivalent to solving the following optimal partial transport problem

$$\inf_{\xi \in \Xi_\le(f,g); \mathcal{M}(\xi)=1} C(\xi) = \inf_{\xi \in \Xi_\le(f,g); \mathcal{M}(\xi)=1} \int_{B_R^n \times B_R^n} |x-y|^2 d\xi(x, y). \tag{34}$$

In particular, since $\mathrm{Vol}(B_R^n) \ge \mathcal{V} > \mathrm{Vol}(B_r)$, it is straightforward to see that $\|\min(f(x), g(x))\|_{L_1} = \mathrm{Vol}(B_r)/\mathcal{V} < 1$, and $\min(\|f(x)\|_{L_1}, \|g(x)\|_{L_1}) \ge 1$. By Theorem 14, the optimization problem $\inf_{\xi \in \Xi_\le(f,g); \mathcal{M}(\xi)=1} C(\xi)$ has a unique solution ξ^*. Now given ξ^*, the optimal solution $\mathbb{P}^*$ of Equation (33) and $\mathbb{P}_{B_r}$ are the first and the second marginals of ξ^*. Thus it is sufficient to prove that the first marginal of ξ^* is a uniform distribution.

Let f_M be the first marginal of ξ^* and $g_M = g$ be the second marginal. We first show that f_M is rotationally invariant. To see that, for any rotation map $\mathcal{R}$, note that $\mathcal{R}(B_R^n) = B_R^n$, $\mathcal{R}(B_r) = B_r$, $f \circ \mathcal{R} = f$, and $g \circ \mathcal{R} = g$. Therefore, $f_M \circ \mathcal{R}$ is the unique optimal solution for the optimization problem

$$\inf_{\xi \in \Xi_\le(f \circ \mathcal{R}, g \circ \mathcal{R}); \mathcal{M}(\xi)=1} \int_{\mathcal{R}(B_R^n) \times \mathcal{R}(B_R^n)} |x-y|^2 d\xi(x, y) = \inf_{\xi \in \Xi_\le(f,g); \mathcal{M}(\xi)=1} \int_{B_R^n \times B_R^n} |x-y|^2 d\xi(x, y).$$

Thus, $f_M \circ \mathcal{R} = f_M$, i.e. f_M is rotationally invariant, up to a measure zero set. For a density function to be rotationally invariant, it is straightforward that its support S is also rotationally invariant, thus is a union of $(n-1)$ spheres. Similarly, one can also prove that T is equivariant under rotations.

We next prove that f_M is a uniform distribution. Note that g_M is a uniform distribution over B_r. Define $\hat{G}(t)$ to be the the cumulative distribution $\hat{g}$ for g_M in the polar coordinate marginalized on the sphere, i.e.,

$$\hat{G}(t) = \int_0^t \frac{1}{\mathrm{Vol}_n B_r^n} \mathrm{Vol}_{n-1}(S_u^{n-1}) du,$$

---

[11]up to a measure zero set

[12]up to a measure zero set

[13]Such property can also be deduced from earlier work, e.g. Theorem 4.3 and Corollary 2.11 from [9]. But [17] is perhaps more direct and accessible.

for every $0 \le t \le r$, and $G(t) = 1$ for $t > r$. Similarly, since f_M is also rotationally invariant, we can also define its cumulative distribution in the polar coordinate marginalized on the sphere. Note that $\mathrm{d}\mu_{f_M} = f_M(x)\mathrm{d}S_r^{n-1}\mathrm{d}r$, let $\hat{f}(r) = \int f_M(x)\mathrm{d}S_r^{n-1}$, thus

$$F(B_t) = \int_{B_t} f_M(x)\mathrm{d}S_u^{n-1}\mathrm{d}u = \int_0^t \int f_M(x)\mathrm{d}S_u^{n-1}\mathrm{d}u = \int_0^t \hat{f}(u)\mathrm{d}u = \hat{F}(t).$$

Finally, note that T is also rotationally invariant, thus $W_2(f_M, \mathbb{P}_{B_r}) = W_2(\hat{f}, \hat{g})$. It is sufficient to prove that $\hat{f}(u) = \mathrm{Vol}_{n-1}(S_u^{n-1})/\mathcal{V}$, thus by rotationally invariant $f_M(x) = 1/\mathcal{V}$ is a uniform distribution.

Note that $\hat{F}(t) \le \hat{G}(t)$ and $\hat{f}(u) = \int f_M(x)\mathrm{d}S_u^{n-1} \le \int f(x)\mathrm{d}S_u^{n-1} = \mathrm{Vol}_{n-1}(S_u^{n-1})/\mathcal{V}$. By a reformulation of the one dimensional Wasserstein distance [44]:

$$W_2(\hat{f}, \hat{g}) = \int_0^1 |\hat{F}^{-1}(t) - \hat{G}^{-1}(t)|^2 \mathrm{d}t$$

$$= \int_0^{r^\#} |x - \hat{G}^{-1}\left(\hat{F}(x)\right)|^2 \mathrm{d}\hat{F}(x), \tag{35}$$

which is just the area between between the graphs of $\hat{F}(r)$ and $\hat{G}(r)$. It is straightforward that the optimal $\hat{f}$ will maximize the growth rate of $\hat{F}$ in order to minimize the area, i.e. $\hat{f}(u) = \int f(x)\mathrm{d}S_u^{n-1} = 1/\mathcal{V}\mathrm{Vol}_{n-1}(S_u^{n-1})$. Therefore, $f_M(x) = 1/\mathcal{V}$ is a uniform distribution over $B_{r^\#}$ where $r^\#$ is the radius of $B_{r^\#}$ such that $\mathrm{Vol}(B_{r^\#}) = \mathcal{V}$. $\qquad\square$

Proof of Lemma 5, second approach. The proof starts in exactly the same way as in the first approach, up to the rotational invariance part. Instead of using the polar coordinate argument, we directly apply by invoking the second statement in Theorem 14, so $f_M = 1_{A_1}f$. But we know that $f(x) = 1/\mathcal{V}$ is a uniform distribution, and the claim follows. $\qquad\square$

Further note that by Equation (35), the optimal transport from $\hat{F}$ to $\hat{G}$ is

$$\hat{T}(u) = \hat{G}^{-1}\left(\hat{F}(u)\right) = \hat{G}^{-1}\left(\frac{1}{\mathcal{V}}\mathrm{Vol}_n(B_u^n)\right) = \left(\frac{\mathrm{Vol}_n(B_{r_U})}{\mathcal{V}}\right)^{1/n} u = \frac{r_U}{r_\mathcal{V}}r,$$

for $0 \le r \le r_M$. Note that T is rotationally symmetric, thus the optimal transport $T(x) = \frac{r_U}{r_\mathcal{V}}x$, for $x \in B_{r_\mathcal{V}}$

Lastly, it remains to prove

$$\inf_{W:\, \mathrm{Vol}_n(W) \ge \mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = \inf_{W:\, \mathrm{Vol}_n(W) = \mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W),$$

which follows the next lemma.

Lemma 6 (Monotonicity of Volume Comparison). *Given two balls B_{r_1} and B_{r_2} such that $\mathrm{Vol}(B_{r_1}) \ge \mathrm{Vol}(B_{r_2})$, then for any $A \subset \mathbb{R}^n$ such that $\mathrm{Vol}(A) \ge \mathrm{Vol}(B_{r_1})$,*

$$W_2(\mathbb{P}(A), \mathbb{P}(B_{r_2})) \ge W_2(\mathbb{P}(B_{r_1}), \mathbb{P}(B_{r_2})).$$

Proof of Lemma 6. We have shown that $W_2(\mathbb{P}(A), \mathbb{P}(B_{r_2})) \ge W_2(\mathbb{P}(B_{r_A}), \mathbb{P}(B_{r_2}))$, where B_{r_A} is a ball with Volume $\mathrm{Vol}(A)$. It remains to prove that

$$W_2(\mathbb{P}(B_{r_A}), \mathbb{P}(B_{r_2})) \ge W_2(\mathbb{P}(B_{r_1}), \mathbb{P}(B_{r_2}))$$

23

Let $T_A(x) = \frac{r_2}{r_A}x$, and $T_1(x) = \frac{r_2}{r_1}x$. By Theorem 14,

$$
\begin{aligned}
W_2^2(\mathbb{P}(B_{r_A}^n), \mathbb{P}(B_{r_2}^n)) &= \int_{B_R^n} |x - T_A(x)|^2 \mathrm{d}\mathbb{P}_{B_{r_A}^n} \\
&= \int_{B_R^n} \left| x - \frac{r_2}{r_A}x \right|^2 \mathrm{d}\mathbb{P}_{B_{r_A}^n} \\
&= \int_{B_R^n} \left(1 - \frac{r_2}{r_A}\right)^2 |x|^2 \, \mathrm{d}\mathbb{P}_{B_{r_A}^n} \\
&\geq \int_{B_R^n} \left(1 - \frac{r_2}{r_1}\right)^2 |x|^2 \, \mathrm{d}\mathbb{P}_{B_{r_A}^n} \\
&= \int_{B_R^n} |x - T_1(x)|^2 \mathrm{d}\mathbb{P}_{B_{r_A}^n} \\
&= W_2^2(\mathbb{P}(B_{r_1}^n), \mathbb{P}(B_{r_2}^n))
\end{aligned}
$$

$\square$

To make Theorem 5 complete, it remains to investigate the remaining cases when $0 < \mathcal{V} < \mathrm{Vol}_n(U)$.

Proof. We claim that when $0 < \mathcal{V} < \mathrm{Vol}(U)$, $\inf_{W: \mathrm{Vol}_n(W)=\mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = 0$, and it is not attained by any set. Let $\mathrm{Vol}_n(W_k) = \mathcal{V}$ and keep $W_k \subset U$ such that the mass of W_k is evenly distributed among the intersection between successively finer rectangular grids and U. Inside each intersection, the two distributions have the same probability mass. Since both are uniform probability distributions, their densities scale inversely proportional to their support sizes inside the intersection. Each little intersection is inside a little cube with width $\frac{2R}{k}$. We take ξ to be the product measure between $\mathbb{P}(U)$ and $\mathbb{P}(W)$. Now, when we compute:

$$
W_2(\mathbb{P}_U, \mathbb{P}_W) = \inf_{\xi \in \Xi(\mathbb{P}_U, \mathbb{P}_W)} \mathbb{E}_{(a,b)\sim\xi}[\|a-b\|_2^2]^{1/2} \leq \mathbb{E}_{(a,b)\sim\xi}[\|a-b\|_2^2]^{1/2}.
$$

The integrand $\|a - b\|_2^2 \leq \sqrt{n}\frac{2R}{k}$. By letting $k \to \infty$ (finer grids), we see that $\inf_{W: \mathrm{Vol}_n(W)=\mathcal{V}} W_2(\mathbb{P}_U, \mathbb{P}_W) = 0$.

However, the infimum is not attained by any set W with $\mathrm{Vol}_n(W) = \mathcal{V} < \mathrm{Vol}_n(U)$. Without loss of generality, we assume $W \subset U$. Then $\mathrm{Vol}_n(U - W) > 0$. So $W_2(\mathbb{P}_U, \mathbb{P}_W) > 0$.

$\square$

F   Proofs for Section 4

We prove the proposition 2 here. We begin with a lemma.

Lemma 7 (One-To-One $\implies$ Perfect Precision). *Let $\mathcal{M}$ be a Riemannian manifold. Let $f : \mathcal{M} \to \mathbb{R}^m$ be an open map. Then f achieves perfect precision.*

Proof. f is an open map, mapping open sets to open sets. For every $U \subset \mathcal{M}$, $f(U)$ is open in $\mathbb{R}^m$. Since $f(U)$ is open and contains $y = f(x)$, there exists $r_V > 0$ such that $V \subset f(U)$. This implies $f^{-1}(V) \subset U$. But then $\mathrm{Precision}^f(U, V) = \frac{\mathrm{Vol}_n(f^{-1}(V)\cap U)}{\mathrm{Vol}_n(f^{-1}(V))} = 1$ for such V and U. $\square$

Proof of Proposition 2. Let $\mathcal{M}$ be an n-dimensional Riemannian manifold and $m \geq 2n$ be the embedding dimension. By the Whitney embedding theorem, there exists a smooth map f such that $f(\mathcal{M})$ embeds into $\mathbb{R}^m$. Thus f is an open map from $\mathcal{M}$ to $f(\mathcal{M})$. We now apply lemma 7 to arrive at the conclusion. $\square$

G Wasserstein many-to-one, discontinuity and cost

In general, we do not have theoretical lower bound for W_2 measure. It is natural to use the sample based Wasserstein distances as substitutes. We perform some preliminary study of this heuristics below.

Recall Wasserstein distance is the minimal cost for mass-preserving transportation between regions. The Wasserstein L^2 distance is:

$$W_2(\mathbb{P}_a, \mathbb{P}_b) = \inf_{\xi \in \Xi(\mathbb{P}_a, \mathbb{P}_b)} \mathbb{E}_{(a,b) \sim \xi}[\|a - b\|_2^2]^{1/2} \tag{36}$$

where $\Xi(\mathbb{P}_a, \mathbb{P}_b)$ denotes all joint distributions $\xi(a, b)$ whose marginal distributions are $\mathbb{P}_a$ and $\mathbb{P}_b$. Intuitively, among all possible ways of transporting the two distributions, it looks for the most efficient one. With the same intuition, we use Wasserstein distance between U and $f^{-1}(V)$[14] to measure precision (See Section 3.2). This not only captures similar overlapping information as the setwise precision: $\frac{Vol_n(f^{-1}(V) \cap U)}{Vol_n(f^{-1}(V))}$, but also captures the shape differences and distances between U and $f^{-1}(V)$. Similarly, Wasserstein distance between $f(U)$ and V may capture the degree of discontinuity. $W_2(\mathbb{P}_{f(U)}, \mathbb{P}_V)$ **captures continuity** and $W_2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)})$ **captures injectivity**.

In practice, we calculate Wasserstein distances between two groups of samples, $\{a_i\}$ and $\{b_j\}$, using algorithms from [6] . Specifically, we solve

$$\min_m \sum_i \sum_j d_{i,j} m_{i \to j} \ ,$$
$$\text{such that} : \ m_{i \to j} \geq 0, \ \sum_i m_{i \to j} = 1, \ \sum_j m_{i \to j} = 1, \tag{37}$$

where $d_{i,j}$ is the distance between a_i and b_j and $m_{i \to j}$ is the mass moved from a_i to b_j. When $\{a_i\} \subset U$ and $\{b_j\} \subset f^{-1}(V)$, it is **Wasserstein many-to-one**. When $\{a_i\} \subset f(U)$ and $\{b_j\} \subset V$, it is **Wasserstein discontinuity**. High many-to-one likely implies low precision, and high discontinuity likely implies low recall. The average of many-to-one and discontinuity is **Wasserstein cost**.

We note that our measures bypass some practical difficulties on using precision and recall as evaluation measures. The first issue was discussed in Section 3.2, where we discussed that precision and recall are always equal when computed naively. This defeats their very purpose for capture both continuity and injectivity. Computing them based on Equation (4) and Equation (5) is more sensible, but it introduces another difficulty in practice due to high dimensionality: the radii r_U and/or r_V need to be quite large in order for some (outlier data point) x to have a reasonable number of neighboring data points. Some x ends up having many neighboring points, while others have very few[15]. This introduces a high variance on the number of neighboring data points across x. Our Wasserstein measures bypass both practical issues: having a fixed number of neighbors won't make $W_2(\mathbb{P}_{f(U)}, \mathbb{P}_V)$ and $W_2(\mathbb{P}_U, \mathbb{P}_{f^{-1}(V)})$ equal. In our experiments, we choose 30 neighboring points for all of U, $f^{-1}(V)$, $f(U)$ and V.

G.1 Preliminary experiments on Wasserstein Measures, Compare Visualization Maps

In this section, we show preliminary results on using Wasserstein measures directly (instead of its lower bound) to choose between dimensionality reduction algorithms. We may interpret this as choosing between different information retrieval systems in the DR visualization context. Figure 4 and 5 show the visualization results of 5 different methods on the S-curve and Swiss roll toy datasets respectively. These include PCA, multidimensional scaling (MDS) [38], locally linear embedding (LLE) [42], Isomap [32] and t-SNE [21]. In the results of PCA and MDS, the mappings squeeze the original data into narrower regions in the 2D projection space. Squeezing naturally implies high degree of many-to-one. At the same time, PCA mapping is linear, the MDS mapping in this case is close to linear, which makes both PCA and MDS has a low discontinuity. For S-curve and Swiss roll, LLE, Isomap and t-SNE all works well in the sense that they successfully unwrapped the manifold. However, when local compression or stretch happens, the Wasserstein discontinuity and

---

[14]The regions U and $f^{-1}(V)$ are given uniform distribution, i.e. their densities are $\frac{1}{Vol_n(U)}$ and $\frac{1}{Vol_n(f^{-1}(V))}$
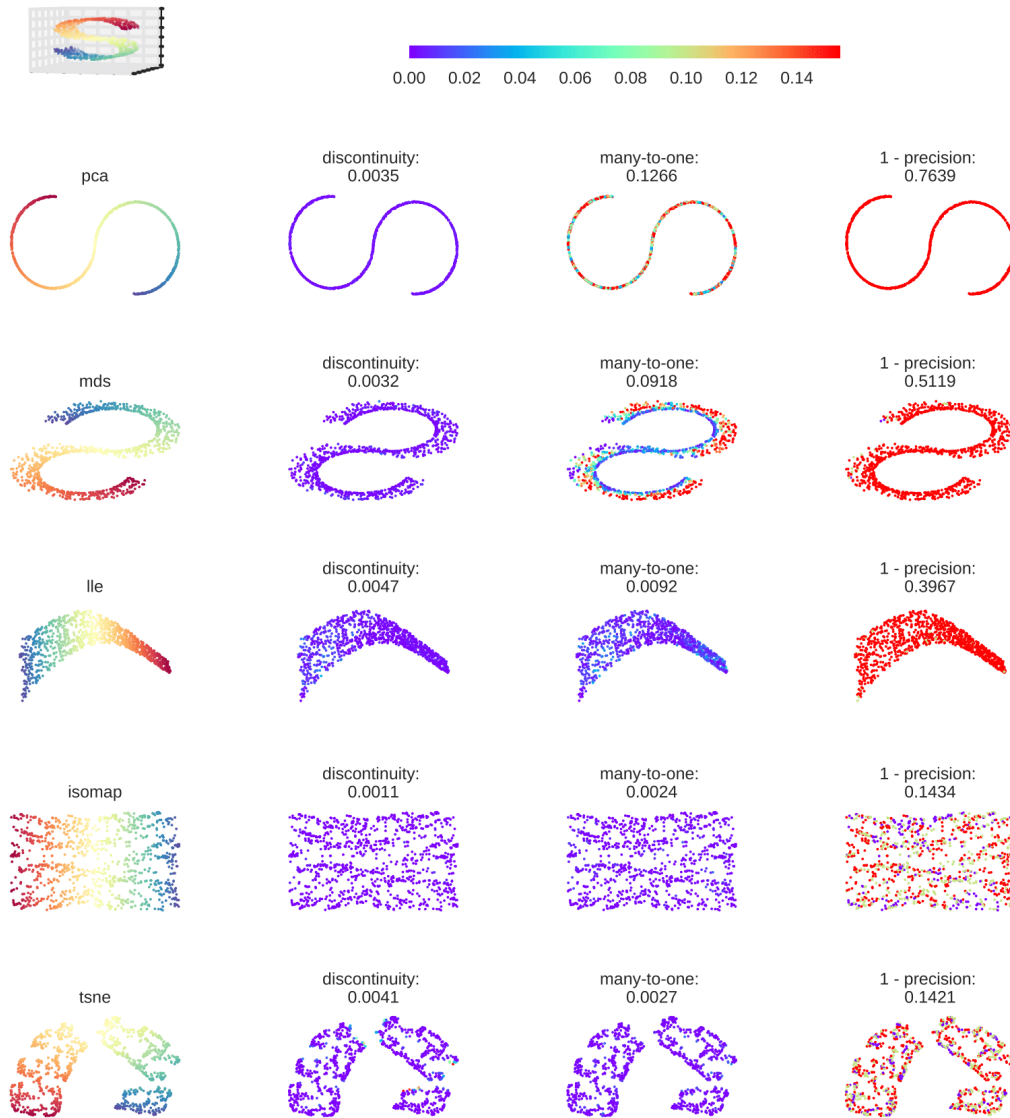
[15]This issue was also discussed in [21].

Figure 4: quality of different methods on S-curve

many-to-one will will increase slightly. For example, in the S-curve LLE results, the right side of data is compressed. Therefore it has a slightly higher many-to-one value, while the discontinuity is still low.

Figure 6 shows the visualization results on MNIST digits. As a linear map, PCA still has a relatively lower discontinuity and higher degree of many-to-one. MDS preserve global distances, at the cost of sacrificing local distances. thus can map nearby points to far away locations, at the same time mapping far a way points together has poor local one-to-one property. So it has both high discontinuity and many-to-one on MNIST digits. Compared with the previous toy example, LLE and Isomap both have a significant performance drop. Among all the methods, t-SNE still have the best local properties for MNIST digits, due to its neighborhood preservation objective.
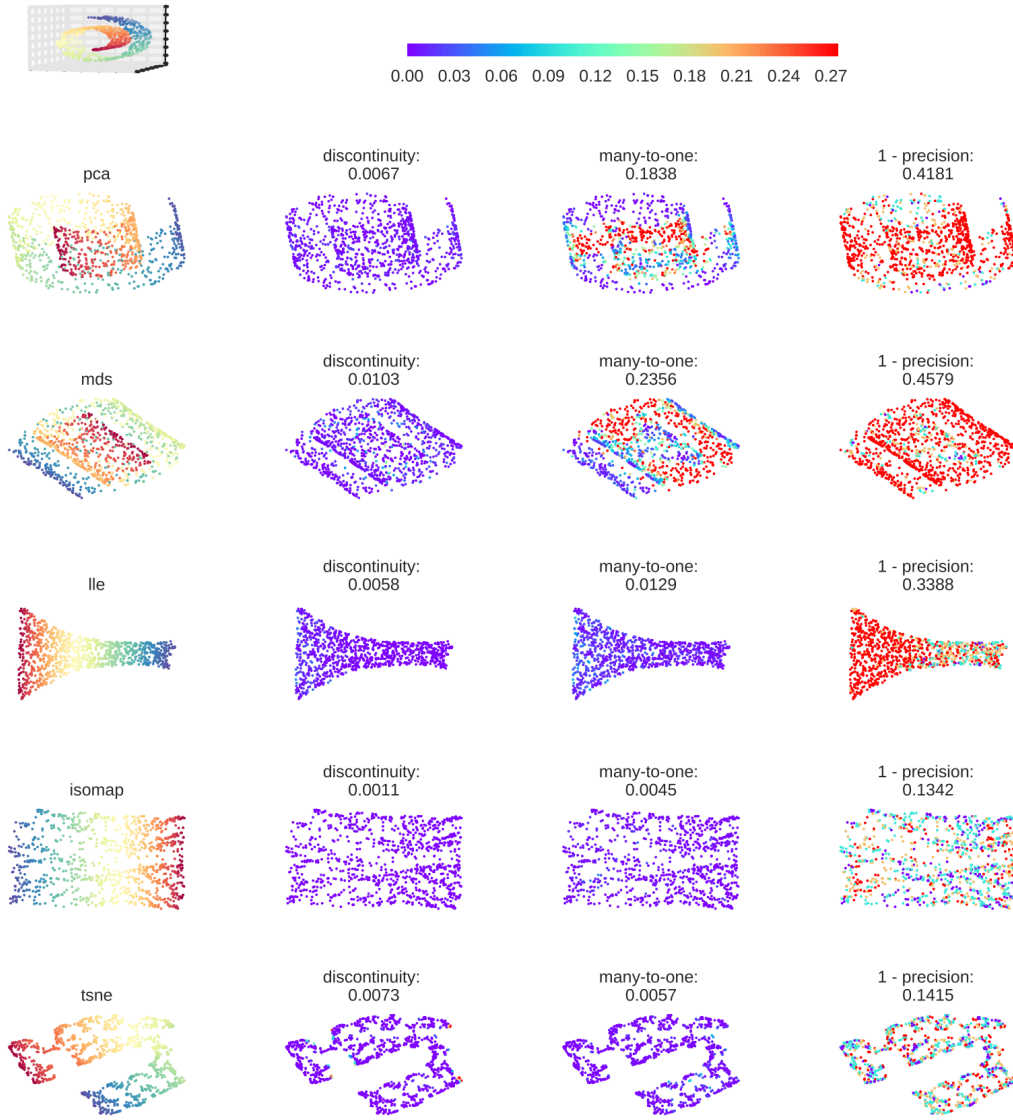
Figure 5: quality of different methods on Swiss roll

G.2 Preliminary experiments precision and recall (continuity v.s. injectivity) tradeoff

Theorem 1 suggests there is a trade-off between precision and recall, or equivalently continuity v.s. injectivity, via Proposition 1. In this section, we attempt to illustrate this tradeoff phenomenon by altering the degree of continuity of a DR algorithm in a practical situation. We choose t-SNE on MNIST because: 1) Heuristically t-SNE's perplexity parameter controls the degree of continuity: a higher perplexity means more neighboring data points will contract together and contraction is a continuous map (respectively, lower perplexity creates more tearing and spliting); 2) the tradeoff may be best seen through DR algorithms that operate at the optimal tradeoff level. t-SNE has proved itself as the de facto standard for visualization in various datasets; 3) As a practical dataset, MNIST visualization is still simple enough that humans can inspect and diagnose.

Fig. 7 shows visualizations with different t-SNE perplexity parameter. Each row is indexed by a different perplexity (perp $= 2, 8, \cdots, 1024$), with the intuition that the t-SNE DR map becomes more continuous with larger perplexity. The middel two columns are colored by our Wasserstein
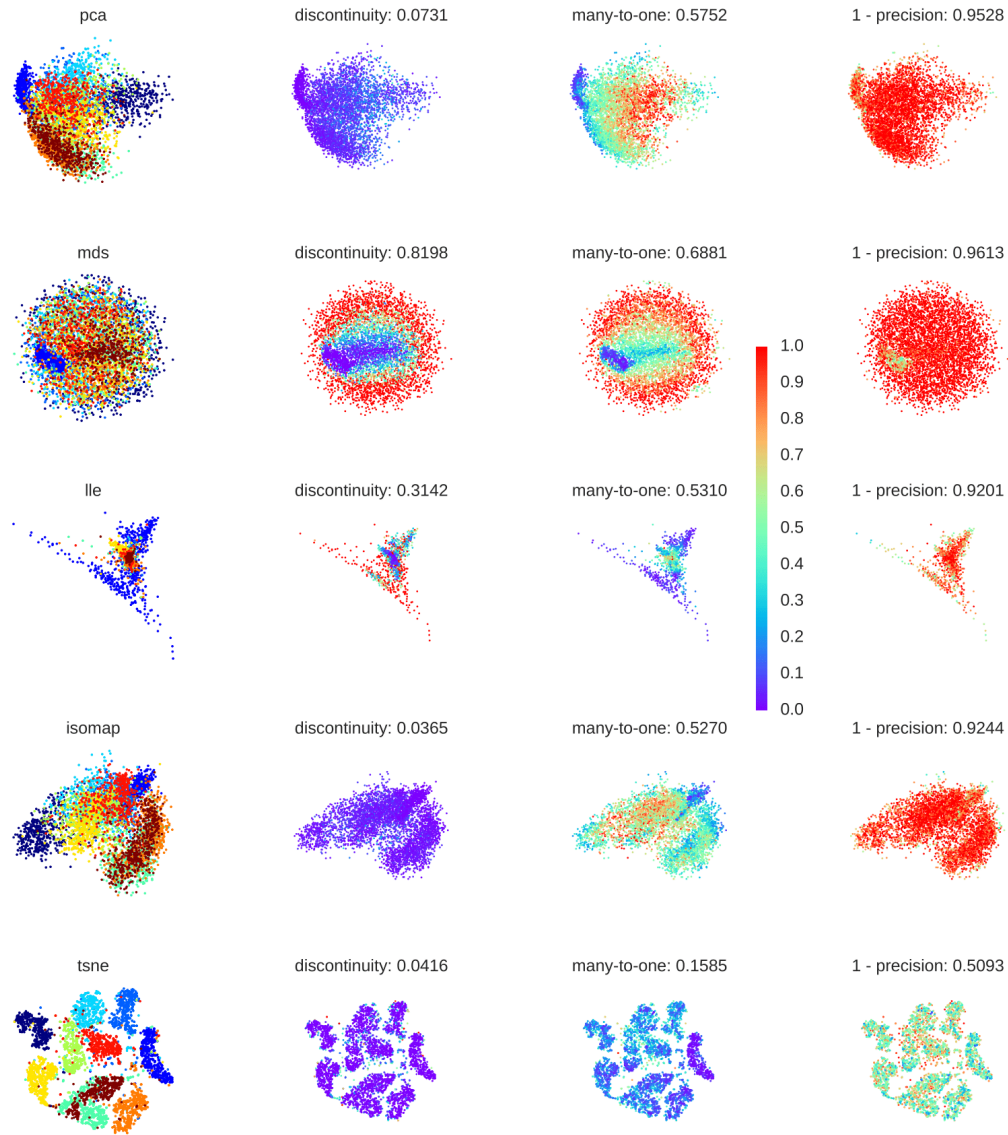
27

Figure 6: quality of different methods on MNIST

measures, with lower discontinuity costs representing more continuous maps (higher recall) and lower many-to-one costs indicating more injective (higher precision) maps. The precision and recall tradeoff can be observed in the perplexity ranging from 32 to 128. As t-SNE becomes more continuous, it is also less injective. In this range, inspection by eye suggests t-SNE gives good visualizations.

Outside of the range of $(32, 128)$ both precision and recall become worse. We interpret this as t-SNE is giving relatively bad visualizations for these choices of parameter, as can be inspected by eye. For example, when perplexity $= 512$ and 1024, t-SNE actually tends to have lower recall while precision worsens. When perplexity < 32, it is less clear whether it is due to: 1) there is a tradeoff but our measures do not capture it. Our neighborhood size is also 30 (comparable or bigger than the perplexity), so the scale may not be fair (on the other hand, choosing neighborhood size smaller than 30 may introduce very high variance in the estimation); 2) t-SNE actually performances worse on both continuity and injectivity, reflected by our measures. By inspection on the visualization, we believe it is probably because t-SNE isn't performing at any optimal level, so tradeoff cannot be seen.
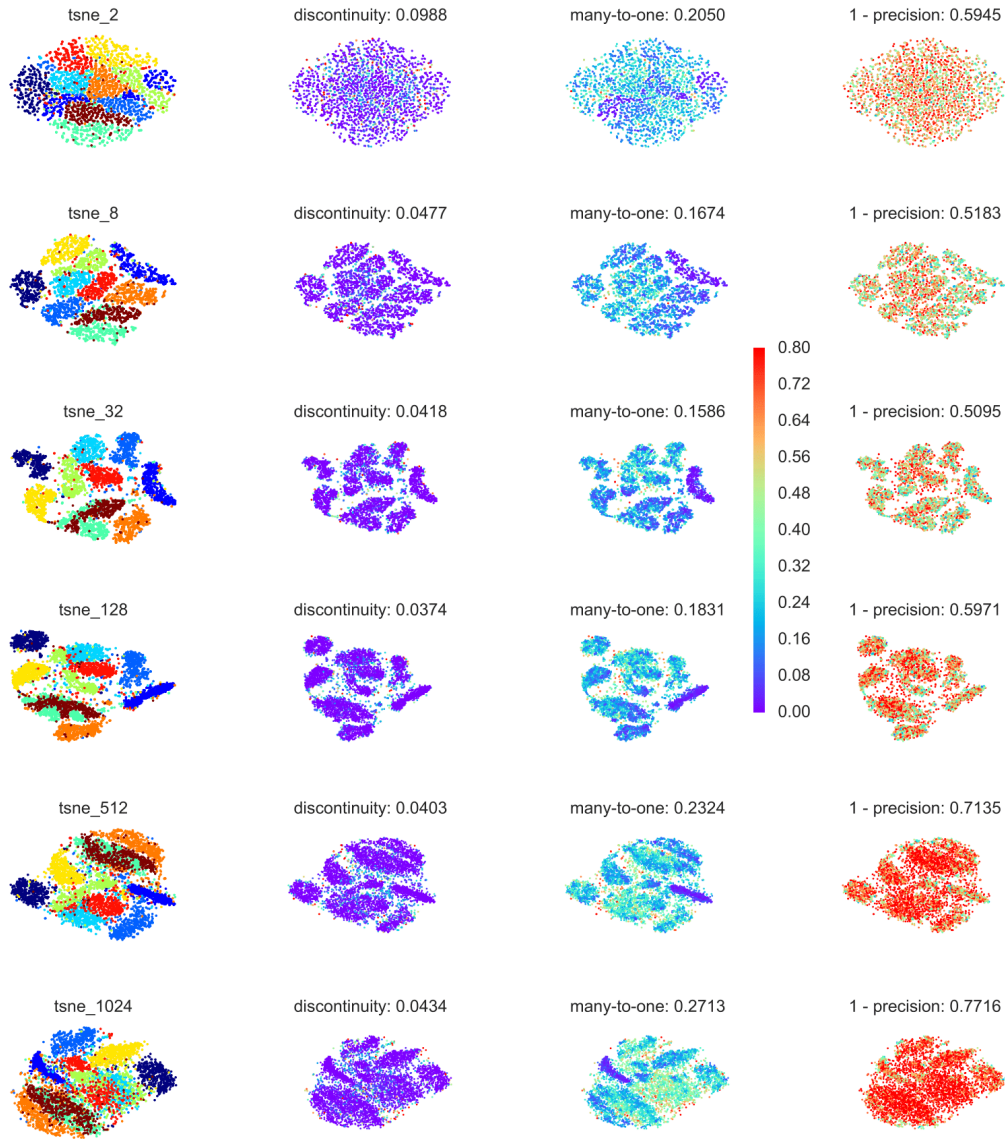
Figure 7: quality of t-SNE with different perplexities on MNIST

Supplementary References

[1] Arseniy Akopyan and Roman Karasev. A tight estimate for the waist of the ball. *Bulletin of the London Mathematical Society*, 49(4):690–693, 2017.

[3] Hannah Alpert and Larry Guth. A family of maps with many small fibers. *Journal of Topology and Analysis*, 7(01):73–79, 2015.

[38] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[9] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and Monge-Ampére obstacle problems. *Annals of mathematics*, 171:673–730, 2010.

[11] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010.

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[42] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[32] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[44] Sergei Sergeevich Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.