Intrinsic Dimension Estimation Using Wasserstein Distances

Adam Block
MIT

Zeyu Jia
MIT

Yury Polyanskiy
MIT

Alexander Rakhlin
MIT

Abstract

It has long been thought that high-dimensional data encountered in many practical machine learning tasks have low-dimensional structure, i.e., the manifold hypothesis holds. A natural question, thus, is to estimate the intrinsic dimension of a given population distribution from a finite sample. We introduce a new estimator of the intrinsic dimension and provide finite sample, non-asymptotic guarantees. We then apply our techniques to get new sample complexity bounds for Generative Adversarial Networks (GANs) depending only on the intrinsic dimension of the data.

1 Introduction

Recently, practical applications of machine learning involve a very large number of features, often many more than there are samples on which to train a model. Despite this imbalance, many modern machine learning models work astonishingly well. One of the more compelling explanations for this behavior is the manifold hypothesis, which posits that, though the data appear to the practitioner in a high-dimensional, ambient space, $\mathbb{R}^D$, they really lie on (or close to) a low dimensional space M of "dimension" $d \ll D$, where we define dimension formally below. A good example to keep in mind is that of image data: each of thousands of pixels corresponds to three dimensions, but we expect that real images have some inherent structure that limits the true number of degrees of freedom in a realistic picture. This phenomenon has been thoroughly explored over the years, beginning with the linear case and moving into the more general, nonlinear regime, with such works as Niyogi *et al.* (2008, 2011); Belkin & Niyogi (2001); Bickel *et al.* (2007); Levina & Bickel (2004); Kpotufe (2011); Kpotufe & Dasgupta (2012); Kpotufe & Garg (2013); Weed *et al.* (2019); Tenenbaum *et al.* (2000); Bernstein *et al.* (2000); Kim *et al.* (2019); Farahmand *et al.* (2007), among many, many others. Some authors have focused on finding representations for these lower dimensional sets (Niyogi *et al.*, 2008; Belkin & Niyogi, 2001; Tenenbaum *et al.*, 2000; Roweis & Saul, 2000; Donoho & Grimes, 2003), while other works have focused on leveraging the low dimensionality into statistically efficient estimators (Bickel *et al.*, 2007; Kpotufe, 2011; Nakada & Imaizumi, 2020; Kpotufe & Dasgupta, 2012; Kpotufe & Garg, 2013; Ashlagi *et al.*, 2021).

In this work, our primary focus is on estimating the intrinsic dimension. To see why this is an important question, note that the local estimators of Bickel *et al.* (2007); Kpotufe (2011); Kpotufe & Garg (2013) and the neural network architecture of Nakada & Imaizumi (2020) all depend in some way on the intrinsic dimension. As noted in Levina & Bickel (2004), while a practitioner may simply apply cross-validation to select the optimal hyperparameters, this can be very costly unless the hyperparameters have a restricted range; thus, an estimate of intrinsic dimension is critical in actually applying the above works. In addition, for manifold learning, where the goal is to construct a representation of the data manifold in a lower dimensional space, the intrinsic dimension is a key parameter in many of the most popular methods (Tenenbaum *et al.*, 2000; Belkin & Niyogi, 2001; Donoho & Grimes, 2003; Roweis & Saul, 2000).

We propose a new estimator, based on distances between probability distributions, as well as provide rigorous, finite sample guarantees for the quality of the novel procedure. Recall that if μ, ν are two measures on a metric space (M, d_M), then the Wasserstein-p distance between μ and ν is

$$W_p^M(\mu,\nu)^p = \inf_{(X,Y)\sim\Gamma(\mu,\nu)} \mathbb{E}\left[d_M(X,Y)^p\right] \tag{1}$$

where $\Gamma(\mu,\nu)$ is the set of all couplings of the two measures. If $M \subset \mathbb{R}^D$, then there are two natural metrics to put on M: one is simply the restriction of the Euclidean metric to M while the other is the geodesic metric in M, i.e., the minimal length of a curve in M that joins the points under consideration. In the sequel, if the metric is simply the Euclidean metric, we leave the Wasserstein distance unadorned to distinguish it from the intrinsic metric. For a thorough treatment of such distances, see Villani (2008). We recall that the Hölder integral probability metric (Hölder IPM) is given by

$$d_{\beta,B}(\mu,\nu) = \sup_{f \in C_B^\beta(\Omega)} \mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(Y)]$$

where $C_B^\beta(\Omega)$ is the Hölder ball defined in the sequel. When $p = \beta = 1$, the classical result of Kantorovich-Rubinstein says that the Wasserstein and Hölder distances agree. It has been known at least since Dudley (1969) that if a space M has dimension d, $\mathbb{P}$ is a measure with support M, and P_n is the empirical measure of n independent samples drawn from $\mathbb{P}$, then $W_1^M(P_n, \mathbb{P}) \asymp n^{-\frac{1}{d}}$. More recently, Weed *et al.* (2019) has determined sharp rates for the convergence of this quantity for higher order Wasserstein distances in terms of the intrinsic dimension of the distribution. Below, we find sharp rates for the convergence of the empirical measure to the population measure with respect to the Hölder IPM: if $\beta < \frac{d}{2}$, then $d_\beta(P_n, \mathbb{P}) \asymp n^{-\frac{\beta}{d}}$ and if $\beta > \frac{d}{2}$ then $d_\beta(P_n, \mathbb{P}) \asymp n^{-\frac{1}{2}}$. These sharp rates are intuitive in that convergence to the population measure should only depend on the intrinsic complexity (i.e. dimension) without reference to the possibly much larger ambient dimension.

The above convergence results are nice theoretical insights, but they have practical value, too. The results of Dudley (1969); Weed *et al.* (2019), as well as our results on the rate of convergence of the Hölder IPM, present a natural way to estimate the intrinsic dimension: take two independent samples, $P_n, P_{\alpha n}$ from $\mathbb{P}$ and consider the ratio of $W_p^M(P_n, \mathbb{P})/W_p^M(P_{\alpha n}, \mathbb{P})$ or $d_\beta(P_n, \mathbb{P})/d_\beta(P_{\alpha n}, \mathbb{P})$; as $n \to \infty$, the first ratio should be about α^d, while the second should be about $\alpha^{\frac{\beta}{d}}$, and so d can be computed by taking the logarithm with respect to α. The first problem with this idea is that we do not know $\mathbb{P}$; to address this, we instead compute the ratios using two independent samples. A more serious issue regards how large n must be in order for the asymptotic regime to apply. As we shall see below, the answer depends on the geometry of the supporting manifold.

We define two estimators: one using the intrinsic distance and the other using Euclidean distance

$$d_n = \frac{\log \alpha}{\log W_1(P_n, P_n') - \log W_1(P_{\alpha n}, P_{\alpha n}')} \qquad \widetilde{d}_n = \frac{\log \alpha}{\log W_1^G(P_n, P_n') - \log W_1^G(P_{\alpha n}, P_{\alpha n}')} \tag{2}$$

where the primes indicate independent samples of the same size and G is a graph-based metric that approximates the intrinsic metric. Before we go into the details, we give an informal statement of our main theorem, which provides finite sample, non-asymptotic guarantees on the quality of the estimator[1]:

Theorem 1 (Informal version of Theorem 22)**.** *Let $\mathbb{P}$ be a measure on $\mathbb{R}^D$ supported on a compact manifold of dimension d. Let τ be the reach of M, an intrinsic geometric quantity defined below. Suppose we have N independent samples from $\mathbb{P}$ where*

$$N = \Omega\left(\tau^{-d} \vee \left(\frac{\operatorname{vol} M}{\omega_d}\right)^{\frac{d+2}{2\gamma}} \vee \left(\log \frac{1}{\rho}\right)^3\right)$$

where ω_d is the volume of a d-dimensional Euclidean unit ball. Then with probability at least $1 - 6\rho$, the estimated dimension $\widetilde{d}_n$ satisfies

$$\frac{d}{1+4\gamma} \leq \widetilde{d}_n \leq (1+4\gamma)d.$$

The same conclusion holds for d_n.

Although the guarantees for d_n and $\widetilde{d_n}$ are similar, empirically $\widetilde{d}_n$ is much better, as explained below. Note that the ambient dimension D never enters the statistical complexity given above. While the exponential dependence on the intrinsic dimension d is unfortunate, it is likely necessary as described below.

While the reach, τ, determines the sample complexity of our dimension estimator, consideration of the injectivity radius, ι, is relevant for practical application. Both geometric quantities are defined formally in

[1] Explicit constants are given in the formal statement of Theorem 22

the following section, but, to understand the intuition, note that, as discussed above, there are two natural metrics we could be placing on $M = \operatorname{supp} \mathbb{P}$, the Euclidean metric and the geodesic distance. The reach is, intuitively, the size of the largest ball with respect to the ambient metric such that we can treat points in M as if they were simply in Euclidean space; the injectivity radius is similar, except it treats neighborhoods with respect to the intrinsic metric. Considering that manifold distances are always at least as large as Euclidean distances, it is unsurprising that $\tau \lesssim \iota$. Getting back to dimension estimation, specializing to the case of $\beta = p = 1$, and recalling (2), there are now two choices for our dimension estimator: we could use Wasserstein distance with respect to the Euclidean metric or Wasserstein distance with respect to the intrinsic metric (which we will denote by W_1^M). We will see that if $\iota \approx \tau$, then the two estimators induced by each of these distances behave similarly, but when $\iota \gg \tau$, the latter is better. While we wish to use $W_1^M(P_n, P_n')$ to estimate the dimension, we do not know the intrinsic metric. As such, we use the kNN graph to approximate this intrinsic metric and introduce the measure $W_1^G(P_n, P_n')$. Note that if we had oracle access to geodesic distance d_M, then the W_1^M-based estimator $\widetilde{d}_n$ would only require $\asymp \iota^{-d}$ samples. However, our kNN estimator of d_M, unfortunately, still requires the τ^{-d} samples. Nevertheless, there is a practical advantage of $\widetilde{d}_n$ in that the metric estimator can leverage all $N = 2(1+\alpha)n$ available samples, so that $\widetilde{d}_n$ works if $N \gtrsim \tau^{-d}$ and only $n \gtrsim \iota^{-d}$, whereas for d_n we require $n \gtrsim \tau^{-d}$ itself.

A natural question: is this more complicated approach necessary? i.e., is $\iota \gg \tau$ on real datasets? We believe that the answer is yes. To see this, consider the case of images of the digit 7 (for example) from MNIST (LeCun & Cortes, 2010). As a demonstration, we sample images from MNIST in datasets of size ranging in powers of 2 from 32 to 2048, calculate the Wasserstein distance between these two samples, and plot the resulting trend. In the right plot, we pool all of the data to estimate the manifold distances, and then use these estimated distances to compute the Wasserstein distance between the empirical distributions. In order to better compare these two approaches, we also plot the residuals to the linear fit that we expect in the asymptotic regime. Looking at Figure 1, it is clear that we are not yet in the asymptotic regime if we simply use Euclidean distances; on the other hand, the trend using the manifold distances is much more clearly linear, suggesting that the slope of the best linear fit is meaningful. Thus we see that in order to get a meaningful dimension estimate from practical data sets, we cannot simply use W_1 but must also estimate the geometry of the underlying distribution; this suggests that $\iota \gg \tau$ on this data manifold. More generally, we note that the injectivity radius, ι, is *intrinsic* to the geometry of the manifold and thus unaffected by the imbedding; in contradistinction, the reach, τ, is *extrinsic* and thus can be made smaller by changing the imbedding. In particular, when the obstruction to the reach being large is a "bottleneck" in the sense that the manifold is imbedded in such a way as to place distant neighborhoods of the manifold close together in Euclidean distance (see Figure 2 for an example), we may expect $\tau \ll \iota$. Intuitively, this matches the notion that the geometry of the data would be simple if we were to have access to the "correct" coordinate system and that the difficulty in understanding the geometry comes from its imbedding in the ambient space.

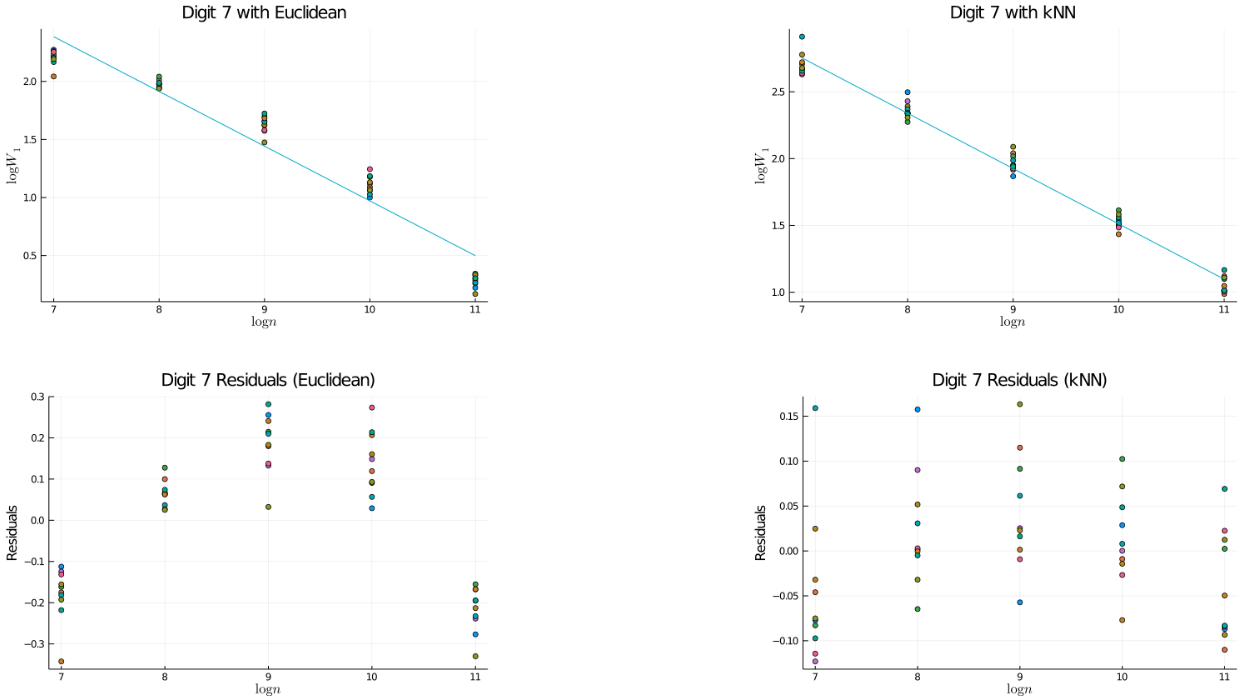

Figure 1: Two log-log plots of comparing how $W_1(P_n, P_n')$ decays to how $W_1^M(P_n, P_n')$ decays as n gets larger, as well as the residuals from a linear fit. The data are images of the digit 7 from MNIST with Wasserstein distances computed with the Sinkhorn algorithm (Cuturi, 2013). The manifold distances are approximated by a k-NN graph, as described in Section 3.

We emphasize that, like many estimators of intrinsic dimension, we do not claim robustness to off-manifold noise (Levina & Bickel, 2004; Farahmand *et al.*, 2007; Kim *et al.*, 2019). Indeed, any "fattening" of the manifold will force any consistent estimator of intrinsic dimension to asymptotically grow to the full, ambient dimension as the number of samples grows. Various works have included off-manifold noise in different ways, often with the assumption that either the noise is known (Koltchinskii, 2000) or the manifold is linear (Niles-Weed & Rigollet, 2019). Methods that do not make these simplifying assumptions are often highly sensitive to scaling parameters that are required inputs in such methods as multi-scale, local SVD (Little *et al.*, 2009). Extensions of our method to such noisy settings are a promising avenue of future research, particularly in understanding the effect of this noise on downstream applications as is done for Lipschitz classification in metric spaces and the resulting dimension-distortion tradeoff found in Gottlieb *et al.* (2016); in this work, however, we confine our theoretical study to the noiseless setting. The primary theoretical advantage of our estimator over that of Levina & Bickel (2004); Farahmand *et al.* (2007) is that we do not require the stringent regularity assumptions for our nonasymptotic rates to hold. We leave it for future empirical works whether this weakening of assumptions allows for a better practical estimator on real-world data sets.

Our main contributions are as follows:

- In Section 3, we introduce a new estimator of intrinsic dimension. In Theorem 22 we prove non-asymptotic bounds on the quality of the introduced estimator. Moreover, unlike the MLE estimator of Levina & Bickel (2004) with non-asymptotic analysis in Farahmand *et al.* (2007), minimal regularity of the density of the population distribution is required for our guarantees and, unlike that suggested in Kim *et al.* (2019), our estimator is both computationally efficient and has sample complexity independent of the ambient dimension.

- In the course of proving Theorem 22, we adapt the techniques of Bernstein *et al.* (2000) to provide new, non-asymptotic bounds on the quality of kNN distance as an estimate of intrinsic distance in

Proposition 24, with explicit sample complexity in terms of the reach of the underlying space. To our knowledge, these are the first such non-asymptotic bounds.

We further note that the techniques we develop to prove the non-asymptotic bounds on our dimension estimator also serve to provide new statistical rates in learning Generative Adversarial Networks (GANs) with a Hölder discriminator class:

- We prove in Theorem 25 that if $\widehat{\mu}$ is a Hölder GAN, then the distance between $\widehat{\mu}$ and $\mathbb{P}$, as measured by the Hölder IPM, is governed by rates dependent only on the intrinsic dimension of the data, independent of the ambient dimension or the dimension of the feature space. In particular, we prove in great generality that if $\mathbb{P}$ has intrinsic dimension d, then the rate of a Wasserstein GAN is $n^{-\frac{1}{d}}$. This improves on the recent work of Schreuder *et al.* (2020).

The work is presented in the order of the above listed contributions, preceded by a brief section on the geometric preliminaries and prerequisite results. We conclude the introduction by fixing notation and surveying some related work.

Notation: We fix the following notation. We always let $\mathbb{P}$ be a probability distribution on $\mathbb{R}^D$ and, whenever defined, we let $d = \dim \operatorname{supp} \mathbb{P}$. We reserve $X_1, \ldots, X_n$ for samples taken from $\mathbb{P}$ and we denote by P_n their empirical distribution. We reserve β for the smoothness of a Hölder class, $\Omega \subset \mathbb{R}^D$ is always a bounded open domain, and Δ is always the intrinsic diameter of a closed set. We also reserve M for a compact manifold. In general, we denote by $\mathcal{S}$ the support of a distribution $\mathbb{P}$ and we reuse $M = \operatorname{supp} \mathbb{P}$ if we restrict ourselves to the case where $\mathcal{S} = M$ is a compact manifold, with Riemannian metric induced by the Euclidean metric. We denote by $\operatorname{vol} M$ the volume of the manifold with respect to its inherited metric and we reserve ω_d for the volume of the unit ball in $\mathbb{R}^d$. When a compact manifold manifold M can be assumed from context, we take the *uniform* measure on M to be the volume measure of M normalized so that M has unit measure.

1.1 Related Work

Dimension Estimation There is a long history of dimension estimation, beginning with linear methods such as thresholding principal components (Fukunaga & Olsen, 1971), regressing k-Nearest-Neighbors (kNN) distances (Pettis *et al.*, 1979), estimating packing numbers (Kégl, 2002; Grassberger & Procaccia, 2004; Camastra & Vinciarelli, 2002), an estimator based solely on neighborhood (but not metric) information that was recently proven consistent (Kleindessner & Luxburg, 2015), and many others. An exhaustive recent survey on the history of these techniques can be found in Camastra & Staiano (2016). Perhaps the most popular choice among current practitioners is the MLE estimator of Levina & Bickel (2004).

The MLE estimator is constructed as the maximum likelihood of a parameterized Poisson process. As worked out in Levina & Bickel (2004), a local estimate of dimension for $k \geq 2$ and $x \in \mathbb{R}^D$ is given by

$$\widehat{m}_k(x) = \left(\frac{1}{k-1} \sum_{j=1}^{k} \log \frac{T_k(x)}{T_j(x)} \right)^{-1}$$

where $T_j(x)$ is the distance between x and its j^{th} nearest neighbor in the data set. The final estimate for fixed k is given by averaging $\widehat{m}_k$ over the data points in order to reduce variance. While not included in the original paper, a similar motivation for such an estimator could be noting that if X is uniformly distributed on a ball of radius R in $\mathbb{R}^d$, then $\mathbb{E}\left[\log \frac{R}{||X||}\right] = \frac{1}{d}$; the local estimator $\widehat{m}_k(x)$ is the empirical version under the assumption that the density is smooth enough to be approximately constant on this small ball. The easy computation is included for the sake of completeness in Appendix E. In Farahmand *et al.* (2007), the authors examined a closely related estimator and provided non-asymptotic guarantees with an exponential dependence on the intrinsic dimension, albeit with stringent regularity conditions on the density.

In addition to the estimators motivated by the volume growth of local balls discussed in the previous paragraph, Kim *et al.* (2019) proposed and analyzed a dimension estimator based on Travelling Salesman Paths (TSP). One major advantage to the TSP estimator is the lack of necessary regularity conditions on the density, requiring only an upper bound of the likelihood of the population density with respect to the volume measure on the manifold. On the other hand, the upper bound on sample complexity that that paper presents depends exponentially on the ambient dimension, which is pessimistic when the intrinsic dimension is substantially smaller. In addition, it is unclear how practical the estimator is due to the necessity of computing a solution to TSP; even ignoring this issue, Kim *et al.* (2019) note that practical tuning of the constants involved in their estimator is difficult and thus deploying their estimator as is on real-world datasets is unlikely.

Manifold Learning The notion of reach was first introduced in Federer (1959), and subsequently used in the machine learning and computational geometry communities in such works as Niyogi *et al.* (2008, 2011); Aamari *et al.* (2019); Amenta & Bern (1999); Fefferman *et al.* (2016, 2018); Narayanan & Mitter (2010); Efimov *et al.* (2019); Boissonnat *et al.* (2019). Perhaps most relevant to our work, Narayanan & Mitter (2010); Fefferman *et al.* (2016) consider the problem of testing membership in a class of manifolds of large reach and derive tight bounds on the sample complexity of this question. Our work does not fall into the purview of their conclusions as we assume that the geometry of the underlying manifold is nice and estimate the intrinsic dimension. In the course of proving bounds on our dimension estimator, we must estimate the intrinsic metric of the data. We adapt the proofs of Tenenbaum *et al.* (2000); Bernstein *et al.* (2000); Niyogi *et al.* (2008) and provide tight bounds on the quality of a k-Nearest Neighbors (kNN) approximation of the intrinsic distance.

Statistical Rates of GANs Since the introduction of Generative Adversarial Networks (GANs) in Goodfellow *et al.* (2014), there has been a plethora of empirical improvements and theoretical analyses. Recall that the basic GAN problem selects an estimated distribution $\widehat{\mu}$ from a class of distributions $\mathcal{P}$ minimizing some adversarially learned distance between $\widehat{\mu}$ and the empirical distribution P_n. Theoretical analyses aim to control the distance between the learned distribution $\widehat{\mu}$ and the population distribution $\mathbb{P}$ from which the data comprising P_n are sampled. In particular statistical rates for a number of interesting discriminator classes have been proven including Besov balls (Uppal *et al.*, 2019), balls in an RKHS (Liang, 2018), and neural network classes (Chen *et al.*, 2020) among others. The latter paper, Chen *et al.* (2020) also considers GANs where the discriminative class is a Hölder ball, which includes the popular Wasserstein GAN framework of Arjovsky *et al.* (2017). They show that if $\widehat{\mu}$ is the empirical minimizer of the GAN loss and the population distribution $\mathbb{P} \ll \mathsf{Leb}_{\mathbb{R}^D}$ then

$$\mathbb{E}\left[d_\beta(\widehat{\mu}, \mathbb{P})\right] \lesssim n^{-\frac{\beta}{2\beta+D}}$$

up to factors polynomial in $\log n$. Thus, in order to beat the curse of dimensionality, one requires $\beta = \Omega(D)$; note that the larger β is, the weaker the IPM is as the Hölder ball becomes smaller. In order to mitigate this slow rate, Schreuder *et al.* (2020) assume that both $\mathcal{P}$ and $\mathbb{P}$ are distributions arising from Lipschitz pushforwards of the uniform distribution on a d-dimensional hypercube; in this setting, they are able to remove dependence on D and show that

$$\mathbb{E}\left[d_\beta(\widehat{\mu}, \mathbb{P})\right] \lesssim L n^{-\frac{\beta}{d}} \vee n^{-\frac{1}{2}}.$$

This last result beats the curse of dimensionality, but pays with restrictive assumptions on the generative model as well as dependence on the Lipschitz constant of the pushforward map. More importantly, the result depends exponentially not on the intrinsic dimension of $\mathbb{P}$ but rather on the dimension of the feature space used to represent $\mathbb{P}$. In practice, state-of-the-art GANs used to produce images often choose d to be on the order of 128, which is much too large for the Schreuder *et al.* (2020) result to guarantee good performance.

2 Preliminaries

2.1 Geometry

In this work, we are primarily concerned with the case of compact manifolds isometrically imbedded in some large ambient space, $\mathbb{R}^D$. We note that this focus is largely in order to maintain simplicity of notation and exposition; extensions to more complicated, less regular sets with intrinsic dimension defined as the Minkowski dimension can easily be attained with our techniques. The key example to keep in mind is that of image data, where each pixel corresponds to a dimension in the ambient space, but, in reality, the distribution lives on a much smaller, imbedded subspace. Many of our results can be easily extended to the non-compact case with additional assumptions on the geometry of the space and tails of the distribution of interest.

Central to our study is the analysis of how complex the support of a distribution is. We measure complexity of a metric space by its entropy:

Definition 2. *Let (X, d) be a metric space. The covering number at scale $\varepsilon > 0$, $N(X, d, \varepsilon)$, is the minimal number s such that there exist points $x_1, \dots, x_s$ such that X is contained in the union of balls of radius ε centred at the x_i. The packing number at scale $\varepsilon > 0$, $D(X, d, \varepsilon)$, is the maximal number s such that there exist points $x_1, \dots, x_s \in X$ such that $d(x_i, x_j) > \varepsilon$ for all $i \neq j$. The entropy is defined as $\log N(X, d, \varepsilon)$.*

We recall the classical packing-covering duality, proved, for example, in (van Handel, 2014, Lemma 5.12):

Lemma 3. *For any metric space X and scale $\varepsilon > 0$,*

$$D(X, d, 2\varepsilon) \leq N(X, d, \varepsilon) \leq D(X, d, \varepsilon).$$

The most important geometric quantity that determines the complexity of a problem is the dimension of the support of the population distribution. There are many, often equivalent ways to define this quantity in general. One possibility, introduced in Assouad (1983) and subsequently used in Dasgupta & Freund (2008); Kpotufe & Dasgupta (2012); Kpotufe & Garg (2013) is that of doubling dimension:

Definition 4. *Let $\mathcal{S} \subset \mathbb{R}^D$ be a closed set. For $x \in \mathcal{S}$, the doubling dimension at x is the smallest d such that for all $r > 0$, the set $B_r(x) \cap \mathcal{S}$ can be covered by 2^d balls of radius $\frac{r}{2}$, where $B_r(x)$ denotes the Euclidean ball of radius r centred at x. The doubling dimension of $\mathcal{S}$ is the supremum of the doubling dimension at x for all $x \in \mathcal{S}$.*

This notion of dimension plays well with the entropy, as demonstrated by the following (Kpotufe & Dasgupta, 2012, Lemma 6):

Lemma 5 ((Kpotufe & Dasgupta, 2012))**.** *Let $\mathcal{S}$ have doubling dimension d and diameter Δ. Then $N(\mathcal{S}, \varepsilon) \leq \left(\frac{\Delta}{\varepsilon}\right)^d$.*

We remark that a similar notion of dimension is that of the *Minkowski dimension*, which is defined as the asymptotic rate of growth of the entropy as the scale tends to zero. Recently, Nakada & Imaizumi (2020) examined the effect that an assumption of small Minkowski dimension has on learning with neural networks; their central statistical result can be recovered as an immediate consequence of our complexity bounds below.

In order to develop non-asymptotic bounds, we need some understanding of the geometry of the support, M. We first recall the definition of the geodesic distance:

Definition 6. *Let $\mathcal{S} \subset \mathbb{R}^D$ be closed. A piecewise smooth curve in $\mathcal{S}$, γ, is a continuous function $\gamma : I \to \mathcal{S}$, where $I \subset \mathbb{R}$ is an interval, such that there exists a partition $I_1, \cdots, I_J$ of I such that γ_{I_j} is smooth as a function to $\mathbb{R}^D$. The length of γ is induced by the imbedding of $\mathcal{S} \subset \mathbb{R}^D$. For points $p, q \in \mathcal{S}$, the intrinsic (or geodesic) distance is*

$$d_{\mathcal{S}}(p, q) = \inf \left\{ \mathsf{length}\,(\gamma) | \gamma(0) = p \text{ and } \gamma(1) = q \text{ and } \gamma \text{ is a piecewise smooth curve in } \mathcal{S} \right\}.$$

It is clear from the fact that straight lines are geodesics in $\mathbb{R}^D$ that for any points $p, q \in \mathcal{S}$, $||p - q|| \leq d_{\mathcal{S}}(p, q)$. We are concerned with two relevant geometric quantities, one extrinsic and the other intrinsic.

Definition 7. *Let $\mathcal{S} \subset \mathbb{R}^D$ be a closed set. Let the medial axis $\mathrm{Med}(\mathcal{S})$ be defined as*

$$\mathrm{Med}(\mathcal{S}) = \left\{ x \in \mathbb{R}^D | \text{ there exist } p \neq q \in \mathcal{S} \text{ such that } ||p - x|| = ||q - x|| = d(x, \mathcal{S}) \right\}.$$

In other words, the medial axis is the set of points in $\mathbb{R}^D$ that have at least two projections to $\mathcal{S}$. Define the reach, $\tau_{\mathcal{S}}$ of $\mathcal{S}$ as $d(\mathcal{S}, \mathrm{Med}(\mathcal{S}))$, the minimal distance between a set and its medial axis.

If $\mathcal{S} = M$ is a compact manifold with the induced Euclidean metric, we define the injectivity radius $\iota = \iota_M$ as the maximal r such that if $p, q \in M$ such that $d_M(p, q) < r$ then there exists a unique length-minimizing geodesic connecting p to q in M.

For more detail on the injectivity radius, see Lee (2018), especially Chapters 6 and 10. The difference between ι_M and τ_M is in the choice of metric with which we equip M. We could choose to equip M with the metric induced by the Euclidean distance $||\cdot||$ or we could choose to use the intrinsic metric d_M defined above. The reach quantifies the maximal radius of a ball with respect to the *Euclidean* distance such that the intersection of this ball with M behaves roughly like Euclidean space. The injectivity radius, meanwhile, quantifies the maximal radius of a ball with respect to the *intrinsic* distance such that this ball looks like Euclidean space. While neither quantity is necessary for our dimension estimator, both figure heavily in the analysis. The final relevant geometric quantity is the sectional curvature. The sectional curvature of M at a point $p \in M$ given two directions tangent to M at p is given by the Gaussian curvature at p of the image of the exponential map applied to a small neighborhood of the origin in the plane determined by the two directions. Intuitively, the sectional curvature measures how tightly wound the manifold is locally around each point. For an excellent exposition on the topic, see (Lee, 2018, Chapter 8).

We now specialize to consider compact, dimension d manifolds M imbedded in $\mathbb{R}^D$ with the induced metric (see Lee (2018) for an accessible introduction to the geometric notions discussed here). One measure of size of the manifold M is the diameter, Δ, with respect to the intrinsic distance defined above. Another notion of size is the volume measure, vol_M. This measure can be defined intrinsically as integration with respect to the volume form, where the volume form can be thought of as the analogue of the Lebesgue differential in standard Euclidean space; for more details see Lee (2018). In our setting, we could equivalently define the volume as the d-dimensional Hausdorff measure as in Aamari *et al.* (2019). Either way, when we refer to a measure μ_M that is uniform on the manifold, we consider the normalization such that $\mu_M(M) = 1$, i.e., $\mu_M(\cdot) = \mathrm{vol}_M(\cdot)/\,\mathrm{vol}(M)$.

With the brief digression into volume concluded, we return to the notion of the reach, which encodes a number of local and global geometric properties. We summarize several of these in the following proposition:

Proposition 8. *Let $M \subset \mathbb{R}^D$ be a compact manifold isometrically imbedded in $\mathbb{R}^D$. Suppose that $\tau = \tau_M > 0$. The following hold:*

(a) (Niyogi et al.*, 2008, Proposition 6.1)] The norm of the second fundamental form of M is bounded by $\frac{1}{\tau}$ at all points $p \in M$.*

(b) (Aamari et al.*, 2019, Proposition A.1 (ii)) The injectivity radius of M is at least $\pi\tau$.*

(c) (Boissonnat et al.*, 2019, Lemma 3) If $p, q \in M$ such that $||p - q|| \leq 2\tau$ then $d_M(p, q) \leq 2\tau \arcsin\left(\frac{||p-q||}{2\tau}\right)$.*

A few remarks are in order. First, note that the Hopf-Rinow Theorem (Hopf & Rinow, 1931) guarantees that M is complete, which is fortuitous as completeness is a necessary, technical requirement for several of our arguments. Second, we note that (c) from Proposition 8 has a simple geometric interpretation: the upper bound on the right hand side is the length of the arc of a circle of radius τ containing points p, q; thus, the maximal distortion of the intrinsic metric with respect to the ambient metric is bounded by the circle of radius τ.

Point (a) in the above proposition demonstrates that control of the reach leads to control of local distortion. From the definition, it is obvious that the reach provides an upper bound for the size of the global notion of a "bottleneck," i.e., two points $p, q \in M$ such that $||p - q|| = 2\tau < d_M(p, q)$. Interestingly, these two local and global notions of distortion are the only ways that the reach of a manifold can be small, as (Aamari *et al.*, 2019, Theorem 3.4) tells us that if the reach of a manifold M is τ, then either there exists a bottleneck of size 2τ or the norm of the second fundamental form is $\frac{1}{\tau}$ at some point. Thus, in some sense, the reach is the "correct" measure of distortion. Note that while (b) above tells us that $\iota_M \gtrsim \tau_M$, there is no comparable upper bound. To see this, consider Figure 2, which depicts a one-dimensional manifold imbedded in $\mathbb{R}^2$. Note that the bottleneck in the center ensures that the reach of this manifold is very small; on the other hand, it is easy to see that the injectivity radius is given by half the length of the entire curve. As the curve can be extended arbitrarily, the reach can be arbitrarily small relative to the injectivity radius.

We now proceed to bound the covering number of a compact manifold using the dimension and the injectivity radius. We note that upper bounds on the covering number with respect to the ambient metric were provided in Niyogi *et al.* (2008); Narayanan & Mitter (2010). A similar bound with less explicit constants can be found in (Kim *et al.*, 2019, Lemma 4).

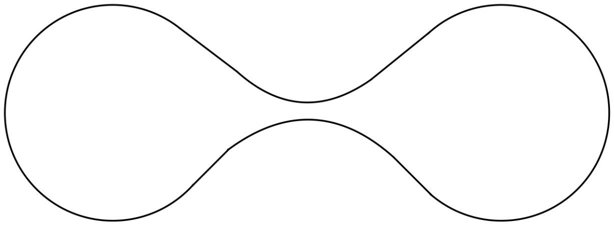

Figure 2: Curve in $\mathbb{R}^2$ where $\tau \ll \iota$.

Proposition 9. *Let $M \subset \mathbb{R}^D$ be an isometrically imbedded, compact, d-dimensional submanifold with injectivity radius $\iota > 0$ such that the sectional curvatures are bounded above by $\kappa_1 \geq 0$ and below by $\kappa_2 \leq 0$. If $\varepsilon < \frac{\pi}{2\sqrt{\kappa_1}} \wedge \iota$ then*

$$N(M, d_M, \varepsilon) \leq \frac{\operatorname{vol} M}{\omega_d} d \left(\frac{\pi}{2}\right)^d \varepsilon^{-d}.$$

If $\varepsilon < \frac{1}{\sqrt{-\kappa_2}} \wedge \iota$ then

$$\frac{\operatorname{vol} M}{\omega_d} d 8^{-d} \varepsilon^{-d} \leq D(M, d_M, 2\varepsilon).$$

Moreover, for all $\varepsilon < \iota$,

$$\frac{\operatorname{vol} M}{\omega_d} d \iota^d (-\kappa_2)^{\frac{d}{2}} e^{-d\iota\sqrt{-\kappa_2}} \varepsilon^{-d} \leq D(M, d_M, \varepsilon).$$

Thus, if $\varepsilon < \tau$, where τ is the reach of M, then

$$\frac{\operatorname{vol} M}{\omega_d} d 8^{-d} \varepsilon^{-d} \leq D(M, d_M, 2\varepsilon) \leq N(M, d_M, \varepsilon) \leq \frac{\operatorname{vol} M}{\omega_d} d \left(\frac{\pi}{2}\right)^d \varepsilon^{-d}.$$

The proof of Proposition 9 can be found in Appendix A and relies on the Bishop-Gromov comparison theorem to leverage the curvature bounds from Proposition 8 into volume estimates for small intrinsic balls, a similar technique as found in Niyogi *et al.* (2008); Narayanan & Mitter (2010). The key point to note is that we have both upper and lower bounds for $\varepsilon < \iota$, as opposed to just the upper bound guaranteed by Lemma 5. As a corollary, we are also able to derive bounds for the covering number with respect to the ambient metric:

Corollary 10. *Let M be as in Proposition 9. For $\varepsilon < \tau$, we can control the covering numbers of M with respect to the Euclidean metric as*

$$\frac{\operatorname{vol} M}{\omega_d} d 16^{-d} \varepsilon^{-d} \leq D(M, ||\cdot||, 2\varepsilon) \leq N(M, ||\cdot||, \varepsilon) \leq \frac{\operatorname{vol} M}{\omega_d} \left(\frac{\pi}{2}\right)^d \varepsilon^{-d}.$$

The proof of Corollary 10 follows from Proposition 9 and the metric comparisons for small scales in Proposition 8; details can be found in Appendix A.

2.2 Hölder Classes and their Complexity

In this section we make the elementary observation that complex function classes restricted to simple subsets can be much smaller than the original class. While such intuition has certainly appeared before, especially in designging esimators that can adapt to local intrinsic dimension, such as Bickel *et al.* (2007); Kpotufe & Dasgupta (2012); Kpotufe (2011); Kpotufe & Garg (2013); Dasgupta & Freund (2008); Steinwart *et al.* (2009); Nakada & Imaizumi (2020), we codify this approach below.

To illustrate the above phenomenon at the level of empirical processes, we focus on Hölder functions in $\mathbb{R}^D$ for some large D and let the "simple" subset be a subspace of dimension d where $d \ll D$. We first recall the definition of a Hölder class:

Definition 11. *For an open domain $\Omega \subset \mathbb{R}^d$ and a function $f : \Omega \to \mathbb{R}$, define the β-Hölder norm as*

$$||f||_{C^\beta(\Omega)} = \max_{0 \le |\gamma| \le |\alpha|} \sup_{x \in \Omega} |D^\gamma f(x)| \vee \sup_{x,y \in \Omega} \frac{\left|D^{\lfloor \beta \rfloor} f(x) - D^{\lfloor \beta \rfloor} f(y)\right|}{||x - y||^{\beta - \lfloor \beta \rfloor}}.$$

Define the Hölder ball of radius B, denoted by $C_B^\beta(\Omega)$, as the set of functions $f : \Omega \to \mathbb{R}$ such that $||f||_{C^\beta(\Omega)} \le B$. If (M, g) is a Riemannian manifold of class $C^{\lfloor \beta \rfloor + 1}$ (see Lee (2018)), and $f : M \to \mathbb{R}$ we define the Hölder norm analogously, replacing $|D^\gamma f(x)|$ with $||\nabla^\gamma f(x)||_g$, where ∇ is the covariant derivative.

It is a classical result of Kolmogorov & Tikhomirov (1993) that, for a bounded, open domain $\Omega \subset \mathbb{R}^D$, the entropy of a Hölder ball scales as

$$\log N\left(C_B^\beta(\Omega), ||\cdot||_\infty, \varepsilon\right) \asymp \left(\frac{B}{\varepsilon}\right)^{\frac{D}{\beta}}$$

as $\varepsilon \downarrow 0$. As a consequence, we arrive at the following result, whose proof can be found in Appendix A for the sake of completeness.

Proposition 12. *Let $\mathcal{S} \subset \Omega \subset \mathbb{R}^d$ be a path-connected closed set contained in an open domain Ω. Let $\widetilde{\mathcal{F}} = C_B^\beta(\Omega)$ and let $\mathcal{F} = \widetilde{\mathcal{F}}|_{\mathcal{S}}$. Then,*

$$D\left(\mathcal{S}, \left(\frac{\varepsilon}{B}\right)^{\frac{1}{\beta}}\right) \le \log D(\mathcal{F}, ||\cdot||_\infty, 2\varepsilon) \le \log N(\mathcal{F}, ||\cdot||_\infty, \varepsilon) \le 3\beta^2 \log\left(\frac{2B}{\varepsilon}\right) N\left(\mathcal{S}, \left(\frac{\varepsilon}{2B}\right)^{\frac{1}{\beta}}\right).$$

Note that the content of the above result is really that of Kolmogorov & Tikhomirov (1993), coupled with the fact that restriction from $\mathbb{R}^d$ to M preserves smoothness.

If we apply the easily proven volumetric bounds on covering and packing numbers for $\mathcal{S}$ a Euclidean ball to Proposition 12, we recover the classical result of Kolmogorov & Tikhomirov (1993). The key insight is that low-dimensional subsets can have covering numbers much smaller than those of a high-dimensional Euclidean ball: if the "dimension" of $\mathcal{S}$ is d, then we expect the covering number of $\mathcal{S}$ to scale like ε^{-d}. Plugging this into Proposition 12 tells us that the entropy of $\mathcal{F}$, up to a factor logarithmic in $\frac{1}{\varepsilon}$, scales like $\varepsilon^{-\frac{d}{\beta}} \ll \varepsilon^{-\frac{D}{\beta}}$. An immediate corollary of Lemma 5 and Proposition 12 is:

Corollary 13. *Let $\mathcal{S} \subset \mathbb{R}^D$ be a closed set of diameter Δ and doubling dimension d. Let $\mathcal{S} \subset \Omega$ open and $\mathcal{F}$ be the restriction of $C_B^\beta(\Omega)$ to $\mathcal{S}$. Then*

$$\log N(\mathcal{F}, ||\cdot||_\infty, \varepsilon) \le 3\beta^2 \left(\frac{2B\Delta^\beta}{\varepsilon}\right)^{\frac{d}{\beta}} \log\left(\frac{2B}{\varepsilon}\right).$$

Proof. Combine the upper bound in Proposition 12 with the bound in Lemma 5. ■

The conclusion of Corollary 13 is very useful for upper bounds as it tells us that the entropy for Hölder balls scales at most like $\varepsilon^{-\frac{d}{\beta}}$ as $\varepsilon \downarrow 0$. If we desire comparable lower bounds, we require some of the geometry discussed above. Combining Proposition 12 and Corollary 10 yields the following bound:

Corollary 14. *Let $M \subset \mathbb{R}^D$ be an isometrically imbedded, compact submanifold with reach $\tau > 0$ and let $\varepsilon \le \tau$. Suppose $\Omega \supset M$ is an open set and let $\mathcal{F}'$ be the restriction of $C_B^\beta(\Omega)$ to M. Then for $\varepsilon \le \tau$,*

$$\frac{\operatorname{vol} M}{\omega_d} d 16^{-d} \left(\frac{2B}{\varepsilon}\right)^{\frac{d}{\beta}} \le \log D(\mathcal{F}', ||\cdot||_\infty, 2\varepsilon) \le \log N(\mathcal{F}', ||\cdot||_\infty, \varepsilon) \le 3\beta^2 \log\left(\frac{2B}{\varepsilon}\right) \frac{\operatorname{vol} M}{\omega_d} d \left(\frac{\pi}{2}\right)^d \left(\frac{2B}{\varepsilon}\right)^{\frac{d}{\beta}}.$$

If we set $\mathcal{F} = C_B^\beta(M)$, then we have that for all $\varepsilon < \iota$,

$$\frac{\operatorname{vol} M}{\omega_d} d\iota^d (-\kappa_2)^{\frac{d}{2}} e^{-d\iota\sqrt{-\kappa_2}} \varepsilon^{-\frac{d}{\beta}} \le \log N(\mathcal{F}, ||\cdot||_\infty, \varepsilon) \le 3\beta^2 \log\left(\frac{2B}{\varepsilon}\right) \frac{\operatorname{vol} M}{\omega_d} d \left(\frac{\pi}{2}\right)^d \varepsilon^{-\frac{d}{\beta}}.$$

In essence, Corollary 14 tells us that the rate of $\varepsilon^{-\frac{d}{\beta}}$ for the growth of the entropy of Hölder balls is sharp for sufficiently small ε. The key difference between the first and second statements is that the first is with respect to an ambient class of functions while the second is with respect to an intrinsic class. To better illustrate the difference, consider the case where $\beta = B = 1$, i.e., the class of Lipschitz functions on the manifold. In both cases, asymptotically, the entropy of Lipschitz functions scales like ε^{-d}; if we restrict to functions that are Lipschitz with respect to the ambient metric, then the above bound only applies for $\varepsilon < \tau$; on the other hand, if we consider the larger class of functions that are Lipschitz with respect to the intrinsic metric, the bound applies for $\varepsilon < \iota$. In the case where $\iota \gg \tau$, this can be a major improvement.

The observations in this section are undeniably simple; the real interest comes in the diverse applications of the general principle, some of which we detail below. As a final note, we remark that our guiding principle of simplifying function classes by restricting them to simple sets likely holds in far greater analysis than is explored here; in particular, Sobolev and Besov classes (see, for example, (Giné & Nickl, 2016, §4.3)) likely exhibit similar behavior.

3 Dimension Estimation

We outlined the intuition behind our dimension estimation in the introduction. In this section, we formally define the estimator and analyse its theoretical performance. We first apply standard empirical process theory and our complexity bounds in the previous section to upper bound the expected Hölder IPM (defined in (1)) between empirical and population distributions:

Lemma 15. *Let $\mathcal{S} \subset \mathbb{R}^D$ be a compact set contained in a ball of radius R. Suppose that we draw n independent samples from a probability measure $\mathbb{P}$ supported on $\mathcal{S}$ and denote by P_n the corresponding empirical distribution. Let P_n' denote an independent identically distributed measure as P_n. Then we have*

$$\mathbb{E}\left[d_{\beta,B}(P_n, \mathbb{P})\right] \leq \mathbb{E}\left[d_{\beta,B}(P_n, P_n')\right] \leq 16B \inf_{\delta>0}\left(2\delta + \frac{3\sqrt{6}}{\sqrt{n}}\beta\sqrt{\log\frac{1}{\delta}}\int_\delta^1 \sqrt{N(\mathcal{S}, ||\cdot||, \varepsilon)}d\varepsilon\right).$$

In particular, there exists a universal constant K such that if $N(\mathcal{S}, ||\cdot||, \varepsilon) \leq C_1\varepsilon^{-d}$ for some $C, d > 0$, then

$$\mathbb{E}\left[d_\beta(P_n, \mathbb{P})\right] \leq C\beta B\left(1 + \sqrt{\log n}\mathbf{1}_{\{d=2\beta\}}\right)\left(n^{-\frac{\beta}{d}} \vee n^{-\frac{1}{2}}\right).$$

holds with $C = KC_1$.

The proof uses the symmetrization and chaining technique and applies the complexity bounds of Hölder functions found above; the details can be found in Appendix E.

We now specialize to the case where $\beta = B = 1$, due to the computational tractability of the resulting Wasserstein distance. Applying Kantorovich-Rubenstein duality (Kantorovich & Rubinshtein, 1958), we see that this special case of Lemma 15 recovers the special $p = 1$ case of Weed *et al.* (2019). From here on, we suppose that $d > 2$ and our metric on distributions is $d_{1,1} = W_1$.

We begin by noting that if we have $2n$, independent samples from $\mathbb{P}$, then we can split them into two data sets of size n, and denote by P_n, P_n' the empirical distributions thus generated. We then note that Lemma 15 implies that if $\operatorname{supp}\mathbb{P} \subset M$ and M is of dimension d, then

$$\mathbb{E}\left[W_1(P_n, P_n')\right] \leq C_{M,d}n^{-\frac{1}{d}}.$$

If we were to establish a lower bound as well as concentration of $W_1(P_n, P_n')$ about its mean, then we could consider the following estimator. Given a data set of size $2(\alpha+1)n$, we can break the data into four samples, P_n, P_n' each of size n and $P_{\alpha n}, P_{\alpha n}'$ of size αn. Then we would have

$$d_n := -\frac{1}{\log_\alpha\left(\frac{W_1(P_{\alpha n}, P_{\alpha n}')}{W_1(P_n, P_n')}\right)} = \frac{\log\alpha}{\log W_1(P_n, P_n') - \log W_1(P_{\alpha n}, P_{\alpha n}')} \approx d.$$

Which distance on M should be used to compute the Wasserstein distance, the Euclidean metric $||\cdot||$ or the intrinsic metric $d_M(\cdot,\cdot)$? As can be guessed from Corollary 14, asymptotically, both will work, but for finite sample sizes when $\iota \gg \tau$, the latter is much better. One problem remains, however: because we are not assuming M to be known, we do not have access to d_M and thus we cannot compute the necessary Wasserstein cost. In order to get around this obstacle, we recall the graph distance induced by a kNN graph:

Definition 16. *Let $X_1, \ldots, X_n \in \mathbb{R}^D$ be a data set and fix $\varepsilon > 0$. We let $G(X, \varepsilon)$ denote the weighted graph with vertices X_i and edges of weight $||X_i - X_j||$ between all vertices X_i, X_j such that $||X_i - X_j|| \leq \varepsilon$. We denote by $d_{G(X,\varepsilon)}$ (or d_G if X, ε are clear from context) the geodesic distance on the graph $G(X, \varepsilon)$. We extend this metric to all of $\mathbb{R}^D$ by letting*

$$d_G(p, q) = ||p - \pi_G(p)|| + d_G(\pi_G(p), \pi_G(q)) + ||q - \pi_G(q)||$$

where $\pi_G(p) \in \operatorname{argmin}_{X_i} ||p - X_i||$.

We now have two Wasserstein distances, each induced by a different metric; to mitigate confusion, we introduce the following notation:

Definition 17. *Let $X_1, \ldots, X_n, X_1', \ldots, X_n' \in \mathbb{R}^D$, sampled independently from $\mathbb{P}$ such that $\operatorname{supp} \mathbb{P} \subset M$. Let P_n, P_n' be the empirical distributions associated to the data X, X'. Let $W_1(P_n, P_n')$ denote the Wasserstein cost with respect to the Euclidean metric and $W_1^M(P_n, P_n')$ denote the Wasserstein cost associated to the manifold metric, as in (1). For a fixed $\varepsilon > 0$, let $W_1^G(P_n, P_n')$ denote the Wasserstein cost associated to the metric $d_{G(\operatorname{supp} P_n \cup \operatorname{supp} P_n', \varepsilon)}$. Let d_n, $\widehat{d}_n$, and $\widetilde{d}_n$ denote the dimension estimators from (3) induced by each of the above metrics.*

Given sample distributions P_n, P_n', we are able to compute $W_1(P_n, P_n')$ and $W_1^G(P_n, P_n')$ for any fixed ε, but not $W_1^M(P_n, P_n')$ because we are assuming that the learner does not have access to the manifold M. On the other hand, adapting techniques from Weed *et al.* (2019), we are able to provide a non-asymptotic lower bound on $W_1(P_n, P_n')$ and $W_1^M(P_n, P_n')$:

Proposition 18. *Suppose that $\mathbb{P}$ is a measure on $\mathbb{R}^D$ such that $\operatorname{supp} \mathbb{P} = M$, where M is a d-dimensional, compact manifold with reach $\tau > 0$ and such that the density of $\mathbb{P}$ with respect to the uniform measure on M is lower bounded by $w > 0$. Suppose that*

$$n > \frac{d \operatorname{vol} M}{4w\omega_d} \left(\frac{\tau}{8}\right)^{-d}.$$

Then, almost surely,

$$W_1(P_n, \mathbb{P}) \geq \frac{1}{32} \left(\frac{d \operatorname{vol} M}{4w\omega_d}\right)^{\frac{1}{d}} n^{-\frac{1}{d}}.$$

If we assume only that

$$n > \left(\frac{d(-\kappa_2)^{\frac{d}{2}} \operatorname{vol} M}{4w\omega_d} e^{d\iota\sqrt{-\kappa_2}}\right) \iota^{-d}$$

then, almost surely,

$$W_1^M(P_n, \mathbb{P}) \geq \frac{1}{32} \left(\frac{d \operatorname{vol} M}{4w\omega_d}\right)^{\frac{1}{d}} (-\kappa_2)^{\frac{1}{2}} e^{\iota\sqrt{-\kappa_2}} n^{-\frac{1}{d}}.$$

An easy proof, based on the techniques (Weed *et al.*, 2019, Proposition 6) can be found in Appendix E. Similarly, we can apply the same proof technique as Lemma 15 to establish the following upper bound:

Proposition 19. *Let $M \subset \mathbb{R}^D$ be a compact manifold with positive reach τ and dimension $d > 2$. Furthermore, suppose that $\mathbb{P}$ is a probability measure on $\mathbb{R}^D$ with $\operatorname{supp} \mathbb{P} \subset M$. Let $X_1, \ldots, X_n, X_1', \ldots, X_n' \sim \mathbb{P}$ be independent with corresponding empirical distributions P_n, P_n'. Then if $\operatorname{diam} M = \Delta$, we have:*

$$\mathbb{E}\left[W_1^M(P_n, \mathbb{P})\right] \leq \mathbb{E}\left[W_1^M(P_n, P_n')\right] \leq C \left(\frac{\operatorname{vol} M}{n\omega_d}\right)^{\frac{1}{d}} \sqrt{\log\left(\frac{n\omega_d \Delta^d}{d \operatorname{vol}_M}\right)}.$$

The full proof is in Appendix E and applies symmetrization and chaining, with an upper bound of Corollary 14. We note, as before, that a similar asymptotic rate is obtained by Weed *et al.* (2019) in a slightly different setting.

We noted above (3) that we required two facts to make our intuition precise. We have just shown that the first holds; we turn now to the second: concentration. To make this rigorous, we need one last technical concept: the T_2-inequality.

Definition 20. *Let μ be a measure on a metric space (M, d). We say that μ satisfies a T_2-inequality with constant c_2 if for all measures $\nu \ll \mu$, we have*

$$W_2(\mu, \nu) \leq \sqrt{2c_2 D(\nu||\mu)}$$

where $D(\nu||\mu) = \mathbb{E}_\mu \left[\log \frac{d\nu}{d\mu}\right]$ is the well-known KL-divergence.

The reason that the T_2 inequality is useful for us is that Bobkov & Götze (1999) tell us that such an inequality implies, and is, by Gozlan *et al.* (2009), equivalent to Lipschitz concentration. We note further that $W_1(P_n, P_n')$ is a Lipschitz function of the dataset and thus concentrates about its mean. The constant in the T_2 inequality depends on the measure μ and upper bounds for specific classes of measures are both well-known and remain an active area of research; for a more complete survey, see Bakry *et al.* (2014). We have the following bound:

Proposition 21. *Let $\mathbb{P}$ be a probability measure on $\mathbb{R}^D$ that has density with respect to the (normalized) volume measure of M, lower bounded by w and upper bounded by W, where M is a d-dimensional manifold with reach $\tau > 0$ and $\operatorname{diam} M = \Delta$. Then we have:*

$$c_2 \leq \frac{2\tau^2}{d-1} \frac{W}{w} \exp\left(d \log 3 + \frac{3d^2\Delta^2}{\tau^2}\right). \tag{3}$$

In order to bound the T_2 constant in our case, we rely on the landmark result of Otto & Villani (2000) that relates c_2 to another functional inequality, the log-Sobolev inequality (Bakry *et al.*, 2014, Chapter 5). There are many ways to control the log-Sobolev constant in various situations, many of which are covered in Bakry *et al.* (2014). We use results from Wang (1997b), which incorporate the intrinsic geometry of the distribution, as our bound. A detailed proof can be found in Appendix B. We note that many other estimates with under slightly different conditions exits, such as that in Wang (1997a), which requires second-order control of the density of the population distribution with respect to the volume measure and the bound in Block *et al.* (2020), which provides control using a measure of nonconvexity. With added assumptions, we can gain much sharper control over c_2; for example, if we assume a positive lower bound on the curvature of the support, we can apply the well-known Bakry-Émery result (Bakry & Émery, 1985) and get dimension-free bounds. As another example, if we may assume stronger contol on the curvature of M beyond that guaranteed by the reach, we can remove the exponential dependence on the reach entirely. For the sake of simplicity and because we already admit an exponential dependence on the intrinsic dimension, we present only the more general bound here. We now provide a non-asymptotic bound on the quality of the estimator $\widetilde{d}_n$.

Theorem 22. *Let $\mathbb{P}$ be a probability measure on $\mathbb{R}^D$ and suppose that $\mathbb{P}$ has a density with respect to the (normalized) volume measure of M lower bounded by w, where M is a d-dimensional manifold with reach $\tau > 0$ such that $d \geq 3$ and $\operatorname{diam} M = \Delta$. Furthermore, suppose that $\mathbb{P}$ satisfies a T_2 inequality with constant c_2. Let $\gamma > 0$ and suppose α, n satisfy*

$$n \geq \max\left[\frac{d \operatorname{vol} M}{4w\omega_d}\left(\frac{8}{\iota}\right)^d, \left(\frac{8c_2}{\Delta^2}\log\frac{1}{\rho}\right)^{\frac{2d}{d-5}}\right]$$
$$\alpha \geq \max\left[\log^{\frac{2}{2\gamma}}\left(\frac{n\omega_d\Delta^d}{d \operatorname{vol} M}\right), (48w)^{\frac{1}{\gamma}}, 3^{\frac{d}{\gamma}}\right]$$
$$\alpha n \geq \frac{d \operatorname{vol} M}{2w\omega_d}\left(\frac{16\pi}{\tau}\right)^d \log\left(\frac{d \operatorname{vol} M}{\rho\omega_d}\left(\frac{16\pi}{\tau}\right)^d\right).$$

Suppose we have $2(\alpha+1)n$ samples drawn independently from $\mathbb{P}$. Then, with probability at least $1-6\rho$, we have

$$\frac{d}{1+3\gamma} \leq \widetilde{d}_n \leq (1+3\gamma)d.$$

If ι is replaced by τ above, we get the same bound with the vanilla estimator d_n replacing $\widetilde{d}_n$.

We note that we have not made every effort to minimize the constants in the statement above, with our emphasis being the dependence of these sample complexity bounds on the relevant geometric quantities. As an immediate consequence of Theorem 22, due to the fact that d is discrete, we can control the probability of error with sufficiently many samples. We may also apply Proposition 21 to replace c_2 with our upper bound in terms of the reach.

Corollary 23. *Suppose we are in the situation of Theorem 22 and that $\mathbb{P}$ has density upper bounded by W with respect to the normalized uniform measure on M. Suppose further that α, n satisfy*

$$\begin{aligned}
n &\geq \max\left[\frac{d\operatorname{vol}M}{4w\omega_d}\left(\frac{8}{\iota}\right)^d, \left(8\frac{2\tau^2}{\Delta^2(d-1)}\frac{W}{w}\exp\left(d\log 3+\frac{3d^2\Delta^2}{\tau^2}\right)\log\frac{1}{\rho}\right)^{\frac{2d}{d-5}}\right] \\
\alpha &\geq \max\left[\log^{2d^2}\left(\frac{n\omega_d\Delta^d}{d\operatorname{vol}M}\right), (48w)^{3d}, 3^{3d^2}\right] \\
\alpha n &\geq \frac{d\operatorname{vol}M}{2w\omega_d}\left(\frac{16\pi}{\tau}\right)^d\log\left(\frac{d\operatorname{vol}M}{\rho\omega_d}\left(\frac{16\pi}{\tau}\right)^d\right).
\end{aligned}$$

Then if we round $\widetilde{d}_n$ to the nearest integer, and denote the resulting estimator by d_n', we have with probability at least $1-6\rho$, $d_n' = d$. Again, replacing ι by τ in the previous display yields the same result with $\widehat{d}_n$ replaced by the vanilla estimator d_n.

Proof. Note that because $d \in \mathbb{N}$, if $\left|\widetilde{d}_n - d\right| \leq \frac{1}{2}$, then rounding $\widehat{d}_n$ to the nearest integer exactly recovers d. Setting $\gamma < \frac{1}{4d}$, and plugging into the result of Theorem 22, along with an application of Proposition 21 to bound c_2, concludes the proof. ∎

While the appearance of ι in Theorem 22 and Corollary 23 may seem minor, it is critical for any practical estimator. While $\alpha n = \Omega\left(\tau^{-d}\right)$, we may take n as small as $\Omega\left(\iota^{-d}\right)$. Thus, using $\widetilde{d}_n$ instead of the naive estimator d_n allows us to leverage the entire data set in estimating the intrinsic distances, even on the small sub-samples. From the proof, it is clear that we want α to be as large as possible; thus if we have a total of N samples, we wish to make n as small as possible. If $\iota \gg \tau$ then we can make n much smaller (scaling like ι^{-d}) than if we were to simply use the Euclidean distance. As a result, on any data set where $\iota \gg \tau$, the sample complexity of $\widetilde{d}_n$ can be much smaller than that of d_n.

There are two parts to the proof of Theorem 22: first, we need to establish that our metric d_G approximates d_M with high probability and thus $\widetilde{d}_n \approx \widehat{d}_n$; second, we need to show that $\widehat{d}_n$ is, indeed, a good estimate of d. The second part follows from Propositions 19 and 18, and concentration; a detailed proof can be found in Appendix C. For the first part of the proof, in order to show that $\widehat{d}_n \approx \widetilde{d}_n$, we demonstrate that $d_M \approx d_G$ in the following result:

Proposition 24. *Let $\mathbb{P}$ be a probability measure on $\mathbb{R}^D$ and suppose that $\operatorname{supp}\mathbb{P} = M$, a geodesically convex, compact manifold of dimension d and reach $\tau > 0$. Suppose that we sample $X_1, \ldots, X_n \sim \mathbb{P}$ independently. Let $\lambda \leq \frac{1}{2}$ and $G = G(X, \tau\lambda)$. If for some $\rho < 1$,*

$$n \geq w_B\left(\frac{\tau\lambda^2}{8}\right)^{-1}\log\frac{N\left(M, d_M, \frac{\tau\lambda^2}{8}\right)}{\rho}$$

where for any $\delta > 0$

$$w_B(\delta) = \inf_{p\in M}\mathbb{P}(B_\delta^M(p))$$

with $B_\delta^M(p)$ the metric ball around p of radius δ. Then, with probability at least $1-\rho$, for all $x, y \in M$,

$$(1-\lambda)\,d_M(x,y) \leq d_G(x,y) \leq (1+\lambda)d_M(x,y).$$

The proof of Proposition 24 follows the general outline of Bernstein *et al.* (2000), but is modified in two key ways: first, we control relevant geometric quantities by τ instead of by the quantities in Bernstein *et al.* (2000); second, we provide a quantitative, nonasymptotic bound on the number of samples needed to get a good approximation with high probability. The details are deferred to Appendix D.

This result may be of interest in its own right as it provides a non-asymptotic version of the results from Tenenbaum *et al.* (2000); Bernstein *et al.* (2000). In particular, if we suppose that $\mathbb{P}$ has a density with respect to the uniform measure on M and this density is bounded below by a constant $w > 0$, then Proposition 24 combined with Proposition 9 tells us that if we have

$$n \gtrsim \frac{\operatorname{vol} M}{w} \left(\tau\lambda^2\right)^{-d} \log\left(\frac{\operatorname{vol} M}{\rho\tau\lambda^2}\right)$$

samples, then we can recover the intrinsic distance of M with distortion λ. We further note that the dependence on τ, λ, d is quite reasonable in Proposition 24. The argument requires the construction of a $\tau\lambda^2$-net on M and it is not difficult to see that one needs a covering at scale proportional to $\tau\lambda$ in order to recover the intrinsic metric from discrete data points. For example, consider Figure 2; were a curve to be added to connect the points at the bottleneck, this would drastically decrease the intrinsic distance between the bottleneck points. In order to determine that the intrinsic distance between these points (without the connector) is actually quite large using the graph metric estimator, we need to set $\varepsilon < \tau$, in which case these points are certainly only connected if there exists a point of distance less than τ to the bottleneck point, which can only occur with high probability if $n = \Omega\left(\tau^{-1}\right)$. We can extend this example to arbitrary dimension d by taking the product of the curve with rS^{d-1} for $r = \Theta(\tau)$; in this case, a similar argument holds and we now need $\Omega\left(\tau^{-d}\right)$ points in order to guarantee with high probability that there exists a point of distance at most τ to one of the bottleneck points. In this way, we see that the τ^{-d} scaling is unavoidable in general. Note that the other estimators of intrinsic dimension mentioned in the introduction, in particular the MLE estimator of Levina & Bickel (2004), implicitly require the accuracy of the kNN distance for their estimation to hold; thus these estimators also suffer from the τ^{-d} sample complexity. Finally, we remark that Kim *et al.* (2019) presents a minimax lower bound for a related hypothesis testing problem and shows that minimax risk is bounded below by a local analogue of the reach raised to a power that depends linearly on the intrinsic dimension.

4 Application of Techniques to GANs

In this section, we note that our techniques are not confined to the realm of dimension estimation and, in fact, readily apply to other problems. As an example, consider the unsupervised learning problem of generative modeling, where we suppose that there are samples $X_1, \dots, X_n \sim \mathbb{P}$ independent and we wish to produce a sample $\widehat{X} \sim \widehat{\mathbb{P}}$ such that $\widehat{\mathbb{P}}$ and $\mathbb{P}$ are close. Statistically, this problem can be expressed by fixing a class of distributions $\mathcal{P}$ and using the data to choose $\widehat{\mu} \in \mathcal{P}$ such that $\widehat{\mu}$ is in some sense close to $\mathbb{P}$. For computational reasons, one wishes $\mathcal{P}$ to contain distributions from which it is computationally efficient to sample; in practice, $\mathcal{P}$ is usually the class of pushforwards of a multi-variate Gaussian distribution by some deep neural network class $\mathcal{G}$. While our statistical results include this setting, they are not restricted and apply for general classes of distributions $\mathcal{P}$.

In order to make the problem more precise, we require some notion of distance between distributions. We use the notion of the Integral Probability Metric (Müller, 1997; Sriperumbudur *et al.*, 2012) associated to a Hölder ball $C_B^\beta(\Omega)$, as defined above. We suppose that $\operatorname{supp} \mathbb{P} \subset \Omega$ and we abbreviate the corresponding IPM distance by $d_{\beta,B}$. Given the empirical distribution P_n, the GAN that we study can be expressed as

$$\widehat{\mu} \in \operatorname*{argmin}_{\mu \in \mathcal{P}} d_{\beta,B}(\mu, P_n) = \operatorname*{argmin}_{\mu \in \mathcal{P}} \sup_{f \in C_B^\beta(\Omega)} \mathbb{E}_\mu[f] - P_n f.$$

In this section, we generalize the results of Schreuder *et al.* (2020). In particular, we derive new estimation rates for a GAN using a Hölder ball as a discriminating class, assuming that the population distribution $\mathbb{P}$ is low-dimensional; like Schreuder *et al.* (2020), we consider the noised and potentially contaminated setting. We have

Theorem 25. *Suppose that $\mathbb{P}$ is a probability measure on $\mathbb{R}^D$ supported on a compact set $\mathcal{S}$ and suppose we have n independent $X_i \sim \mathbb{P}$ with empirical distribution P_n. Let η_i be independent, centred random variables on $\mathbb{R}^D$ such that $\mathbb{E}\left[||\eta_i||^2\right] \leq \sigma^2$. Suppose we observe $\widetilde{X}_i$ such that for at least $(1-\varepsilon)n$ of the $\widetilde{X}_i$, we have $\widetilde{X}_i = X_i + \eta_i$; let the empirical distribution of the $\widetilde{X}_i$ be $\widetilde{P}_n$. Let $\mathcal{P}$ be a known set of distributions and define*

$$\widehat{\mu} \in \operatorname*{argmin}_{\mu \in \mathcal{P}} d_{\beta,B}(\mu, \widetilde{P}_n).$$

Then if there is some C_1, d such that $N(\mathcal{S}, ||\cdot||, \delta) \leq C_1 \varepsilon^{-d}$, we have

$$\mathbb{E}\left[d_{\beta,B}(\widehat{\mu}, \mathbb{P})\right] \leq \inf_{\mu \in \mathcal{P}} d_{\beta,B}(\mu, \mathbb{P}) + B(\sigma + 2\varepsilon) + C\beta B\sqrt{\log n}\left(n^{-\frac{\beta}{d}} \vee n^{-\frac{1}{2}}\right)$$

where C is a constant depending linearly on C_1.

We note that the $\log n$ factor can be easily removed for all cases $\beta \neq \frac{d}{2}$ by paying slightly in order to increase the constants; for the sake of simplicity, we do not bother with this argument here. The proof of Theorem 25 is similar in spirit to that of Schreuder *et al.* (2020), which in turn follows Liang (2018), with details in Appendix E. The key step is in applying the bounds in Lemma 15 to the arguments of Liang (2018).

We compare our result to the corresponding theorem (Schreuder *et al.*, 2020, Theorem 2). In that work, the authors considered a setting where there is a known intrinsic dimension d and the population distribution $\mathbb{P} = g_{\#}\mathcal{U}\left([0,1]^d\right)$, the push-forward by an L-Lipschitz function g of the uniform distribution on a d-dimensional hypercube; in addition, they take $\mathcal{P}$ to be the set of push-forwards of $U\left([0,1]^d\right)$ by functions in some class $\mathcal{F}$, all of whose elements are L-Lipschitz. Their result, (Schreuder *et al.*, 2020, Theorem 2), gives an upper bound of

$$\mathbb{E}\left[d_{\beta,1}(\widehat{\mu}, \mathbb{P})\right] \leq \inf_{\mu \in \mathcal{P}} d_{\beta,1}(\mu, \mathbb{P}) + L(\sigma + 2\varepsilon) + cL\sqrt{d}\left(n^{-\frac{\beta}{d}} \vee n^{-\frac{1}{2}}\right). \tag{4}$$

Note that our result is an improvement in two key respects. First, we do not treat the intrinsic dimension d as known, nor do we force the dimension of the feature space to be the same as the intrinsic dimension. Many of the state-of-the-art GAN architectures on datasets such as ImageNet use a feature space of dimension 128 or 256 (Wu *et al.*, 2019); the best rate that the work of Schreuder *et al.* (2020) can give, then would be $n^{-\frac{1}{128}}$. In our setting, even if the feature space is complex, if the true distribution lies on a much lower dimensional subspace, then it is the true, intrinsic dimension, that determines the rate of estimation. Secondly, note that the upper bound in (4) depends on the Lipschitz constant L; as the function classes used to determine the push-forwards are essentially all deep neural networks in practice, and the Lipschitz constants of such functions are exponential in depth, this can be a very pessimistic upper bound; our result, however, does not depend on this Lipschitz constant, but rather on properties intrinsic to the probability distribution $\mathbb{P}$. This dependence is particularly notable in the noisy regime, where σ, ε do not vanish; the large multiplicative factor of L in this case would then make the bound useless.

We conclude this section by considering the case most often used in practice: the Wasserstein GAN.

Corollary 26. *Suppose we are in the setting of Theorem 25 and $\mathcal{S}$ is contained in a ball of radius R for $R \geq \frac{1}{2}$. Then,*

$$\mathbb{E}\left[W_1(\widehat{\mu}, \mathbb{P})\right] \leq \inf_{\mu \in \mathcal{P}} W_1(\mu, \mathbb{P}) + \sigma + 2R\varepsilon + CR\sqrt{\log n}n^{-\frac{1}{d}}.$$

The proof of the corollary is almost immediate from Theorem 25. With additional assumptions on the tails of the η_i, we can turn our expectation into a high probability statement. In the special case with neither noise nor contamination, i.e. $\sigma = \varepsilon = 0$, we get that the Wasserstein GAN converges in Wasserstein distance at a rate of $n^{-\frac{1}{d}}$, which we believe explains in large part the recent empirical success in modern Wasserstein-GANs.

Acknowledgements

We acknowledge support from the NSF through award DMS-2031883 and from the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning. We acknowledge the support from NSF under award DMS-1953181, NSF Graduate Research Fellowship support under Grant No. 1122374, and support from the MIT-IBM Watson AI Lab.

References

Aamari, Eddie, Kim, Jisu, Chazal, Frédéric, Michel, Bertrand, Rinaldo, Alessandro, Wasserman, Larry, *et al.* 2019. Estimating the reach of a manifold. *Electronic journal of statistics*, **13**(1), 1359–1399.

Amenta, Nina, & Bern, Marshall. 1999. Surface reconstruction by Voronoi filtering. *Discrete & Computational Geometry*, **22**(4), 481–504.

Arjovsky, Martin, Chintala, Soumith, & Bottou, Léon. 2017. Wasserstein generative adversarial networks. *Pages 214–223 of: International conference on machine learning*. PMLR.

Ashlagi, Yair, Gottlieb, Lee-Ad, & Kontorovich, Aryeh. 2021. Functions with average smoothness: structure, algorithms, and learning. *Pages 186–236 of: Conference on Learning Theory*. PMLR.

Assouad, Patrice. 1983. Plongements lipschitziens dans r^n. *Bulletin de la Société Mathématique de France*, **111**, 429–448.

Bakry, Dominique, & Émery, Michel. 1985. Diffusions hypercontractives. *Pages 177–206 of: Seminaire de probabilités XIX 1983/84*. Springer.

Bakry, Dominique, Gentil, Ivan, & Ledoux, Michel. 2014. *Analysis and Geometry of Markov Diffusion Operators*. Springer International Publishing.

Belkin, Mikhail, & Niyogi, Partha. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Pages 585–591 of: Nips*, vol. 14.

Bernstein, Mira, Silva, Vin De, Langford, John C., & Tenenbaum, Joshua B. 2000. *Graph Approximations to Geodesics on Embedded Manifolds.*

Bickel, Peter J, Li, Bo, *et al.* 2007. Local polynomial regression on unknown manifolds. *Pages 177–186 of: Complex datasets and inverse problems*. Institute of Mathematical Statistics.

Block, Adam, Mroueh, Youssef, Rakhlin, Alexander, & Ross, Jerret. 2020. Fast mixing of multi-scale langevin dynamics underthe manifold hypothesis. *arXiv preprint arXiv:2006.11166*.

Bobkov, Sergej G, & Götze, Friedrich. 1999. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, **163**(1), 1–28.

Boissonnat, Jean-Daniel, Lieutier, André, & Wintraecken, Mathijs. 2019. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *Journal of applied and computational topology*, **3**(1), 29–58.

Camastra, Francesco, & Staiano, Antonino. 2016. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, **328**, 26–41.

Camastra, Francesco, & Vinciarelli, Alessandro. 2002. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on pattern analysis and machine intelligence*, **24**(10), 1404–1407.

Chen, Minshuo, Liao, Wenjing, Zha, Hongyuan, & Zhao, Tuo. 2020. Statistical Guarantees of Generative Adversarial Networks for Distribution Estimation.

Cuturi, Marco. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Pages 2292–2300 of: Advances in neural information processing systems.*

Dasgupta, Sanjoy, & Freund, Yoav. 2008. Random projection trees and low dimensional manifolds. *Pages 537–546 of: Proceedings of the fortieth annual ACM symposium on Theory of computing.*

Donoho, David L, & Grimes, Carrie. 2003. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, **100**(10), 5591–5596.

Dudley, Richard Mansfield. 1969. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, **40**(1), 40–50.

Efimov, Kirill, Adamyan, Larisa, & Spokoiny, Vladimir. 2019. Adaptive nonparametric clustering. *IEEE Transactions on Information Theory*, **65**(8), 4875–4892.

Farahmand, Amir Massoud, Szepesvári, Csaba, & Audibert, Jean-Yves. 2007. Manifold-adaptive dimension estimation. *Pages 265–272 of: Proceedings of the 24th international conference on Machine learning.*

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

Fefferman, Charles, Mitter, Sanjoy, & Narayanan, Hariharan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, **29**(4), 983–1049.

Fefferman, Charles, Ivanov, Sergei, Kurylev, Yaroslav, Lassas, Matti, & Narayanan, Hariharan. 2018. Fitting a putative manifold to noisy data. *Pages 688–720 of: Conference On Learning Theory.* PMLR.

Fukunaga, Keinosuke, & Olsen, David R. 1971. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, **100**(2), 176–183.

Gigli, Nicola, & Ledoux, Michel. 2013. From log Sobolev to Talagrand: a quick proof. *Discrete and Continuous Dynamical Systems-Series A*, dcds–2013.

Giné, Evarist, & Nickl, Richard. 2016. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge University Press.

Goodfellow, Ian J, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, & Bengio, Yoshua. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661.*

Gottlieb, Lee-Ad, Kontorovich, Aryeh, & Krauthgamer, Robert. 2016. Adaptive metric dimensionality reduction. *Theoretical Computer Science*, **620**, 105–118.

Gozlan, Nathael, *et al.* 2009. A characterization of dimension free concentration in terms of transportation inequalities. *The Annals of Probability*, **37**(6), 2480–2498.

Grassberger, Peter, & Procaccia, Itamar. 2004. Measuring the strangeness of strange attractors. *Pages 170–189 of: The Theory of Chaotic Attractors.* Springer.

Gray, Alfred. 2004. *Tubes.* Birkhäuser Basel.

Holley, Richard, & Stroock, Daniel W. 1986. Logarithmic Sobolev inequalities and stochastic Ising models.

Hopf, Heinz, & Rinow, Willi. 1931. Über den Begriff der vollständigen differentialgeometrischen Fläche. *Commentarii Mathematici Helvetici*, **3**(1), 209–225.

Kantorovich, Leonid Vitaliyevich, & Rubinshtein, SG. 1958. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, **13**(7), 52–59.

Kégl, Balázs. 2002. Intrinsic dimension estimation using packing numbers. *Pages 681–688 of: NIPS.* Citeseer.

Kim, Jisu, Rinaldo, Alessandro, & Wasserman, Larry. 2019. Minimax Rates for Estimating the Dimension of a Manifold. *Journal of Computational Geometry*, **10**(1).

Kleindessner, Matthäus, & Luxburg, Ulrike. 2015. Dimensionality estimation without distances. *Pages 471–479 of: Artificial Intelligence and Statistics*. PMLR.

Kolmogorov, A. N., & Tikhomirov, V. M. 1993. epsilon-Entropy and epsilon-Capacity of Sets In Functional Spaces. *Pages 86–170 of: Mathematics and Its Applications*. Springer Netherlands.

Koltchinskii, Vladimir I. 2000. Empirical geometry of multivariate data: a deconvolution approach. *Annals of statistics*, 591–629.

Kpotufe, Samory. 2011. k-NN regression adapts to local intrinsic dimension. *Pages 729–737 of: Proceedings of the 24th International Conference on Neural Information Processing Systems*.

Kpotufe, Samory, & Dasgupta, Sanjoy. 2012. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, **78**(5), 1496–1515.

Kpotufe, Samory, & Garg, Vikas K. 2013. Adaptivity to Local Smoothness and Dimension in Kernel Regression. *Pages 3075–3083 of: NIPS*.

LeCun, Yann, & Cortes, Corinna. 2010. MNIST handwritten digit database.

Lee, John M. 2018. *Introduction to Riemannian manifolds*. Springer.

Levina, Elizaveta, & Bickel, Peter. 2004. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, **17**, 777–784.

Liang, Tengyuan. 2018. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*.

Little, Anna V, Jung, Yoon-Mo, & Maggioni, Mauro. 2009. Multiscale estimation of intrinsic dimensionality of data sets. *In: 2009 AAAI Fall Symposium Series*.

Müller, Alfred. 1997. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 429–443.

Nakada, Ryumei, & Imaizumi, Masaaki. 2020. Adaptive Approximation and Generalization of Deep Neural Network with Intrinsic Dimensionality. *Journal of Machine Learning Research*, **21**(174), 1–38.

Narayanan, Hariharan, & Mitter, Sanjoy. 2010. Sample complexity of testing the manifold hypothesis. *Pages 1786–1794 of: Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 2*.

Niles-Weed, Jonathan, & Rigollet, Philippe. 2019. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*.

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

Niyogi, Partha, Smale, Stephen, & Weinberger, Shmuel. 2011. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, **40**(3), 646–663.

Otto, Felix, & Villani, Cédric. 2000. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, **173**(2), 361–400.

Pettis, Karl W, Bailey, Thomas A, Jain, Anil K, & Dubes, Richard C. 1979. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Transactions on pattern analysis and machine intelligence*, 25–37.

Roweis, Sam T, & Saul, Lawrence K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, **290**(5500), 2323–2326.

Schreuder, Nicolas, Brunel, Victor-Emmanuel, & Dalalyan, Arnak. 2020. Statistical guarantees for generative models without domination. *arXiv preprint arXiv:2010.09237*.

Sriperumbudur, Bharath K, Fukumizu, Kenji, Gretton, Arthur, Schölkopf, Bernhard, Lanckriet, Gert RG, *et al.* 2012. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, **6**, 1550–1599.

Steinwart, Ingo, Hush, Don R, Scovel, Clint, *et al.* 2009. Optimal Rates for Regularized Least Squares Regression. *Pages 79–93 of: COLT*.

Tenenbaum, Joshua B, De Silva, Vin, & Langford, John C. 2000. A global geometric framework for nonlinear dimensionality reduction. *science*, **290**(5500), 2319–2323.

Uppal, Ananya, Singh, Shashank, & Póczos, Barnabás. 2019. Nonparametric density estimation & convergence rates for gans under besov ipm losses. *arXiv preprint arXiv:1902.03511*.

van Handel, Ramon. 2014. *Probability in high dimension*. Tech. rept. PRINCETON UNIV NJ.

Villani, Cédric. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

Wang, Feng-Yu. 1997a. Logarithmic Sobolev inequalities on noncompact Riemannian manifolds. *Probability theory and related fields*, **109**(3), 417–424.

Wang, Feng-Yu. 1997b. On estimation of the logarithmic Sobolev constant and gradient estimates of heat semigroups. *Probability theory and related fields*, **108**(1), 87–101.

Weed, Jonathan, Bach, Francis, *et al.* 2019. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, **25**(4A), 2620–2648.

Wu, Yan, Donahue, Jeff, Balduzzi, David, Simonyan, Karen, & Lillicrap, Timothy. 2019. Logan: Latent optimisation for generative adversarial networks. *arXiv preprint arXiv:1912.00953*.

A Proofs from Section 2

Proof of Proposition 12. We apply the method from the classic paper (Kolmogorov & Tikhomirov, 1993), following notation introduced there as applicable. For the sake of simplicity, we assume that β is an integer; the generalization to $\beta \notin \mathbb{N}$ is analogous to that in Kolmogorov & Tikhomirov (1993). Let $\Delta^\beta = \frac{\varepsilon}{2B}$ and let $x_1, \dots, x_s$ be a Δ-connected Δ net on $\mathcal{S}$. For $0 \leq k \leq \beta$ and $1 \leq i \leq s$, define

$$\gamma^k(f) = \left\lfloor \frac{\left|\left|D^k f(x_i)\right|\right|}{\varepsilon_k} \right\rfloor \qquad\qquad \varepsilon_k = \frac{\varepsilon}{\Delta^k}$$

where $||\cdot||$ is the norm on tensors induced by the ambient (Euclidean) metric and D^k is the k^{th} application of the covariant derivative. Let $\gamma(f) = \left(\gamma_i^k(f)\right)_{i,k}$ be the matrix of all $\gamma_i^k(f)$ and let U_γ be the set of all f such that $\gamma(f) = \gamma$. Then the argument in the proof of (Kolmogorov & Tikhomirov, 1993, Theorem XIV) applies *mutatis mutandis* and we note that U_γ are 2ε neighborhoods in the Hölder norm. Thus it suffices to bound the number of possible γ. As in Kolmogorov & Tikhomirov (1993), we note that the number of possible values for γ_1^k is at most $\frac{2B}{\varepsilon_k}$. Given the row $\left(\gamma_i^k\right)_{0 \leq k \leq \beta}$, there are at most $(4e+2)^{\beta+1}$ values for the next row. Thus the total number of possible γ is bounded by

$$\left((4e+2)^{\beta+1}\right)^s \prod_{k=1}^{\beta} \frac{2B}{\varepsilon_k} = (4e+2)^{(\beta+1)s} \prod_{k=1}^{\beta} \frac{2B}{\varepsilon} \left(\frac{\varepsilon}{2B}\right)^{\frac{k}{\beta}} = (4e+2)^{(\beta+1)s} \left(\frac{2B}{\varepsilon}\right)^{\frac{\beta}{2}}.$$

By definition of the covering number and the fact that $\mathcal{S}$ is path-connected, we may take

$$s = N(\mathcal{S}, \Delta) = N\left(\mathcal{S}, \left(\frac{\varepsilon}{2B}\right)^{\frac{1}{\beta}}\right).$$

Taking logarithms and noting that $\log(4e+2) \leq 3$ concludes the proof of the upper bound.

The middle inequality is Lemma 3. For the lower bound, we again follow Kolmogorov & Tikhomirov (1993). Define

$$\varphi(x) = \begin{cases} a \prod_{i=1}^{D} \left(1 - x_i^2\right)^{\frac{\beta}{2}} & \left|\left|x\right|\right|_\infty \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

with a a constant to be set. Choose a 2Δ-separated set $x^1, \ldots, x^s s$ with $\Delta = \left(\frac{\varepsilon}{2B}\right)^{\frac{1}{\beta}}$ and consider the set of functions

$$g_\sigma = \sum_{i=1}^{s} \sigma_i \Delta^\beta \varphi\left(\frac{x - x^i}{\Delta}\right)$$

where $\sigma_i \in \{\pm 1\}$ and σ varies over all possible sets of signs. The results of Kolmogorov & Tikhomirov (1993) guarantee that the g_σ form a 2ε-separated set in $\mathcal{F}$ if a is chosen such that $g_\sigma \in \mathcal{F}$ and there are 2^s such combinations. By definition of packing numbers, we may choose

$$s = D\left(\mathcal{F}, \left(\frac{\varepsilon}{B}\right)^{\frac{1}{\beta}}\right).$$

This concludes the proof of the lower bound. ■

Proof of Proposition 9. We note first that the second statement follows from the first by applying (b) and (c) to Proposition 8 to control the curvature and injectivity radius in terms of the reach. Furthermore, the middle inequality in the last statement follows from Lemma 3. Thus we prove the first two statements.

A volume argument yields the following control:

$$N\left(M, \left|\left|\cdot\right|\right|_g, r\right) \leq \frac{\operatorname{vol} M}{\inf_{p \in M} \operatorname{vol} B_{\frac{\varepsilon}{2}}(p)}$$

where $B_{\frac{\varepsilon}{2}}(p)$ is the ball around p of radius $\frac{\varepsilon}{2}$ with respect to the metric g. Thus it suffices to lower bound the volume of such a ball. Because $\varepsilon < \iota$, we may apply the Bishop-Gromov comparison theorem (Gray, 2004, Theorem 3.17) to get that

$$\operatorname{vol} B_\varepsilon(p) \geq \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\varepsilon \left(\frac{\sin\left(t\sqrt{\kappa_1}\right)}{\sqrt{\kappa_1}}\right)^{d-1} dt = \omega_d \int_0^\varepsilon \left(\kappa_1^{-\frac{1}{2}} \sin\left(t\sqrt{\kappa_1}\right)\right)^{d-1} dt$$

where κ_1 is an upper bound on the sectional curvature. We note that for $t \leq \frac{\pi}{2\sqrt{\kappa_1}}$, we have $\sin\left(t\sqrt{\kappa_1}\right) \geq \frac{2}{\pi} t\sqrt{\kappa_1}$ and thus

$$\operatorname{vol} B_\varepsilon(p) \geq \omega_d \int_0^\varepsilon \left(\frac{2}{\pi} t\right)^{d-1} dt = \frac{\omega_d}{d}\left(\frac{2}{\pi}\right)^{d-1} \varepsilon^d.$$

The upper bound follows from control on the sectional curvature by τ, appearing in (Aamari *et al.*, 2019, Proposition A.1), which, in turn, is an easy consequence of applying the Gauss formula to (a) of Proposition 8.

We lower bound the packing number through an analogous argument as the upper bound for the covering number, this time with an upper bound on the volume of a ball of radius ε, again from (Gray, 2004, Theorem 3.17), but this time using a lower bound on the sectional curvature. In particular, we have for $\varepsilon < \iota$,

$$\operatorname{vol} B_\varepsilon(p) \leq \omega_d \int_0^\varepsilon \left(\frac{\sin\left(t\sqrt{\kappa_2}\right)}{\sqrt{\kappa_2}}\right)^{d-1} dt = \omega_d \int_0^\varepsilon \left(\frac{\sinh\left(t\sqrt{-\kappa_2}\right)}{\sqrt{-\kappa_2}}\right)^{d-1} dt$$

where κ_2 is a lower bound on the sectional curvature. Note that for $t \leq \frac{1}{\sqrt{-\kappa_2}}$, we have

$$\frac{\sinh\left(t\sqrt{-\kappa_2}\right)}{\sqrt{-\kappa_2}} \leq \cosh(2)t \leq 4t.$$

Thus,

$$\operatorname{vol} B_\varepsilon(p) \leq \omega_d \int_0^\varepsilon (4t)^{d-1} dt = \frac{\omega_d}{d} 4^d \varepsilon^d.$$

The volume argument tells us that

$$N\left(M, ||\cdot||_g, r\right) \geq \frac{\operatorname{vol} M}{\sup_{p \in M} \operatorname{vol} B_r(p)}$$

and the result follows.

If we wish to extend the range of ε, we pay with a constant exponential exponential in d, reflecting the growth in volume of balls in negatively curved spaces. In particular, we can apply the same argument and note that as $\frac{\sinh(x)}{x}$ is increasing, we have

$$\frac{\sinh\left(t\sqrt{-\kappa_2}\right)}{\sqrt{-\kappa_2}} \leq \frac{\sinh(\iota\sqrt{-\kappa_2})}{\iota\sqrt{-\kappa_2}} t \leq \frac{e^{\iota\sqrt{-\kappa_2}}}{\iota\sqrt{-\kappa_2}} t$$

for all $t < \iota$. Thus for all $\varepsilon < \iota$. We have:

$$N(M, ||\cdot||_g, \varepsilon) \geq \frac{\operatorname{vol} M}{\omega_d} d\iota^d (-\kappa_2)^{\frac{d}{2}} e^{-d\iota\sqrt{-\kappa_2}} \varepsilon^{-d}$$

as desired. ■

Proof of Corollary 10. Let $B_\varepsilon^{\mathbb{R}^D}(p)$ be the set of points in $\mathbb{R}^D$ with Euclidean distance to p less than ε and let $B_\varepsilon^M(p)$ be the set of points in M with intrinsic (geodesic) distance to p less than ε. Then, if $\varepsilon \leq 2\tau$, combining the fact that straight lines are geodesics in $\mathbb{R}^D$ and (d) from Proposition 8 gives

$$B_\varepsilon^M(p) \subset B_\varepsilon^{\mathbb{R}^D}(p) \cap M \subset B^M_{2\tau \arcsin\left(\frac{\varepsilon}{2\tau}\right)}(p)$$

In particular, this implies

$$N\left(M, d_M, 2\tau \arcsin\left(\frac{\varepsilon}{2\tau}\right)\right) \leq N(M, ||\cdot||, \varepsilon) \leq N(M, d_M, \varepsilon)$$
$$D\left(M, d_M, 2\tau \arcsin\left(\frac{\varepsilon}{2\tau}\right)\right) \leq D(M, ||\cdot||, \varepsilon) \leq D(M, d_M, \varepsilon)$$

whenever $\varepsilon \leq 2\tau$. Thus, applying Proposition 9, we have

$$N(M, ||\cdot||, \varepsilon) \leq N(M, d_M, \varepsilon) \leq \frac{\operatorname{vol} M}{\omega_d} d \left(\frac{\pi}{2}\right)^d \varepsilon^{-d}$$

and similarly,

$$D(M, ||\cdot||, 2\varepsilon) \geq D\left(M, d_M, 2\tau \arcsin\left(\frac{\varepsilon}{\tau}\right)\right) \geq \frac{\operatorname{vol} M}{\omega_d} d16^{-d} \varepsilon^{-d}$$

using the fact that $\arcsin(x) \leq 2x$ for $x \geq 0$. The result follows. ■

B Proof of Proposition 21

As stated in the body, we bound the T_2 constant c_2 by the log-Sobolev constant of the same measure. We thus first define a log-Sobolev inequality:

Definition 27. *Let μ be a measure on M. We say that μ satisfies a log-Sobolev inequality with constant c_{LS} if for all real valued, differentiable functions with mean 0 $f : M \to \mathbb{R}$, we have:*

$$\int_M f^2 \log(f^2) d\mu \leq c_{LS} \int_M ||\nabla f||^2 d\mu$$

where ∇ is the Levi-Civita connection and $||\cdot||$ is the norm with respect to the Riemannian metric.

While in the main body we cited Otto & Villani (2000) for the Otto-Villani theorem, we actually need a slight strengthening of this result. For technical reasons, Otto & Villani (2000) required the density of μ to have two derivatives; more recent works have eliminated that assumption. We have:

Theorem 28 (Theorem 5.2 from Gigli & Ledoux (2013))**.** *Suppose that μ satisfies the log-Sobolev inequality with constant c_{LS}. Then μ satisfies the T_2 inequality with constant $c_2 \leq 2c_{LS}$.*

We now recall the key estimate from Wang (1997b) that controls the log-Sobolev constant for the uniform measure on a compact manifold M[2]:

Theorem 29 (Theorem 3.3 from Wang (1997b))**.** *Let M be a compact, d-dimensional manifold with diameter Δ. Suppose that $Ric_M \succeq -K$ for some $K \in \mathbb{R}$. Let μ be the uniform measure on M (i.e., the volume measure normalized so that $\mu(M) = 1$). Then μ satisfies a log-Sobolev inequality with*

$$c_{LS} \leq \left(\frac{d+2}{d}\right)^d \frac{e^{2K(d+1)\Delta^2} - 1}{K} e^{1+d\Delta^2 K_+}.$$

We are now ready to complete the proof.

Proof of Proposition 21. By the Holly-Stroock perturbation theorem (Holley & Stroock, 1986), we know that if μ is the uniform measure on M normalized such that $\mu(M) = 1$, and μ satisfies a log-Sobolev inequality with constant c'_{LS} then $\mathbb{P}$ satisfies a log-Sobolev constant with $c_{LS} \leq \frac{W}{w} c'_{LS}$. By (a) from Proposition 8, we have that the sectional curvatures of M are all bounded below by $-\frac{2}{\tau^2}$ and thus $Ric_M \succeq -(d-1)\frac{2}{\tau^2}$ (for the relationship between the Ricci tensor and the sectional curvatures, see Lee (2018)). Noting that $\frac{d+2}{d} \leq 3$ and plugging into the results of Theorem 29, we get that

$$c'_{LS} \leq \frac{2\tau^2}{d-1} \exp\left(d \log 3 + \frac{3\Delta^2 d^2}{\tau^2}\right).$$

Combining this with the Holly-Stroock result and Theorem 28 concludes the proof. ■

C Proof of Theorem 22

We first prove the following lemma on the concentration of $W_1(P_n, P'_n)$.

Lemma 30. *Suppose that $\mathbb{P}$ is a probability measure on (T, d) and that it satisfies a $T_2(c_2)$-inequality. Let $X_1, \ldots, X_n, X'_1, \ldots, X'_n$ denote independent samples with corresponding empirical distributions P_n, P'_n. Then the following inequalities hold:*

$$\mathbb{P}\left(|W_1(P_n, P'_n) - \mathbb{E}\left[W_1(P_n, P'_n)\right]| \geq t\right) \leq 2e^{-\frac{nt^2}{8c_2}}$$
$$\mathbb{P}\left(|W_1(P_n, P'_n) - \mathbb{E}\left[W_1(P_n, P'_n)\right]| \leq t\right) \leq 2e^{-\frac{nt^2}{8c_2}}.$$

Proof. We note that by Gozlan *et al.* (2009), in particular the form of the main theorem stated in (van Handel, 2014, Theorem 4.31), it suffices to show that, as a function of the data, $W_1(P_n, P'_n)$ is $\frac{2}{\sqrt{n}}$-Lipschitz. Note that by symmetry, it suffices to show a one-sided inequality. By the triangle inequality,

$$W_1(P_n, P'_n) \leq W_1(P_n, \mu) + W_1(P'_n, \mu)$$

for any measure μ and thus it suffices to show that $W_1(P_n, \mu)$ is $\frac{1}{\sqrt{n}}$-Lipschitz in the X_i. By (van Handel, 2014, Lemma 4.34), there exists a bijection between the set of couplings between P_n and μ and the set of

[2] We remark that some works, including Wang (1997b), define the log-Sobolev constant to be the inverse of our c_{LS}. We translate their theorem into our terms by taking the reciprocol.

ordered n-tuples of measures $\mu_1, \ldots, \mu_n$ such that $\mu = \frac{1}{n}\sum_i \mu_i$. Thus we see that if $X, \widetilde{X}$ are two data sets, then

$$\begin{aligned}
W_1(P_n, \mu) - W_1(\widetilde{P}_n, \mu) &\le \sup_{\frac{1}{n}\sum_{i=1}^n \mu_i = \mu} \left[\frac{1}{n} \sum_{i=1}^n \int \left(d(X_i, y) - d(\widetilde{X}_i, y) \right) d\mu_i(y) \right] \\
&\le \sup_{\frac{1}{n}\sum_{i=1}^n \mu_i = \mu} \left[\frac{1}{n} \sum_{i=1}^n \int d(X_i, \widetilde{X}_i) d\mu_i(y) \right] \\
&= \frac{1}{n} \sum d(X_i, \widetilde{X}_i) \\
&\le \frac{1}{n} \sqrt{n \sum_{i=1}^n d(X_i, \widetilde{X}_i)^2} \le \frac{1}{\sqrt{n}} d^{\otimes n}(X, \widetilde{X}).
\end{aligned}$$

The identical argument applies to W_1^M. ■

We are now ready to show that $\widehat{d}_n$ is a good estimator of d.

Proposition 31. *Suppose we are in the situation of Theorem 22 and we have*

$$\begin{aligned}
n &\ge \max\left(\frac{d \operatorname{vol} M}{4w\omega_d} \left(\frac{\iota}{8}\right)^{-d}, \left(\frac{8c_2}{\Delta^2} \log \frac{1}{\rho} \right)^{\frac{d}{2d-5}} \right) \\
\alpha &\ge \max\left(\log^{\frac{d}{2\gamma}} \left(\frac{n\omega_d \Delta^d}{d \operatorname{vol} M} \right), (Cw)^{\frac{1}{\gamma}} \right)
\end{aligned}$$

Then with probability at least $1 - 4\rho$, we have

$$\frac{d}{1+3\gamma} \le \widehat{d}_n \le (1+3\gamma)d.$$

Proof. By Proposition 19 and Lemma 30, we have that with probability at least $1 - e^{-\frac{nt^2}{8c_2}}$, we have

$$W_1^M(P_n, P_n') \le C \left(\frac{\operatorname{vol} M}{n\omega_d} \right)^{\frac{1}{d}} \sqrt{\log\left(\frac{n\omega_d}{d \operatorname{vol}_M} \right)} + t.$$

By Proposition 18 and Lemma 30 and the left hand side of Proposition 19, we have that with probability at least $1 - e^{-\frac{\alpha n t^2}{8c_2}}$,

$$W_1^M(P_{\alpha n}, P_{\alpha n}') \ge \frac{1}{32} \left(\frac{d \operatorname{vol} M}{4w\omega_d} \right)^{\frac{1}{d}} (\alpha n)^{-\frac{1}{d}} - t$$

all under the assumption that

$$n > \frac{d \operatorname{vol} M}{4w\omega_d} \left(\frac{\iota}{8} \right)^{-d}.$$

Setting $t = \Delta(\alpha n)^{-\frac{5}{4d}}$, we see that, as $\alpha > 1$, with probability at least $1 - 2e^{-\frac{nt^2}{8c_2}}$, we simultaneously have

$$\begin{aligned}
W_1^M(P_n, P_n') &\le C \left(\frac{\operatorname{vol} M}{n\omega_d} \right)^{\frac{1}{d}} \sqrt{\log\left(\frac{n\omega_d \Delta^d}{d \operatorname{vol}_M} \right)} \\
W_1^M(P_{\alpha n}, P_{\alpha n}') &\ge \frac{1}{64} \left(\frac{d \operatorname{vol} M}{4w\omega_d} \right)^{\frac{1}{d}} (\alpha n)^{-\frac{1}{d}}.
\end{aligned}$$

Thus, in particular,

$$\frac{W_1^M(P_n, P_n')}{W_1^M(P_{\alpha n}, P_{\alpha n}')} \le \frac{C \ \left(\frac{\operatorname{vol} M}{n\omega_d} \right)^{\frac{1}{d}} \sqrt{\log\left(\frac{n\omega_d}{d \operatorname{vol} M} \right)}}{\frac{1}{64} \left(\frac{d \operatorname{vol} M}{4w\omega_d} \right)^{\frac{1}{d}} (\alpha n)^{-\frac{1}{d}}} \le C w^{\frac{1}{d}} \alpha^{\frac{1}{d}} \sqrt{\log\left(\frac{n\omega_d \Delta^d}{d \operatorname{vol} M} \right)}$$

Thus we see that

$$\begin{aligned}\widehat{d}_n &= \frac{\log \alpha}{\log \frac{W_1(P_n, P_n')}{W_1(P_{\alpha n}, P_{\alpha n}')}} \\ &\geq \frac{\log \alpha}{\frac{1}{d}\log \alpha + + \frac{1}{d}\log w + \frac{1}{2}\log\log\left(\frac{n\omega_d \Delta^d}{d \operatorname{vol} M}\right)} \\ &= \frac{d}{1 + \frac{\log(Cw) + \frac{d}{2}\log\log\left(\frac{n\omega_d\Delta^d}{d\operatorname{vol} M}\right)}{\log \alpha}}\end{aligned}$$

Now, if

$$\begin{aligned} n &\geq \max\left(\frac{d \operatorname{vol} M}{4w\omega_d}\left(\frac{\tau}{8}\right)^{-d}, \left(\frac{8c_2^2}{\Delta^2}\log\frac{1}{\rho}\right)^{\frac{d}{2d-5}}\right) \\ \alpha &\geq \max\left(\log^{\frac{d}{2\gamma}}\left(\frac{n\omega_d\Delta^d}{d\operatorname{vol} M}\right), (Cw)^{\frac{1}{\gamma}}\right)\end{aligned}$$

Then with probability at least $1 - 2\rho$,

$$\widehat{d}_n \geq \frac{d}{1 + 2\gamma}.$$

An identical proof holds for the other side of the bound and thus the result holds. ■

We are now ready to prove the main theorem using Proposition 31 and Proposition 24.

Proof of Theorem 22. Note first that

$$w\left(\frac{\iota\lambda^2}{8}\right) \geq \frac{w\omega_d}{d}\left(\frac{\pi}{2}\right)^{-d}\left(\frac{\iota\lambda^2}{8}\right)^d \tag{5}$$

$$N\left(M, d_M, \frac{\iota\lambda^2}{8}\right) \leq \frac{\operatorname{vol} M}{\omega_d} d \left(\frac{\pi}{2}\right)^d \left(\frac{\iota\lambda^2}{8}\right)^{-d} \tag{6}$$

by Proposition 9. Setting $\lambda = \frac{1}{2}$, we note that by Proposition 24, if the total number of samples

$$2(\alpha + 1)n \geq \left(\frac{w\omega_d}{d}\left(\frac{\pi}{2}\right)^{-d}\left(\frac{\iota\lambda^2}{8}\right)^d\right)^{-1}\log\left(\frac{\operatorname{vol} M}{\rho\omega_d} d \left(\frac{\tau}{16\pi}\right)^{-d}\right)$$

then with probability at least $1 - \rho$, we have

$$\frac{1}{2}d_M(p,q) \leq d_G(p,q) \leq \frac{3}{2}d_M(p,q)$$

for all $p, q \in M$. Thus by the proof of Proposition 31 above,

$$\frac{W_1^M(P_n, P_n')}{W_1^M(P_{\alpha n}, P_{\alpha n}')} \leq \frac{1+\lambda}{1-\lambda} C w^{\frac{1}{d}} \alpha^{\frac{1}{d}} \sqrt{\log\left(\frac{n\omega_d\Delta^d}{d\operatorname{vol} M}\right)}.$$

Thus as long as $\alpha \geq \left(\frac{1+\lambda}{1-\lambda}\right)^{\frac{d}{\gamma}} = 3^{\frac{d}{\gamma}}$, then we have with probability at least $1 - 3\rho$,

$$\widetilde{d}_n \geq \frac{d}{1 + 3\gamma}.$$

A similar computation holds for the other bound.

To prove the result for d_n, note that if we replace the ιs by τ in (5) and (6), then the result still holds by the second part of Proposition 9. Then the identical arguments apply, *mutatis mutandis*, after skipping the step of approximating d_M by d_G. ■

D Metric Estimation Proofs

In order to state our result, we need to consider the minimal amount of probability mass that $\mathbb{P}$ puts on any intrinsic ball of a certain radius in M. To formalize this notion, we define, for $\delta > 0$,

$$w_B(\delta) = \inf_{p \in M} \mathbb{P}\left(B_\delta^M(p)\right).$$

We need a few lemmata:

Lemma 32. *Fix $\varepsilon > 0$ and a set of $x_i \in M$ and form $G(x, \varepsilon)$. If the set of x_i form a δ-net for M such that $\delta \leq \frac{\varepsilon}{4}$, then for all $x, y \in M$,*

$$d_G(x,y) \leq \left(1 + \frac{4\delta}{\varepsilon}\right) d_M(x,y).$$

Proof. This is a combination of (Bernstein *et al.*, 2000, Proposition 1) and (Bernstein *et al.*, 2000, Theorem 2). ■

Lemma 33. *Let $0 < \lambda < 1$ and let $x, y \in M$ such that $||x - y|| \leq 2\tau\lambda(1-\lambda)$. Then*

$$(1-\lambda) d_M(x,y) \leq ||x-y|| \leq d_M(x,y).$$

Proof. Note that $2\tau\lambda(1-\lambda) \leq \frac{\tau}{2}$ so we are in the situation of Proposition 8 (e). Let $\ell = d_M(x,y)$. Rearranging the bound in Proposition 8 (e) yields

$$\ell\left(1 - \frac{\ell}{2\tau}\right) \leq ||x-y|| \leq \ell.$$

Thus it suffices to show that

$$\frac{\ell}{2\tau} \leq \lambda.$$

Again applying Proposition 8, we see that

$$\ell \leq \tau\left(1 - \sqrt{1 - \frac{2\,||x-y||}{\tau}}\right).$$

Rearranging and plugging in $||x-y|| \leq 2\tau\lambda(1-\lambda)$ concludes the proof. ■

The next lemma is a variant of (Niyogi *et al.*, 2008, Lemma 5.1).

Lemma 34. *Let $w_B(\delta)$ be as in Proposition 24 and let $N(M,\delta)$ be the covering number of M at scale δ. If we sample $n \geq w\left(\frac{\delta}{2}\right)^{-1} \log \frac{N\left(M, \frac{\delta}{2}\right)}{\rho}$ points independently from $\mathbb{P}$, then with probability at least $1 - \rho$, the points form a δ-net of M.*

Proof. Let $y_1, \ldots, y_N$ be a minimal $\frac{\delta}{2}$-net of M. For each y_i the probability that x_i is not in $B_{\frac{\delta}{2}}(y_i)$ is bounded by $1 - w_B\left(\frac{\delta}{2}\right)$ by definition. By independence, we have

$$\mathbb{P}\Big(\forall i\ x_j \notin B_{\frac{\delta}{2}}(y_i)\Big) \leq \Big(1-)Bw\left(\frac{\delta}{2}\right)\Big)^n \leq e^{-nw_B\left(\frac{\delta}{2}\right)}.$$

By a union bound, we have

$$\mathbb{P}\Big(\exists i \text{ such that } \forall j\ x_j \notin B_{\frac{\delta}{2}}(y_i)\Big) \leq N\left(M, \frac{\delta}{2}\right) e^{-nw_B\left(\frac{\delta}{2}\right)}. \tag{7}$$

If n satisfies the bound in the statement then the right hand side (7) is controlled by ρ. ■

Note that for any measure $\mathbb{P}$, a simple union bound tells us that $w_B(\delta) \leq N(M, \delta)^{-1}$ and that equality, up to a constant, is achieved for the uniform measure. This is within a log factor of the obvious lower bound given by the covering number on the number of points required to have a δ-net on M.

With these lemmata, we are ready to conclude the proof:

Proof of Proposition 24. Let $\varepsilon = \tau\lambda \leq 2\tau\lambda(1-\lambda)$ by $\lambda \leq \frac{1}{2}$. Let $\delta = \frac{\lambda\varepsilon}{4} = \frac{\tau\lambda^2}{4}$. By Lemma 34, with high probability, the x_i form a δ-net on M; thus for the rest of the proof, we fix a set of x_i such that this condition holds. Now we may apply Lemma 32 to yield the upper bound $d_G(x, y) \leq (1+\lambda) d_M(x, y)$.

For the lower bound, for any points $p, q \in M$ there are points x_{j_0}, x_{j_m} such that $d_M(p, x_{j_0}) \leq \delta$ and $d_M(q, x_{j_m}) \leq \delta$ by the fact that the x_i form a δ-net. Let $x_{j_1}, \dots, x_{j_{m-1}}$ be a geodesic in G between x_{j_0} and x_{j_m}. By Lemma 33 and the fact that edges only exist for small weights, we have

$$\begin{aligned}
d_M(p,q) &\leq d_M\left(p, x_{j_0}\right) + d_M\left(x_{j_m}, q\right) + \sum_{i=1}^m d_M\left(x_{j_{i-1}}, x_{j_i}\right) \\
&\leq (1-\lambda)^{-1}\left(||p - x_{j_0}|| + ||x_{j_m} - q|| + \sum_{i=1}^m \left|\left|x_{j_{i-1}} - x_{j_i}\right|\right| \right) \\
&= (1-\lambda)^{-1} d_G(p,q).
\end{aligned}$$

Rearranging concludes the proof. ■

E Miscellany

Proof of Lemma 15. By symmetrization and chaining, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i) - f(X_i')\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right] &\leq 2\inf_{\delta>0}\left[8\delta + \frac{8\sqrt{2}}{\sqrt{n}}\int_\delta^B \sqrt{\log N(\mathcal{F}, ||\cdot||_\infty, \varepsilon)} d\varepsilon\right] \\
&\leq 2B\inf_{\delta>0}\left[8\delta + \frac{8\sqrt{2}}{\sqrt{n}}\int_\delta^1 \sqrt{\log N\left(\mathcal{F}, ||\cdot||_\infty, \frac{\varepsilon}{2R}\right)} d\varepsilon\right] \\
&\leq 2B\inf_{\delta>0}\left[8\delta + \frac{8\sqrt{2}}{\sqrt{n}}\int_\delta^1 \sqrt{3\beta^2 \log\frac{1}{\varepsilon} N(\mathcal{S}, ||\cdot||, \varepsilon)} d\varepsilon\right]
\end{aligned}$$

where the last step follows from Proposition 12. The first statement follows from noting that $\sqrt{\log\frac{1}{\varepsilon}}$ is decreasing in ε, and thus allowing it to be pulled from the integral. If $\beta > \frac{d}{2}$, the second statement follows from plugging in $\delta = 0$ and recovering a rate of $n^{-\frac{1}{2}}$. If $\beta < \frac{d}{2}$, then the second statement follows from plugging in $\delta = n^{-\frac{\beta}{d}}$. ■

Proof of Proposition 18. We follow the proof of (Weed *et al.*, 2019, Proposition 6) and use their notation. In particular, let

$$N_\varepsilon\left(\mathbb{P}, \frac{1}{2}\right) = \inf\left\{N(S, d_M, \varepsilon) | S \subset M \text{ and } \mathbb{P}(S) \geq \frac{1}{2}\right\}.$$

Applying a volume argument in the identical fashion to Proposition 9, but lower bounding the probability of a ball of radius ε by w multiplied by the volume of said small ball, we get that

$$N_\varepsilon\left(\mathbb{P}, \frac{1}{2}\right) \geq \frac{\operatorname{vol} M}{2w\omega_d} d8^{-d}\varepsilon^{-d}$$

if $\varepsilon \leq \tau$. Let

$$\varepsilon = \left(\frac{\operatorname{vol} M}{4w\omega_d} d8^{-d}\right)^{\frac{1}{d}} n^{-\frac{1}{d}}$$

and assume that

$$n > \frac{\operatorname{vol} M}{4 w \omega_d} d 8^{-d} \left(\tau\right)^{-d}$$

Let

$$S = \bigcup_{1 \leq i \leq n} B^M_{\frac{\varepsilon}{2}}(X_i).$$

Then because

$$N_\varepsilon\left(\mathbb{P}, \frac{1}{2}\right) > n$$

by our choice of ε, we have that $\mathbb{P}(S) < \frac{1}{2}$. Thus if $X \sim \mathbb{P}$ then we have with probability at least $\frac{1}{2}$, $d_M(X, \{X_1, \dots, X_n\}) \geq \frac{\varepsilon}{2}$. Thus the Wasserstein distance between $\mathbb{P}$ and P_n is at least $\frac{\varepsilon}{4}$. The first result follows. We may apply the identical argument, instead using intrinsic covering numbers and the bound in Proposition 9 to recover the second statement. ■

Proof of Proposition 19. By Kantorovich-Rubenstein duality and Jensen's inequality, we have

$$\mathbb{E}\left[W_1^M(P_n, \mathbb{P})\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}\left[f(X_i)\right]\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) - f(X_i')\right] = \mathbb{E}\left[W_1^M(P_n, P_n')\right]$$

where $\mathcal{F}$ is the class of functions on M that are 1-Lipschitz with respect to d_M. Note that, by translation invariance, we may take the radius of the Hölder ball $\mathcal{F}$ to be Δ. By symmetrization and chaining,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) - f(X_i')\right] &\leq 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i)\right] \leq 2 \inf_{\delta > 0}\left[8\delta + \frac{8\sqrt{2}}{\sqrt{n}} \int_\delta^\Delta \sqrt{\log N(\mathcal{F}, ||\cdot||_\infty, \varepsilon)} d\varepsilon\right] \\
&\leq \inf_{\delta > 0}\left[8\delta + \frac{8\sqrt{2}}{\sqrt{n}} \int_\delta^\Delta \sqrt{3 \log\left(\frac{2\Delta}{\varepsilon}\right) \frac{d \operatorname{vol} M}{\omega_d} \left(\frac{\pi}{2}\right)^d \left(\frac{2}{\varepsilon}\right)^{\frac{d}{2}}} d\varepsilon\right] \\
&\leq 2\Delta \inf_{\delta > 0}\left[8\delta + \frac{8\sqrt{6}}{\sqrt{n}} \sqrt{\frac{d \operatorname{vol} M}{\omega_d}} \left(\frac{\pi}{2}\right)^{\frac{d}{2}} \sqrt{\log \frac{1}{\delta}} \int_\delta^1 \left(\frac{\Delta}{\varepsilon}\right)^{-\frac{d}{2}} d\varepsilon\right]
\end{aligned}$$

where the last step comes from Corollary 14 and noting that after recentering, $\mathcal{F}$ contains functions f such that $||f||_{L^\infty(M)} \leq \Delta$ and $||\nabla f||_{L^\infty(M)} \leq 1$. Setting

$$\delta = \frac{\pi}{2}\left(\frac{d \operatorname{vol} M}{n \omega_d \Delta^d}\right)^{\frac{1}{d}}$$

gives

$$\mathbb{E}\left[W_1^M(P_n, P_n')\right] \leq C \left(\frac{\operatorname{vol} M}{n \omega_d}\right)^{\frac{1}{d}} \sqrt{\log\left(\frac{n \omega_d \Delta^d}{d \operatorname{vol}_M}\right)}$$

for some $C \leq 48$, which concludes the proof. ■

Proof of Theorem 25. By bounding the supremum of sums by the sum of suprema and the construction of $\widehat{\mu}$,

$$\begin{aligned}
d_{\beta,B}(\widehat{\mu}, \mathbb{P}) &\leq d_{\beta,B}(\widehat{\mu}, \widetilde{P}_n) + d_{\beta,B}(\widetilde{P}_n, \mathbb{P}) \leq \inf_{\mu \in \mathcal{P}} d_{\beta,B}(\mu, \widetilde{P}_n) + d_{\beta,B}(\widetilde{P}_n, \mathbb{P}) \\
&\leq \inf_{\mu \in \mathcal{P}} d_{\beta,B}(\mu, \mathbb{P}) + 2 d_{\beta,B}(\widetilde{P}_n, \mathbb{P}) \\
&\leq \inf_{\mu \in \mathcal{P}} d_{\beta,B}(\mu, \mathbb{P}) + 2 d_{\beta,B}(\widetilde{P}_n, P_n) + 2 d_{\beta,B}(P_n, \mathbb{P}).
\end{aligned}$$

Taking expectations and applying Lemma 15 bounds the last term. The middle term can be bounded as follows:

$$\begin{aligned}
d_{\beta,B}(\widetilde{P}_n, P_n) &= \sup_{f \in C_B^\beta(\Omega)} \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(\widetilde{X}_i) \leq \sup_{f \in C_B^\beta(\Omega)} \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X_i + \eta_i) + 2B\varepsilon \\
&\leq \sup_{f \in C_B^\beta(\Omega)} \frac{1}{n} \sum_{i=1}^{n} B \, ||\eta_i|| + 2B\varepsilon
\end{aligned}$$

where the first inequality follows from the fact that if $f \in C_B^\beta(\Omega)$ then $||f||_\infty \leq B$ and the contamination is at most ε. The second inequality follows from the fact that f is B-Lipschitz. Taking expectations and applying Jensen's inequality concludes the proof. ■

Proof of Corollary 26. Applying Kantorovich-Rubenstein duality, the proof follows immediately from that of Theorem 25 by setting $\beta = 1$, with the caveat that we need to bound B and the Lipschitz constant separately. The Lipschitz constant is bounded by 1 by Kantorovich duality. The class is translation invariant, and so $|||f||_\infty - \mathbb{E}[f]| \leq 2R$ by the fact that the Euclidean diameter of $\mathcal{S}$ is bounded by $2R$. The result follows. ■

Lemma 35. *Let X be distributed uniformly on a centred (ℓ^2) ball in $\mathbb{R}^d$ of radius R. Then,*

$$\mathbb{E}\left[\log \frac{R}{||X||}\right] = \frac{1}{d}.$$

Proof. Note that by scaling it suffices to prove the case $R = 1$. By changing to polar coordinates,

$$\begin{aligned}
\mathbb{E}\left[\log \frac{1}{||X||}\right] &= \frac{\int_{S^1} \int_0^1 \left(\log \frac{1}{r}\right) r^{d-1} dr d\theta}{\int_{S^1} \int_0^1 r^{d-1} dr d\theta} \\
&= -d \int_0^1 (\log r) \, r^{d-1} dr.
\end{aligned}$$

Substituting $u = \log r$ and applying integration by parts then gives

$$-d \int_0^1 (\log r) \, r^{d-1} dr = \left[\frac{r^d}{d} - r^d \log r\right] \Bigg|_{r=0}^{r=1} = \frac{1}{d}$$

as desired. ■