Geometry-Aware Optimal Transport: Fast Intrinsic Dimension and Wasserstein Distance Estimation

Ferdinand Genans [1] **Olivier Wintenberger** [1 2]

Abstract

Solving large scale Optimal Transport (OT) in machine learning typically relies on sampling measures to obtain a tractable discrete problem. While the discrete solver's accuracy is controllable, the rate of convergence of the discretization error is governed by the intrinsic dimension of our data. Therefore, the true bottleneck is the knowledge and control of the sampling error. In this work, we tackle this issue by introducing novel estimators for both sampling error and intrinsic dimension. The key finding is a simple, tuning-free estimator of $\mathrm{OT}_c(\rho, \widehat{\rho})$ that utilizes the semi-dual OT functional and, remarkably, requires no OT solver. Furthermore, we derive a fast intrinsic dimension estimator from the multi-scale decay of our sampling error estimator. This framework unlocks significant computational and statistical advantages in practice, enabling us to (i) quantify the convergence rate of the discretization error, (ii) calibrate the entropic regularization of Sinkhorn divergences to the data's intrinsic geometry, and (iii) introduce a novel, intrinsic-dimension-based Richardson extrapolation estimator that strongly debiases Wasserstein distance estimation. Numerical experiments demonstrate that our geometry-aware pipeline effectively mitigates the discretization error bottleneck while maintaining computational efficiency.

1. Introduction

Many modern machine learning objects are most naturally modeled as probability distributions: think images (as distributions of pixel intensities), point clouds, word embeddings, or empirical datasets viewed through their sampling measures. A principled way to compare and manipulate such objects is provided by Optimal Transport (OT), which seeks the lowest-cost plan for moving mass from one distribution to another. Through this formulation, one obtains the family of Wasserstein (Earth Mover's) distances and associated transport plans, which have become effective building blocks in complex ML pipelines (Peyré et al., 2019), including generative modeling (An et al., 2019; Chen et al., 2019), speeding up diffusion models (Li et al., 2023), domain adaptation (Courty et al., 2014), and natural language processing (Kusner et al., 2015).

These applications typically operate in the large-scale regime: data may arise from continuous distributions or from discrete distributions supported on an intractably large number of points. In practice, one therefore works with discretized approximations obtained by sampling. This simple step, which replaces unknown measures with empirical ones, has spurred two complementary research threads: (i) Statistical OT, which quantifies the error induced by this discretization; and (ii) Computational OT, which designs solvers that scale to the resulting large discrete problems.

Statistical OT in theory. A central question is how well OT quantities computed on samples approximate their population counterparts. The line of work initiated by (Dudley, 1969) analyzed the 1-Wasserstein discrepancy $W_1(\rho_n, \rho)$ when a continuous measure ρ is replaced by its empirical counterpart ρ_n built from n i.i.d. samples with uniform weights. This was later generalized in (Fournier & Guillin, 2015), showing that $W_p(\rho_n, \rho)$ typically scales as $\mathcal{O}(n^{-1/d})$ in $\mathbb{R}^d$ for all $p \geq 1$, making explicit the curse of dimensionality: the number of samples required to accurately approximate OT quantities grows exponentially with dimension. Estimating the transport *map* itself also suffers from this curse, as established in (Hütter & Rigollet, 2021). Several strategies mitigate these effects: exploiting low intrinsic dimensional structure (Weed & Bach, 2017), or focusing on special settings such as semi-discrete OT (Pooladian et al., 2023). We refer to the recent book (Chewi et al., 2024) for a comprehensive overview.

[1] LPSM, Sorbonne Université, Paris, France [2] Wolfgang Pauli Institute, Vienna, Austria. Correspondence to: Ferdinand Genans <genans.ferdinand@gmail.com>.

Preprint. February 3, 2026.

OT and Intrinsic Dimension. A key insight mitigating the curse of dimensionality is that real-world data rarely fully occupy the ambient space $\mathbb{R}^d$. As established in (Weed & Bach, 2017), if the measure ρ is supported on a manifold $\mathcal{M}$ of lower intrinsic dimension $d_{\mathcal{M}} \ll d$, the estimation error improves significantly to $\mathcal{O}(n^{-1/d_{\mathcal{M}}})$. Furthermore, OT exhibits a favorable multiscale behavior: the approximation error is not governed by a single global dimension, but rather tracks the covering numbers of the support at varying radii, often yielding superior non-asymptotic performance. Crucially, when using OT to compare two different measures, this complexity adapts to the *minimum* intrinsic dimension of the two measures (Hundrieser et al., 2024), and these properties extend to the entropic setting (Stromme, 2024).

Computational OT. On the algorithmic side, discrete OT can be posed as a linear program, but generic LP solvers are prohibitive at scale. The practical impact of OT in ML stems from specialized, structure-exploiting algorithms. A turning point was (Cuturi, 2013), which introduced entropically regularized OT (EOT) solved efficiently by Sinkhorn iterations with complexity $\mathcal{O}(n^2/\varepsilon^2)$ to approximate discrete OT with ε-accuracy. Subsequent work improved constants and convergence guarantees, including primal-dual methods (Dvurechensky et al., 2018; Lin et al., 2019) and recent advances achieving $\mathcal{O}(n^2/\varepsilon)$ (Blanchet et al., 2023), which appears optimal for this class of problems.

The Sinkhorn Divergence: statistical-computational bridge and its limits. A principled attempt to bridge statistical and computational OT lies in the study of *Sinkhorn Divergences* (Ramdas et al., 2017; Genevay et al., 2018; Feydy et al., 2019). By explicitly defining the discrepancy through the discrete Sinkhorn divergence, one achieves a parametric estimation rate of $\mathcal{O}(\varepsilon^{-d/2} n^{-1/2})$ (Genevay et al., 2019; Mena & Niles-Weed, 2019; Chizat et al., 2020). Consequently, the curse of dimensionality is not removed but rather shifted into the regularization parameter ε. Crucially, the exponent d governing this trade-off is the *intrinsic* dimension $d_{\mathcal{M}}$ of the data manifold (Stromme, 2024), and a well-calibrated speed at which $\varepsilon \to 0$ nearly achieves the minimax rates to estimate the Wasserstein distance.

Missing Pieces. Despite this progress, a fully operational link between statistical theory and computational practice remains elusive. While Sinkhorn Divergences offer a theoretical trade-off, calibrating it well requires knowledge of the typically unknown underlying intrinsic dimension $d_{\mathcal{M}}$ to select the optimal ε. As a result, practitioners are left without concrete guidance: they often drive numerical solvers to very high precision, implicitly treating the sampling error as negligible, or tune ε heuristically without addressing the bias-variance trade-off. To apply OT with calibrated, problem-dependent accuracy, one requires a practical way to estimate the discretization error and the intrinsic dimension directly from the available samples.

Contributions. To address these challenges, we propose a geometry-aware framework that estimates both the discretization error and the intrinsic dimension directly from samples, without relying on expensive ground-truth computations. Our specific contributions are:

- **A solver-free discretization error estimator.** We introduce a novel estimator for the quantization error $\mathrm{OT}_c(\rho, \rho_n^*)$, where ρ_n^* represents the empirical measure with weights optimally adjusted to minimize the transport cost. Our key theoretical insight is that the semi-dual formulation of this problem admits a closed-form solution. This allows us to estimate the error via simple Monte Carlo integration, bypassing the need for an OT solver entirely. The resulting estimator is parameter-free, highly parallelizable on GPUs, and computationally inexpensive.
- **A scalable intrinsic dimension estimator.** Leveraging the convergence behavior of our error estimator, we derive an estimator of the empirical intrinsic dimension d_{int} of the data. By analyzing the decay of the error as a function of sample size, our estimator captures the multi-scale geometric structure of the distribution, consistent with the theoretical findings in (Weed & Bach, 2017). Crucially, this method scales linearly (i.e., $\mathcal{O}(n)$) in both time and space $\mathcal{O}(n)$, making it applicable to large-scale datasets.
- **Debiased Wasserstein estimation via Diagonal Richardson Extrapolation.** We propose a new extrapolation scheme that combines our intrinsic dimension estimate with Sinkhorn Divergences. While previous work (Chizat et al., 2020) utilized Richardson extrapolation solely on the regularization parameter ε, we introduce a *diagonal* approach that links ε to the sample size n and the intrinsic dimension d_{int}. This joint debiasing strategy effectively cancels out both the first-order entropic bias and the statistical discretization error. This scheme permits us to have a convergence of $o(n^{-2/(d_{\text{int}}+4)})$ for W_2^2, compared to $o(n^{-2/(d+8)})$ for the ε-only Richardson scheme.

We validate our pipeline on synthetic manifolds and real-world datasets (MNIST, CIFAR), demonstrating that our dimension estimator is accurate and that our extrapolation scheme yields Wasserstein estimates with significantly reduced bias compared to standard baselines.

Notations. We note $\|\cdot\|$ the Euclidean norm and ω_d the volume of the unit d-dimensional ball. For $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. $\mathcal{P}(\mathbb{R}^d)$ is the set of probabilities in $\mathbb{R}^d$, and for $\rho \in \mathcal{P}(\mathbb{R}^d)$, $\mathrm{Supp}(\rho)$ is its support. Δ_d represents the probability simplex in $\mathbb{R}^d$: $\Delta_d := \{w \in \mathbb{R}^d_+ : \sum_{i=1}^d w_i = 1\}$. $\mathcal{O}(\cdot)$ and $o(\cdot)$ are the usual approximation orders. We use $f \lesssim g$ if there exists a universal constant $C > 0$ such that $f(\cdot) \leq Cg(\cdot)$. We write $a \asymp b$ if both $a \lesssim b$ and $b \lesssim a$.

2. Preliminaries

2.1. Optimal Transport: Primal, Dual, and Semi-Dual

We consider probability measures supported on a compact metric space $\mathcal{X} \subset \mathbb{R}^d$. Given source and target measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the Kantorovich optimal transport problem is defined as:

$$\mathrm{OT}_c(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\pi(x, y), \tag{1}$$

where $\Pi(\mu, \nu)$ denotes the set of joint distributions with marginals μ and ν. When $c(x, y) = \|x - y\|^p$ for $p \geq 1$, the quantity $\mathrm{OT}_c(\mu, \nu)^{1/p}$ defines the p-Wasserstein distance, denoted $W_p(\mu, \nu)$ (Villani, 2009).

Dual and Semi-Dual Formulations. The Kantorovich problem admits a dual formulation involving continuous potential functions $f, g \in C(\mathcal{X})$:

$$\mathrm{OT}_c(\mu, \nu) = \sup_{\substack{f,g \\ f(x)+g(y) \leq c(x,y)}} \int f d\mu + \int g d\nu. \tag{2}$$

By defining the c-transform of a potential g as $g^c(x) := \inf_y [c(x, y) - g(y)]$, one can eliminate the constraint $f(x) \leq c(x, y) - g(y)$ by setting $f = g^c$. This yields the **semi-dual formulation**, which depends on a single potential g:

$$\mathrm{OT}_c(\mu, \nu) = \sup_{g \in C(\mathcal{X})} \int g^c d\mu + \int g d\nu. \tag{3}$$

This semi-dual form is crucial for our analysis as, even if μ is continuous, having ν discrete converts the constrained continuous problem into a finite dimension problem, a property we will leverage to construct our error estimator.

2.2. Entropic Regularization and Sinkhorn Divergences

To overcome the computational cost of exact OT, one commonly employs Entropic Regularization. We define the regularized objective in its primal form by adding a Kullback-Leibler penalty:

$$\mathrm{OT}_{c,\varepsilon}(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int c d\pi + \varepsilon \mathrm{KL}(\pi | \mu \otimes \nu), \tag{4}$$

where $\varepsilon > 0$ is the regularization strength. To correct for the bias introduced by entropy (where $\mathrm{OT}_{c,\varepsilon}(\mu, \mu) \neq 0$), we can use the the **Sinkhorn Divergence**:

$$S_\varepsilon(\mu, \nu) := \mathrm{OT}_{c,\varepsilon}(\mu, \nu) - \frac{1}{2}\mathrm{OT}_{c,\varepsilon}(\mu, \mu) - \frac{1}{2}\mathrm{OT}_{c,\varepsilon}(\nu, \nu), \tag{5}$$

which, for $c = \|\cdot\|^2$, is a divergence that interpolates between OT (as $\varepsilon \to 0$) and Maximum Mean Discrepancy (as $\varepsilon \to \infty$) (Genevay et al., 2018; Feydy et al., 2019) .

2.3. Statistical Rates and Intrinsic Dimension

When approximating μ and ν via empirical measures $\widehat{\mu}_n$ and $\widehat{\nu}_n$, the discretization error was classically shown to scale with the ambient dimension d. Informally, for p-Wasserstein costs and Sinkhorn Divergences:

$$\mathbb{E}[|W_p^p(\widehat{\mu}_n, \widehat{\nu}_n) - W_p^p(\mu, \nu)|] = \mathcal{O}(n^{-p/d}),$$
$$\mathbb{E}[|S_\varepsilon(\widehat{\mu}_n, \widehat{\nu}_n) - S_\varepsilon(\mu, \nu)|] = \mathcal{O}(n^{-1/2}\varepsilon^{-d/2}).$$

Crucially, subsequent work established that d is not the ambient dimension, but the **minimum intrinsic dimension** of the two measures, both for OT (Weed & Bach, 2017; Hundrieser et al., 2024) and EOT (Groppe & Hundrieser, 2024; Stromme, 2024). More precisely, the rate tracks the ϵ-covering number $\mathcal{N}_\epsilon$, defining an effective dimension $d_\epsilon \approx \min_{\rho\in\mu,\nu} \frac{\log \mathcal{N}_\epsilon(\rho)}{-\log \epsilon}$. This quantity exhibits a *multi-resolution behavior*: OT and EOT adapt to the geometry at varying scale of observation, capturing d_ε in the non-asymptotic regime where the discretization error is still significant. This allows the estimator to exploit the effective geometry at coarse scales, in regimes where $d_\epsilon > \lim_{\varepsilon\to 0} d_\epsilon$.

3. A Simple Estimator of the Wasserstein Discretization Error

As discussed in Section 2, the reliability of OT in machine learning hinges on controlling the error incurred when replacing continuous measures with discrete samples. A fundamental insight into this error comes from the triangle inequality of the Wasserstein distance. For any two measures μ, ν and their discrete approximations $\widehat{\mu}, \widehat{\nu}$, the estimation error is bounded by:

$$|W_p(\mu,\nu) - W_p(\widehat{\mu},\widehat{\nu})| \le W_p(\mu,\widehat{\mu}) + W_p(\nu,\widehat{\nu}).$$

This decomposition reveals that the total error is governed by how well discrete samples capture the geometry of the underlying distributions. While asymptotic theory predicts a rate of $n^{-1/d}$, these bounds offer limited practical guidance for finite datasets. Estimating the term $\mathrm{OT}_c(\rho,\widehat{\rho})$ for $\rho = \mu, \nu$ directly from data serves two critical purposes: (i) it provides a concrete, non-asymptotic quantification of the approximation quality; and (ii) the rate at which this error decays with sample size acts as an indicator of the *intrinsic dimension* of the measure.

In this section, we introduce a fast, solver-free estimator for this one-sample discretization error $\mathrm{OT}_c(\rho,\widehat{\rho}^*_n)$, where $\widehat{\rho}^*_n$ represents the optimal discrete approximation of ρ given a fixed support of n points.

3.1. Wasserstein discretization error and semi-discrete OT

Consider a measure ρ and a fixed support set $X = (X_1, ..., X_n)$ sampled i.i.d from ρ. For any weight vector $\mathbf{a} \in \Delta_n$, let $\widehat{\rho}_{m,\mathbf{a}} = \sum_{i=1}^n a_i \delta_{X_i}$ be the corresponding discrete measure. The computation of $\mathrm{OT}_c(\rho,\widehat{\rho}_{m,\mathbf{a}})$ constitutes a semi-discrete optimal transport problem. Even with continuous ρ, this problem admits a finite-dimensional semi-dual formulation (3):

$$\mathrm{OT}_c(\rho,\widehat{\rho}_{m,\mathbf{a}}) = \max_{\mathbf{g}\in\mathbb{R}^n} \left[\int_{\mathbb{R}^d} \mathbf{g}^c(x)\mathrm{d}\rho(x) + \sum_{i=1}^n a_i g_i \right]. \tag{6}$$

Solving (6) typically requires expensive stochastic gradient descent or semi-discrete solvers (Genevay et al., 2018; Kitagawa et al., 2016), making it impractical if the goal is merely to evaluate the quality of the discretization when one uses $\mathrm{OT}_c(\widehat{\mu}_n,\widehat{\nu}_n)$ as a computationally tractable proxy of $\mathrm{OT}_c(\mu,\nu)$. Remarkably, the following proposition demonstrates that we can bypass optimization entirely. By selecting the optimal weights for the fixed support, the dual potential collapses to the zero vector, yielding a closed-form solution.

Proposition 3.1. *Let c be a continuous cost. Given $\rho \in \mathcal{P}(\mathbb{R}^d)$ and a support $X = (x_1,\dots,x_n) \in (\mathbb{R}^d)^n$, define the weights $\mathbf{w}_n = (w_1,\dots,w_n)^\top$ by*

$$w_i \;=\; \rho\big(\{x\in\mathbb{R}^d; \min_j c(x,x_j) = c(x,x_i)\}\big),$$

and set $\widehat{\rho}^*_n = \sum_{i=1}^n w_i\delta_{x_i}$. Then $\widehat{\rho}^*_n$ provides the best approximation of ρ supported on X, that is,

$$\widehat{\rho}^*_n \in \mathrm{argmin}_{\mathbf{a}\in\Delta_n} \Big\{ \mathrm{OT}_c(\rho,\widehat{\rho}^{\mathbf{a}}_n) \;\Big|\; \widehat{\rho}^{\mathbf{a}}_n = \sum_{i=1}^n a_i\delta_{x_i} \Big\},$$

and the minimizer is unique. Moreover, the zero vector $\mathbf{0}_n$ is the maximizer of the semi-dual (6) at $\widehat{\rho}^*_n$. That is,

$$\mathrm{OT}_c(\rho,\widehat{\rho}^*_n) \;=\; \int_{\mathbb{R}^d} \mathbf{0}^c_n(x)d\rho(x).$$

Proof. By definition of $\mathbf{w}_n$, we have that the vector $\mathbf{0}_n$ is the Kantorovich potential of the semi-discrete OT problem between ρ and ρ_n. Indeed, the first-order optimality condition, stating that $\mathbf{g}$ is optimal if and only if $\nabla H(\mathbf{g}) = \mathbf{0}_n$ reads $\rho(\{\mathbf{g}^c(x) = c(x,X_i) - g_i\}) = w_i$ which is exactly how we defined the weights w_i fixing $g_i = 0$, $i \in [\![1,n]\!]$.

Since $\mathbf{0}_n$ is its Kantorovich potential, the optimal transport reads for all $x \in \mathbb{R}^d : x \mapsto X_i = \mathrm{argmin}_{X_i} c(x, X_i)$. That is, every point is sent to its closest neighbor from in X. For any other probability weight vector, another $\mathbf{g}' \in \mathbb{R}^n$ such that $\mathbf{g}' \notin \{\lambda \mathbf{1}_n, \lambda \in \mathbb{R}\}$ will be an optimal vector, which will lead to an OT plan where a strictly positive mass of ρ will not be sent to its nearest neighbor, therefore increasing the OT cost. □

3.2. A Solver-Free Monte Carlo Estimator

The key insight of Proposition 3.1 is that for the optimally weighted measure $\widehat{\rho}_n^*$, the semi-dual objective is maximized at $\mathbf{0}_n$. This simplifies the OT cost to the integral of the c-transform of the zero vector, which is simply the distance to the nearest neighbor in the support X. Consequently, we can approximate $\mathrm{OT}_c(\rho, \widehat{\rho}_n^*)$ using a direct Monte Carlo (MC) estimator $\widehat{\mathrm{OT}}_N^n$ using N additional samples $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} \rho$:

$$\widehat{\mathrm{OT}}_N^n := \frac{1}{N}\sum_{k=1}^{N} \mathbf{0}_n^c(X_k) \;=\; \frac{1}{N}\sum_{k=1}^{N} \min_{j \in [\![1,n]\!]} c(X_k, x_j). \tag{7}$$

This estimator is computationally efficient, fully parallelizable on GPUs, and requires no OT solver. Furthermore, if explicit weights are required, they also admit a natural MC estimator:

$$w_i \;\approx\; \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}\left\{\arg\min_{j} c(X_k, x_j) = i\right\} \;=:\; \widehat{w}_{i,N}.$$

The following theorem establishes that these estimators concentrate exponentially fast around their true values.

Proposition 3.2. *Suppose that ρ has a bounded support and that the cost function is continuous. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have:*

$$\left|\mathrm{OT}_c(\rho, \widehat{\rho}_n^*) - \widehat{\mathrm{OT}}_N^n\right| \leq \underbrace{\sqrt{\tfrac{2\widehat{\sigma}^2_{\mathbf{0}^c} \log(2/\delta)}{N}} + \tfrac{7C_\rho \log(2/\delta)}{3(N-1)}}_{E_{MC,OT}},$$

where $C_\rho = \sup_{x,y \in supp(\rho)} c(x,y)$ and $\widehat{\sigma}^2_{\mathbf{0}^c}$ is the usual biased sample variance estimator

$$\frac{1}{N}\sum_{k=1}^{N}\left(\min_{j \in [\![1,n]\!]} c(X_k, x_j) - \widehat{\mathrm{OT}}_N^n\right)^2.$$

Also, for any $i \in [\![1, n]\!]$, with probability $1 - \delta$, we have

$$|w_i - \widehat{w}_{N,i}| \leq \underbrace{\sqrt{\tfrac{2\widehat{w}_{N,i}(1-\widehat{w}_{N,i}) \log(2/\delta)}{N}} + \tfrac{7\log(2/\delta)}{3(N-1)}}_{E_{MC,w_i}}.$$

Proof. Under our assumptions, we can bound a.s. $\mathbf{0}_n^c(\cdot) = \min_{j \in [\![1,n]\!]} c(\cdot, x_j)$ by C_ρ. Therefore, the empirical Bernstein bound in Theorem 4 in (Maurer & Pontil, 2009) yields the first inequality. The same theorem applies to the second inequality, since $w_i \in [0,1]$ for all i. □

We conclude with two practical observations. First, although the bound depends on the diameter C_ρ, this term appears linearly and does not affect the convergence rate. Second, while standard practice often assumes uniform weights ($\widehat{\rho}_n^{\mathrm{unif}} = \frac{1}{m}\sum \delta_{x_i}$), our estimator $\widehat{\mathrm{OT}}_N^n$ targets the optimally weighted measure $\widehat{\rho}_n^*$. Since $\mathrm{OT}_c(\rho, \widehat{\rho}_n^*) \leq \mathrm{OT}_c(\rho, \widehat{\rho}_n^{\mathrm{unif}})$, our method provides a tight lower bound on the discretization error achievable on the support X. More importantly, because $\widehat{\rho}_n^*$ optimally adapts the mass to the geometry of the support, the decay rate of $\mathrm{OT}_c(\rho, \widehat{\rho}_n^*)$ as n increases is driven purely by the intrinsic geometry of ρ. This property makes $\widehat{\mathrm{OT}}_N^n$ the key building block for our fast intrinsic dimension estimator, detailed in the next section.

4. Discretization Error and Intrinsic Dimension Estimation

4.1. Geometric setting: Multi-scale behavior and Covering Numbers

A recurring theme in machine learning is that high-dimensional data often exhibits low-dimensional structure. While the *manifold hypothesis*, positing that data concentrates on a smooth submanifold $\mathcal{M} \subset \mathbb{R}^d$, provides a useful baseline, real-world data rarely adheres strictly to such idealized assumptions. Instead, complex distributions often display a multi-resolution behavior, where the effective geometric complexity depends on the scale of observation or the desired approximation accuracy.

To illustrate this, consider $\rho \in \mathcal{P}(\mathbb{R}^{10})$ constructed as the following mixture: $\rho := \frac{1}{2}\mathcal{U}([0,1]^2 \times \{0\}^8) + \frac{1}{2}\mathcal{U}([1,2]^8 \times \{0\}^2)$. Strictly speaking, the support has a topological dimension of 8. However, the behavior of the optimal transport discretization error reveals a more nuanced reality. If one aims for a coarse approximation (e.g., a discretization error of 50%), the problem may behave effectively like a 2-dimensional problem. Conversely, demanding a refined accuracy (e.g., less than 10% error) forces the discretization to fill the 8-dimensional volume, causing the error rate to suffer from the curse of dimensionality associated with $d = 8$.

This example highlights that for probability measures, the relevant notion of complexity is not a static integer but a dynamic quantity related to the *covering numbers* of the support at different radii (Weed & Bach, 2017). Consequently, we replace rigid geometric definitions (such as the manifold hypothesis) with a functional assumption based on the convergence rate of the Wasserstein distance itself.

We assume that within a specific range of sample sizes $n \in [n_{\min}, n_{\max}]$, the discretization error is governed by an effective intrinsic dimension d_{int}.

Assumption 4.1 (Effective Intrinsic Dimension)**.** Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ be a probability measure with finite diameter. We assume that for $n \in [n_{\min}, n_{\max}]$, the convergence rate of the 1-Wasserstein discretization error is governed by a dimension parameter $s = d_{\text{int}}$. Specifically, there exist constants $C_1, C_2 > 0$ depending essentially on Diam(ρ), such that $C_2 \leq cC_1$ for a universal constant $c \geq 1$, and:

Lower bound (Quantization limit): For any discrete measure σ_n supported on at most n points, the approximation error is lower-bounded by the dimension:

$$W_1(\rho, \sigma_n) \geq C_1 n^{-1/s}. \tag{8}$$

Upper bound (Statistical performance): The empirical measure $\widehat{\rho}_n$ constructed from n i.i.d. samples satisfies:

$$\mathbb{E}[W_1(\rho, \widehat{\rho}_n)] \leq C_2 n^{-1/s'}, \tag{9}$$

for all $s' \geq s$. We refer to the exponent s as the **intrinsic dimension** d_{int} of ρ in the regime $[n_{\min}, n_{\max}]$.

This assumption posits that d_{int} captures the ”difficulty” of the Wasserstein discretization error of ρ with a finite number of points. In the asymptotic limit ($n_{\min} \to \infty$), if ρ admits a density with respect to a manifold of dimension $d_{\mathcal{M}}$, then s coincides with $d_{\mathcal{M}}$. We refer to the appendix for further discussion, and bounds on the constants in specific settings.

4.2. Leveraging our discretization estimator to estimate the intrinsic dimension.

To estimate the intrinsic dimension, we leverage the scaling behavior of the one-sample discretization error established in Assumption 4.1. Using our Monte Carlo estimator $\widehat{\text{OT}}_N^n$ (with $c(x,y) = \|x - y\|$) defined in Section 3, we evaluate the error at two distinct support sizes, n and ηn (with $\eta > 1$). The ratio of these estimates reveals the exponent governing the convergence rate, leading to the following estimator:

$$\widehat{d}^* := \frac{\log \eta}{\log \widehat{\text{OT}}_N^n - \log \widehat{\text{OT}}_N^{\eta n}},$$

where N denotes the number of samples used in the MC approximation. Evaluating $\widehat{d}^*$ reduces to computing two MC estimates, yielding a linear complexity in n. The statistical guarantee of our method is given in the following theorem.

Theorem 4.2. *Suppose that Assumption 4.1 holds and that* $d_{\text{int}} > 2$ *and let Diam*(ρ) *denote the diameter of* ρ*'s support.*

Then, for $\gamma, \xi > 0$ *and* $n\eta \le n_{\max}$, *using*

$$\begin{aligned} \eta &\ge \left(\frac{(1+\xi)C_2}{C_1}\right)^{\frac{2d_{\text{int}}}{\gamma}}, \\ n &\ge \left(\frac{\mathrm{Diam}(\rho)^2}{2\xi^2 C_2}\log(4/\delta)\right)^{\frac{d_{\text{int}}}{d_{\text{int}}-2}}, \\ N &\ge \left[\frac{16\, d_{\text{int}}^2}{\gamma^2(\log\eta)^2} \vee 1\right] \frac{2\,\mathrm{Diam}(\rho)^2}{C_1^2}\eta^{2/d_{\text{int}}} n^{2/d_{\text{int}}} \log\left(\frac{8}{\delta}\right), \end{aligned}$$

we have with probability $1-\delta$

$$\frac{d_{\text{int}}}{1+\gamma} \le \widehat{d}^* \le (1+\gamma)d_{\text{int}}\,.$$

Comparison to the literature. The concept of leveraging the multi-scale decay of Wasserstein error for dimension estimation was pioneered in (Block et al., 2022). However, their estimator relies on the two-sample empirical distance:

$$\widehat{d} \;=\; \frac{\log\eta}{\log W_1(\widehat{\rho}_n, \widehat{\rho}'_n) - \log W_1(\widehat{\rho}_{\eta n}, \widehat{\rho}'_{\eta n})}\,,$$

where $\widehat{\rho}_n$ and $\widehat{\rho}'_n$ are independent empirical measures of n points. This approach rests on the fact that $\mathbb{E}[W_1(\widehat{\rho}_n, \widehat{\rho}'_n)] \approx \mathbb{E}[W_1(\widehat{\rho}_n, \rho)]$, but it faces a severe computational bottleneck: evaluating $W_1(\widehat{\rho}_n, \widehat{\rho}'_n)$ requires solving a full discrete optimal transport problem. Standard Linear Programming solvers scale as $\mathcal{O}(n^3)$, while approximate Sinkhorn solvers scale as $\mathcal{O}(n^2/\varepsilon)$. This prohibitive cost restricts such estimation to small datasets or requires substantial approximations that may bias the dimension estimate.

In contrast, our estimator $\widehat{d}^*$ relies on the one-sample error to the optimal support, which we approximate via Monte Carlo without any optimization. The resulting complexity is linear $\mathcal{O}(n)$, allowing for dimension estimation on large-scale datasets where standard OT solvers are intractable.

4.3. Experiments

We compare our estimator, denoted as Semi-Discrete W_1, against the Discrete W_1 estimator based on (Block et al., 2022) and standard baselines from the `scikit-dimension` package (Bac et al., 2021): MLE, CorrInt, lPCA, TwoNN.

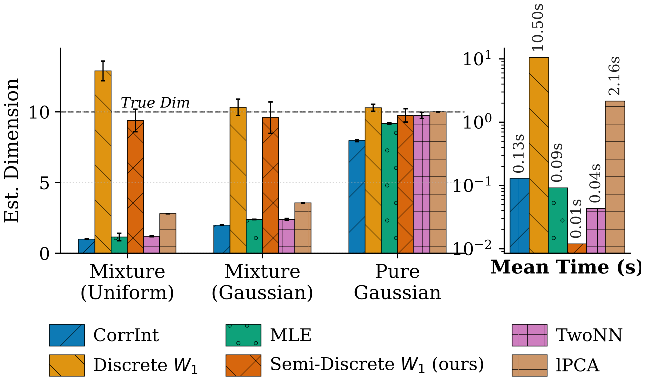

Figure 1. Intrinsic Dimension Estimation Benchmark. Comparison of our Semi-Discrete W_1 estimator against the Discrete W_1 baseline and standard geometric estimators. The dashed line represents the ground truth effective dimension $d_{\text{int}} = 10$. Our estimator matches the robustness of discrete OT on mixtures but runs orders of magnitude faster (< 0.1s vs ~ 10s).

Experimental Setup: We generate data in $\mathbb{R}^{20}$ under three configurations designed to have an effective Wasserstein dimension of $d_{\text{int}} = 10$: (1) a mixture of hypercubes (80% 2D, 20% 10D); (2) a mixture of low-rank Gaussians (80% rank 2, 20% rank 10); and (3) a single rank-10 Gaussian. OT-based estimators compare supports of size $n = 2000$ and $\eta n = 3000$, while baselines use a fixed size of 3000. Experiments were repeated 20 times.

Figure 1 highlights the performance differences. While all methods succeed on the simple manifold (Config 3), standard geometric estimators fail to capture the global transport complexity of the mixtures (Configs 1 & 2). In contrast, both OT-based methods robustly recover $d_{\text{int}} \approx 10$. Crucially, our Semi-Discrete estimator is drastically more efficient than the Discrete Wasserstein-based estimator: it computes the estimate in under 0.1 seconds, compared to $\simeq 10$ seconds on GPU for the Discrete W_1 baseline which requires solving exact OT problems. For reference, CPU run times for `scikit-dimension` baselines are also provided.

5. Coupled Bias Reduction: The Diagonal Richardson Estimator

Sinkhorn divergence serves as a fast proxy for OT, supported by efficient solvers like `geomloss` (Feydy et al., 2019) which offer linear memory usage and high GPU performance, provided ε is not vanishingly small. However, estimating the unregularized Optimal Transport cost from finite samples with the Sinkhorn divergence presents a dual challenge: the estimator is biased by both the regularization parameter ε (entropic bias) and the finite sample size n (statistical bias). To further reduce the entropic bias, (Chizat et al., 2020) introduced a Richardson Extrapolation scheme on ε, proposing the estimator $\widehat{R}_{\varepsilon,n} = 2\widehat{S}_{\varepsilon,n} - \widehat{S}_{\sqrt{\varepsilon},n}$. Although this approach theoretically improves the approximation error to $\mathbb{E}[|\widehat{R}_{\varepsilon,n} - W_2^2|] = o(n^{-2/(d+8)})$ (with $\varepsilon \asymp n^{-1/(d+8)}$). The first improvement is naturally to replace d in ε with our estimator $\widehat{d}^*$ of d_{int}. Then, even if the sample approximation is not the asymptotic limiting factor, its large constant pre-factor can render entropic debiasing ineffective in practice by targeting a term that is not the dominant source of error.

To better tackle both the entropic and statistical bias, we propose a *coupled* strategy—the **Diagonal Richardson Estimator**—which links the regularization schedule ε directly to the sample size n. By tackling both biases simultaneously, this method leverages our intrinsic dimension estimator to calibrate the debiasing optimally.

Total bias decomposition. It is well known in statistical OT that the entropic regularization should decay with the number of samples; i.e., $\varepsilon_n \asymp n^{-a}$ for some exponent $a > 0$. The total expected error decomposes into:

$$\mathbb{E}[\widehat{S}_{\varepsilon_n,n}] - W_2^2 = \underbrace{\left(\mathbb{E}[\widehat{S}_{\varepsilon_n,n}] - S_{\varepsilon_n}\right)}_{\text{Statistical Bias}} + \underbrace{\left(S_{\varepsilon_n} - W_2^2\right)}_{\text{Entropic Bias}} .$$

We characterize the asymptotic expansions of both terms. The entropic bias expansion (10) follows from Proposition 4 in (Chizat et al., 2020) (assuming finite Fisher information of the measures), while the statistical bias expansion (11) is derived from Theorem 2 in (Stromme, 2024), which establishes the lower complexity adaptation of the Sinkhorn divergence:

$$\text{Bias}_{\text{ent}}(n) = C_3 n^{-2a} + o_{\text{ent}}(n^{-2a}) , \tag{10}$$

$$\text{Bias}_{\text{stat}}(n) \leq C_4 n^{-(1-ad_{\text{int}})/2} + o_{\text{stat}}(n^{-(1-ad_{\text{int}})/2}) . \tag{11}$$

Remark 5.1. The constant governing the statistical bias is formally of the form $c(\varepsilon)n^{-1/2}$, depending on the second derivative of the Sinkhorn divergence. The upper bound $c(\varepsilon) \lesssim \varepsilon^{-d_{\text{int}}/2}$ was obtained independently in (Stromme, 2024; Groppe & Hundrieser, 2024), Example 5, through the rate $n^{-(1-ad_\epsilon)/2}$ when $\varepsilon \asymp n^{-a}$ on the standard error. This rate is solely governed by the statistical bias since $\widehat{S}_{\varepsilon,n} - \mathbb{E}[\widehat{S}_{\varepsilon,n}] = \mathcal{O}_{\mathcal{P}}(1/\sqrt{n})$ uniformly in $\varepsilon \geq 0$ in Proposition 4 of (Chizat et al., 2020). The sharpness of the rate of the statistical bias was not perfectly established, though discussions are provided for our case of interest $\varepsilon = n^{-a}$ and d_ϵ replaced by d_{int}. Therefore, to be completely rigorous, we use the equality in (11) as an assumption, as it is compulsory for our Richardson method to effectively cancel the bias term.

Assumption 5.2 (Sharpness of Statistical Bias)**.** We assume that the statistical estimation error of the Sinkhorn divergence admits a sharp asymptotic expansion. That is, equality holds in (11):

$$\mathbb{E}[\widehat{S}_{\varepsilon_n,n}] - S_{\varepsilon_n} \;=\; C_4 n^{-(1-ad_{\text{int}})/2} + o_{\text{stat}}(n^{-(1-ad_{\text{int}})/2}) ,$$

where $C_4 \neq 0$ and d_{int} is the intrinsic dimension of the data.

Optimal schedule and the Diagonal Estimator. To construct an efficient estimator, we seek a rate $\varepsilon \asymp n^{-a}$ that balances the convergence rates of the entropic and statistical biases. Equating the exponents $4a = 1 - ad_{\text{int}}$ yields the optimal decay rate $a = \frac{1}{d_{\text{int}}+4}$. Under this schedule, a *diagonal* Richardson extrapolation can eliminate the first-order terms of both biases simultaneously.

We define our Diagonal Richardson Estimator using sample sizes n and $2n$ as:

$$\widehat{R}_{2n}^{\text{Diag}} := w_{2n}\widehat{S}_{\varepsilon_{2n},2n} + w_n\widehat{S}_{\varepsilon_n,n} \ ,$$

where the weights are determined by the common decay rate $\gamma = \frac{2}{d_{\text{int}}+4}$ of the bias terms:

$$w_{2n} = \frac{2^\gamma}{2^\gamma - 1}, \quad w_n = \frac{-1}{2^\gamma - 1}.$$

The following proposition establishes the improved convergence rate of this coupled estimator.

Proposition 5.3. *Under Assumption 5.2, and using the schedule $\varepsilon_n \asymp n^{-1/(d_{\text{int}}+4)}$, the Diagonal Richardson estimator satisfies:*

$$\mathbb{E}\left[|\widehat{R}_{2n}^{Diag} - W_2^2|\right] = o(n^{-\frac{2}{d_{\text{int}}+4}}) \ .$$

Furthermore, if the higher-order terms satisfy $o_{ent}(n^{-2a}) = \mathcal{O}(n^{-4a})$ and $o_{stat}(n^{-(1-ad_{\text{int}})}) = \mathcal{O}(n^{-2+2ad_{\text{int}}})$, the rate improves to $\mathcal{O}(n^{-\frac{4}{d_{\text{int}}+4}})$.

Therefore, this rate is slightly better than the ε-only Richardson, which is $(n^{-\frac{2}{d_{\text{int}}+8}})$ using d_{int}.

Proof sketch. Let $R^{\text{Diag}} = \lim_{n\to\infty}\mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}]$. The error decomposes as:

$$\begin{aligned}\mathbb{E}\left[|\widehat{R}_{2n}^{\text{Diag}} - W_2^2|\right] &\le \mathbb{E}\left[|\widehat{R}_{2n}^{\text{Diag}} - \mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}]|\right] \\ &+ |\mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}] - R^{\text{Diag}}| + |R^{\text{Diag}} - W_2^2| \\ &= \mathcal{O}(1/\sqrt{n}) + o_{\text{stat}}(n^{-\frac{2}{d_{\text{int}}+4}}) + o_{\text{ent}}(n^{-\frac{2}{d_{\text{int}}+4}}).\end{aligned}$$

The stochastic error scales as $\mathcal{O}(n^{-1/2})$, which is negligible compared to the bias terms for $d_{\text{int}} > 2$. The choice of weights cancels the leading order terms $n^{-\frac{2}{d_{\text{int}}+4}}$ in the bias, leaving only the higher-order remainder. Remark that if Assumption 5.2 is not fulfilled, we still have the asymptotically (in d_{int}) minimax rate $\mathcal{O}(n^{-2/(d_{\text{int}}+4)})$.

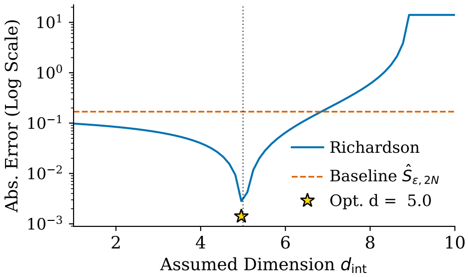

Figure 2. **Sensitivity to Intrinsic Dimension.** We perform Diagonal Richardson extrapolation, $2n = 2000$, on a source synthetic mixture in $\mathbb{R}^{10}$ (90% mass on a 5D Gaussian, 10% on a 1D Gaussian) to a 10D Gaussian. By varying the dimension parameter d used in the extrapolation weights, we observe that the estimation error is minimized exactly at the dominant intrinsic dimension $d = 5$. This confirms that correctly calibrating the schedule to the effective geometry is essential for optimal debiasing.

Bagged Diagonal Richardson. While the Diagonal Richardson estimator $\widehat{R}_{2n}^{\text{Diag}}$ successfully reduces the bias, this improvement comes at the cost of increased variance. This variance increase comes from extrapolation involving a weighted difference from the coefficients w_{2n} and w_n. Consequently, the noise from the subsampled term $\widehat{S}_{\varepsilon_n,n}$ is amplified in the final estimate. Furthermore, relying on a single random subset of size n is statistically inefficient, as it discards information from the unused samples for the low-resolution component. To mitigate this variance increase, we employ a bagging strategy that averages the low-resolution estimator over K independent subsamples, while retaining the full dataset for the high-resolution term. The *Bagged Richardson Estimator* is defined as:

$$\widehat{R}_{K,2n}^{\text{Bagged}} := w_{2n}\widehat{S}_{\varepsilon_{2n},2n} + \frac{w_n}{K}\sum_{k=1}^{K}\widehat{S}_{\varepsilon_n^{(k)},n} \ . \tag{12}$$

This approach stabilizes the estimator by driving down the variance of the subtractive term. We quantify this stability in terms of the variance inflation relative to the standard Sinkhorn estimator. The proof, detailed in the Appendix, builds on the second-order Hadamard differentiability of the Sinkhorn divergence established in (Goldfeld et al., 2024).

Proposition 5.4 (Variance Stability)**.** *The variance of the Bagged Richardson estimator relates to the variance of the standard Sinkhorn estimator $\widehat{S}_{\varepsilon_{2n},2n}$ as:*

$$\mathrm{Var}(\widehat{R}^{\mathit{Bagged}}_{K,2n}) = \mathrm{Var}(\widehat{S}_{\varepsilon_{2n},2n}) \times \left(1 + \tfrac{1}{(2^\gamma - 1)^2 K}\right) + o(1).$$

Crucially, as the number of bags $K \to \infty$, the variance overhead vanishes. The bagged estimator thus achieves the improved bias rate of the Richardson scheme while converging to the optimal variance floor dictated by the full sample size $2n$.

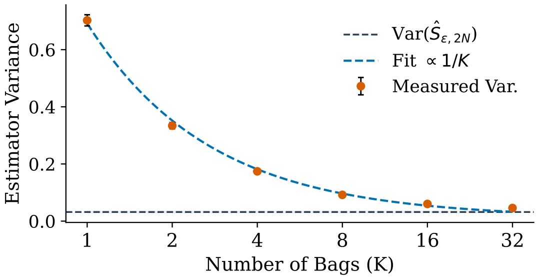

Figure 3. **Variance Reduction via Bagging.** Using the same mixture setting as in Figure 2, we measure the variance of the Bagged Richardson estimator as a function of the number of bags K. The variance is computed over 20 runs and repeated 10 times. We observe a clear $1/K$ decay in the variance overhead, which converges asymptotically to the variance floor of the standard Sinkhorn estimator $\widehat{S}_{\varepsilon_{2n},2n}$ (dashed line). This confirms that bagging eliminates the stability cost of extrapolation.

5.1. From OT Geometry to Wasserstein Distance Estimation

We validate our full ”geometry-aware” pipeline by estimating W_2^2 on image manifolds (MNIST, FashionMNIST, CIFAR10). To overcome the lack of ground truth for empirical measures, we construct synthetic target measures $\nu = (T^*)_{\#}\mu$ using a strictly convex potential $\psi_\theta(x) = \frac{1}{2}\|x\|^2 + \mathrm{ICNN}(x)$. By Brenier's theorem, $T^* = \nabla\psi_\theta$ is the unique optimal map, providing an exact reference distance $W_2^2(\mu,\nu) = \mathbb{E}[\|x - T^*(x)\|^2]$.

Experimental Setup. For each dataset, with $n = 1000$, we first estimate the intrinsic dimension $\widehat{d}^*$ (Sec. 4) and plug it into our Bagged Diagonal Richardson estimator ($K = 12$) to calibrate the bias cancellation weights. We compare our approach against the standard Sinkhorn divergence (”Base”), and the standard ε-Richardson extrapolation (”Eps-Rich”) (Chizat et al., 2020).

Table 1. **Benchmark of Wasserstein estimation W_2^2.**

Dataset	Method	Error ± std	Avg Time(s)
CIFAR10	Base	1.210 ± 0.02	0.94
	n-Diag-Rich	**0.089** ± 0.03	2.85
	$\frac{n}{4}$-Diag-Rich	0.201 ± 0.03	0.89
	Eps-Rich	1.007 ± 0.02	1.56
Fashion MNIST	Base	0.852 ± 0.02	0.84
	n-Diag-Rich	**0.259** ± 0.02	2.89
	$\frac{n}{4}$-Diag-Rich	0.355 ± 0.04	0.91
	Eps-Rich	0.849 ± 0.02	1.26
MNIST	Base	0.788 ± 0.01	0.78
	n-Diag-Rich	**0.219** ± 0.02	3.30
	$\frac{n}{4}$-Diag-Rich	0.420 ± 0.04	0.95
	Eps-Rich	0.706 ± 0.01	1.40

Results. Table 1 confirms that incorporating intrinsic geometry is essential for accurate estimation. Standard ε-extrapolation fails to address the statistical bias dominating in high dimensions, yielding errors comparable to the baseline, even when using

$\widehat{d}^*$. In contrast, our diagonal estimator, calibrated via $\widehat{d}^*$, reduces the error by an order of magnitude (e.g., $1.21 \to 0.09$ on CIFAR10). This validates that coupling the regularization schedule to the intrinsic dimension ($\varepsilon \asymp n^{-1/(\widehat{d}^*+4)}$) effectively mitigates the discretization bottleneck. Furthermore, even with a reduced budget (reduced sample size $\frac{n}{4}$ - bagging $K = 12$), our method significantly outperforms baselines while matching their computational cost.

6. Conclusion and Future Work

In this work, we have established that explicitly estimating the intrinsic dimension of data is key to overcoming the statistical limits of discrete Optimal Transport. We showed that our Diagonal Richardson estimator, calibrated via a solver-free intrinsic dimension estimate, significantly improves the accuracy of Wasserstein distance estimation on high-dimensional manifolds. This framework opens several directions for future research. First, the validity of our bias cancellation relies on the sharpness of the statistical error expansion (Assumption 5.2), a property that warrants further theoretical analysis. Second, while we focused on the scalar transport cost, extending this ”geometry-aware” debiasing strategy and analysis to Kantorovich potentials and gradients remains a promising avenue for improving training stability in generative modeling.

References

An, D., Guo, Y., Lei, N., Luo, Z., Yau, S.-T., and Gu, X. Ae-ot: A new generative model based on extended semi-discrete optimal transport. *ICLR 2020*, 2019.

Bac, J., Mirkes, E. M., Gorban, A. N., Tyukin, I., and Zinovyev, A. Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021.

Blanchet, J., Jambulapati, A., Kent, C., and Sidford, A. Towards optimal running times for optimal transport. *Operations Research Letters*, 2023.

Block, A., Jia, Z., Polyanskiy, Y., and Rakhlin, A. Intrinsic dimension estimation using wasserstein distance. *Journal of Machine Learning Research*, 23(313):1–37, 2022.

Chen, Y., Telgarsky, M., Zhang, C., Bailey, B., Hsu, D., and Peng, J. A gradual, semi-discrete approach to generative network training via explicit wasserstein minimization. In *International Conference on Machine Learning*, pp. 1071–1080. PMLR, 2019.

Chewi, S., Niles-Weed, J., and Rigollet, P. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3, 2024.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

Courty, N., Flamary, R., and Tuia, D. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pp. 274–289. Springer, 2014.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances In Neural Information Processing Systems*, 26, 2013.

Dudley, R. M. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn’s algorithm. In *International Conference On Machine Learning*, pp. 1367–1376, 2018.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019.

Fournier, N. and Guillin, A. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. Limit theorems for entropic optimal transport maps and sinkhorn divergence. *Electronic Journal of Statistics*, 18(1):980–1041, 2024.

Groppe, M. and Hundrieser, S. Lower complexity adaptation for empirical entropic optimal transport. *Journal of Machine Learning Research*, 25(344):1–55, 2024.

Hundrieser, S., Staudt, T., and Munk, A. Empirical optimal transport between different measures adapts to lower complexity. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 60, pp. 824–846. Institut Henri Poincaré, 2024.

Hütter, J.-C. and Rigollet, P. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2), 2021.

Kitagawa, J., Mérigot, Q., and Thibert, B. A newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21, 03 2016. doi: 10.4171/JEMS/889.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

Li, Z., Li, S., Wang, Z., Lei, N., Luo, Z., and Gu, D. X. Dpm-ot: a new diffusion probabilistic model based on optimal transport. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 22624–22633, 2023.

Lin, T., Ho, N., and Jordan, M. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pp. 3982–3991. PMLR, 2019.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample-variance penalization. In *22nd Annual Conference on Learning Theory (COLT)*, 2009.

Mena, G. and Niles-Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32, 2019.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Pooladian, A.-A., Divol, V., and Niles-Weed, J. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *International Conference on Machine Learning*, pp. 28128–28150. PMLR, 2023.

Ramdas, A., García Trillos, N., and Cuturi, M. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Stromme, A. J. Minimum intrinsic dimension scaling for entropic optimal transport. In *International Conference on Soft Methods in Probability and Statistics*, pp. 491–499. Springer, 2024.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25, 06 2017. doi: 10.3150/18-BEJ1065.

A. Intrinsic Dimension and Wasserstein Distances

We recall some notions and properties related to the intrinsic dimension and multi-resolution behavior of the Wasserstein dimension and Sinkhorn divergences, which are present in (Weed & Bach, 2017; Stromme, 2024).

A.1. Notions.

Definition (δ-covering number of a set). Given a set $S \subseteq X$, the δ-covering number of S, denoted $\mathcal{N}_\delta(S)$, is the minimum m such that there exists m closed balls $B_1, \dots, B_m$ of diameter δ such that $S \subseteq \bigcup_{1 \leq i \leq m} B_i$. The δ-dimension of S is the quantity

$$d_\delta(S) := \frac{\log \mathcal{N}_\delta(S)}{-\log \delta}\,.$$

The following definition is particularly useful when working with measures instead of sets, allowing one to ignore a small fraction of the mass. The following definition appears first in (Dudley, 1969).

Definition ((δ,τ)-covering number of a measure). Given a measure μ on X, the (δ,τ)-covering number is

$$\mathcal{N}_\delta(\mu,\tau) := \inf\{\mathcal{N}_\delta(S) : \mu(S) \geq 1-\tau\}$$

and the (δ,τ)-dimension is

$$d_\delta(\mu,\tau) := \frac{\log \mathcal{N}_\delta(\mu,\tau)}{-\log \delta}\,.$$

The limits are defined by

$$\mathcal{N}_\varepsilon(\mu) := \mathcal{N}_\varepsilon(\mu, 0)\,, \quad d_\varepsilon(\mu) := d_\varepsilon(\mu,0)\,.$$

Definition (Wasserstein dimensions). The upper and lower Wasserstein dimensions are respectively

$$\begin{aligned} d_p^*(\mu) &= \inf\left\{ s \in (2p,\infty) : \limsup_{\delta\to 0} d_\delta\left(\mu, \delta^{\frac{sp}{s-2p}}\right) \leq s \right\}, \\ d_*(\mu) &= \lim_{\tau\to 0} \liminf_{\delta\to 0} d_\delta(\mu,\tau)\,. \end{aligned}$$

A.2. Properties

Proposition A.1 (Prop.5 and Corollary 1 (Weed & Bach, 2017))**.** *Let $p \in [1,\infty)$. Suppose there exist $\delta' \leq 1$ and $s > 2p$ such that $d_\delta(\mu, \delta^{\frac{sp}{s-2p}}) \leq s$ for all $\varepsilon \leq \delta'$. Then there exist constants C_1, C_2 such that:*

$$\mathbb{E}[W_p^p(\mu,\widehat{\mu}_n)] \leq C_1 n^{-p/s} + C_2 n^{-1/2}. \tag{13}$$

In particular, $\mathbb{E}[W_p(\mu,\widehat{\mu}_n)] \lesssim n^{-1/s}$. As a corollary, if $s > d_p^*(\mu)$, we have $\mathbb{E}[W_p(\mu,\widehat{\mu}_n)] \lesssim n^{-1/s}$.

We now derive a slightly different proposition, based on Proposition 6 of (Weed & Bach, 2017), with a similar proof.

Proposition A.2 (Local Lower Bound for Wasserstein Rates)**.** *Let μ be a probability measure on $\mathbb{R}^D$ and let $p \in [1,\infty)$. Suppose there exist constants $\tau \in (0,1)$, $A > 0$, $s > 0$, and a scale interval $[a,b] \subset (0,1]$ such that the (δ,τ)-packing number satisfies*

$$\mathcal{N}'_\delta(\mu,\tau) \geq A\delta^{-s}, \quad \forall \delta \in [a,b]. \tag{14}$$

If the sample size n satisfies the validity condition $Ab^{-s} < n < Aa^{-s}$, then for any discrete measure ν supported on at most n points (including the empirical measure $\widehat{\mu}_n$), we have the following lower bound:

$$W_p^p(\mu,\nu) \geq \left(\tau 4^{-p} A^{p/s}\right) n^{-p/s}. \tag{15}$$

Proof. Let n satisfy the validity condition. We define the critical radius $\varepsilon_n := (n/A)^{-1/s}$. By the condition $Ab^{-s} < n < Aa^{-s}$, it follows explicitly that $a < \varepsilon_n < b$.

Consider an arbitrary $\varepsilon < \varepsilon_n$ such that $\varepsilon \in [a, b]$. By the packing assumption, the number of disjoint balls of radius ε that can be packed into the support of μ (capturing mass $1 - \tau$) satisfies $N := \mathcal{N}'_\varepsilon(\mu, \tau) \geq A\varepsilon^{-s}$. Since $\varepsilon < \varepsilon_n$, we strictly have

$$N \geq A\varepsilon^{-s} > A\varepsilon_n^{-s} = n. \tag{16}$$

Let $\{B_1, \dots, B_N\}$ denote this packing of disjoint balls. Since ν is supported on at most n points, by the Pigeonhole Principle, at least $N - n$ of these balls do not contain any support points of ν. Let $\mathcal{I}_{\text{empty}}$ denote the indices of these empty balls.

The Wasserstein distance $W_p^p(\mu, \nu)$ involves transporting the mass from these empty balls to the support of ν. For any ball B_i with $i \in \mathcal{I}_{\text{empty}}$, the mass $\mu(B_i)$ must be moved a distance of at least ε to reach any point in $\text{supp}(\nu)$ (which lies outside B_i). Following the sharp derivation in Weed & Bach (2017) (Proposition 6), the aggregated transport cost is lower bounded by:

$$W_p^p(\mu, \nu) \geq \tau \left(\frac{\varepsilon}{4}\right)^p = \tau 4^{-p} \varepsilon^p. \tag{17}$$

Since this inequality holds for any $\varepsilon < \varepsilon_n$, we take the limit $\varepsilon \to \varepsilon_n$. Substituting $\varepsilon_n = (n/A)^{-1/s}$ yields:

$$W_p^p(\mu, \nu) \geq \tau 4^{-p} \left[(n/A)^{-1/s}\right]^p = \tau 4^{-p} A^{p/s} n^{-p/s}. \tag{18}$$

□

A.3. Minimum Intrinsic Dimension scaling

We recall here the *Minimum Intrinsic Dimension* scaling of the Sinkhorn divergence, which we leverage for the Wasserstein distance estimation.

Theorem (MID scaling for the Sinkhorn divergences). For numerical constants independent of all problem parameters,

$$\mathbb{E}[|S_\varepsilon(\widehat{\mu}, \widehat{\nu}) - S_\varepsilon(\mu, \nu)|] \lesssim (1 + \varepsilon) \sqrt{\frac{\mathcal{N}(\mu, \frac{\varepsilon}{L}) \wedge \mathcal{N}(\nu, \frac{\varepsilon}{L})}{n}}.$$

The dimensional quantity in this estimate is contained in the minimum covering numbers at scale ε, demonstrating the MID scaling phenomenon. Here, the constant L refers to the Lipschitz bound on the cost, which is assured to be finite for $\|\cdot\|^p$ costs, in the bounded setting.

Connection to Intrinsic Dimension. In this work, we characterize the intrinsic dimension at a fixed scale through the convergence rate of $\mathbb{E}[W_1(\mu, \widehat{\mu}_n)]$ and its corresponding lower bounds. We now establish the link between this characterization and the MID scaling bound for Sinkhorn divergences presented above.

Consider the cost function $c = \|\cdot\|$, which implies a Lipschitz constant $L = 1$. Suppose that over a scale interval $\varepsilon \in [\varepsilon_a, \varepsilon_b]$, the measure is sufficiently regular such that the dimension is stable across mass thresholds; specifically, $d_\varepsilon(\mu, \tau) = d_\varepsilon(\mu, 0)$ for all $\tau \in (0, 1/2]$. Under these conditions, Propositions A.1 and A.2 provide matching upper and lower bounds on the Wasserstein distance convergence rate with $d_{\text{int}} = d_\varepsilon$ for a range $n \in [N_{\min}, N_{\max}]$. This allows us to identify the intrinsic dimension d_{int} directly with the scale-dependent covering dimension d_ε, which is precisely the quantity governing the MID scaling in Theorem 2.

B. Proof of Theorem 4.2: Dimension Estimation Control

Proof. Fix $n \geq n_{\min}$ and $\eta > 1$. For each $k \in \{n, \eta n\}$, we introduce the notations

$$W_k^* := \text{OT}_c(\rho, \widehat{\rho}_k^*), \qquad \widehat{W}_k := \widehat{\text{OT}}_N^k = \frac{1}{N} \sum_{i=1}^N \min_{1 \leq j \leq k} c(X_i, x_j),$$

where $x_1, \dots, x_k \overset{\text{i.i.d.}}{\sim} \rho$ define the support of $\widehat{\rho}_k^*$ and $X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} \rho$ are independent Monte Carlo samples (independent of the x_j's). Set $c(\cdot) = \|\cdot\|$ and notice that $\text{Diam}(\rho) := \sup_{x,y \in \text{Supp}(\rho)} c(x, y)$.

Monte Carlo concentration for $\widehat{W}_k$. Conditionally on $(x_1, \ldots, x_k)$, the random variables $\min_{j \le k} c(X_i, x_j)$ are i.i.d. and lie in $[0, \mathrm{Diam}(\rho)]$. Hence, Hoeffding's inequality gives, for any $t > 0$,

$$\mathbb{P}\left(|\widehat{W}_k - W_k^*| \ge t\right) \le 2\exp\left(-\frac{2Nt^2}{\mathrm{Diam}(\rho)^2}\right).$$

Define the Monte Carlo accuracy level

$$\kappa_N := \mathrm{Diam}(\rho)\sqrt{\frac{\log(8/\delta)}{2N}}.$$

Then $\mathbb{P}(|\widehat{W}_k - W_k^*| > \kappa_N) \le \delta/4$ for each k. By a union bound over $k \in \{n, \eta n\}$, the event

$$\mathcal{E}_{\mathrm{MC}} := \left\{|\widehat{W}_n - W_n^*| \le \kappa_N\right\} \cap \left\{|\widehat{W}_{\eta n} - W_{\eta n}^*| \le \kappa_N\right\}$$

satisfies

$$\mathbb{P}(\mathcal{E}_{\mathrm{MC}}) \ge 1 - \delta/2.$$

Two-sided control of W_k^* at scales n and ηn. From the bounds provided in Assumption 4.1, we have for all $k \in [n_{\min}, n_{\max}/\eta]$,

$$W_k^* \ge C_1 k^{-1/d_{\mathrm{int}}} \qquad \text{a.s.} \tag{19}$$

and for any $t > 0$, using the boundness of the semi-dual optimize $\mathbf{0}_c \le \mathrm{Diam}(\rho)$,

$$\mathbb{P}\left(W_k^* \ge C_2 k^{-1/d_{\mathrm{int}}} + t\right) \le \exp\left(-\frac{2kt^2}{\mathrm{Diam}(\rho)^2}\right),$$

where $\mathrm{Diam}(\rho)$ is the diameter of $\mathrm{Supp}(\rho)$. Taking $t = \xi C_2 k^{-1/d_{\mathrm{int}}}$ for $\xi > 0$ yields

$$\mathbb{P}\left(W_k^* \ge (1+\xi) C_2 k^{-1/d_{\mathrm{int}}}\right) \le \exp\left(-\frac{2k\,\xi^2 C_2^2\, k^{-2/d_{\mathrm{int}}}}{\mathrm{Diam}(\rho)^2}\right) = \exp\left(-\frac{2\xi^2 C_2^2}{\mathrm{Diam}(\rho)^2}\, k^{1-2/d_{\mathrm{int}}}\right).$$

For $k^{1-2/d_{\mathrm{int}}} \ge \max\left\{{n^*}^{1-2/d_{\mathrm{int}}}, \frac{\mathrm{Diam}(\rho)^2}{2\xi^2 C_2^2}\log(4/\delta)\right\}$, the event

$$\mathcal{E}_{\mathrm{Q}} := \{W_k^* \le (1+\xi) C_2 k^{-1/d_{\mathrm{int}}}\} \cap \{W_{\eta k}^* \le (1+\xi) C_2 (\eta k)^{-1/d_{\mathrm{int}}}\}$$

satisfies $\mathbb{P}(\mathcal{E}_{\mathrm{Q}}) \ge 1 - \delta/2$. On $\mathcal{E}_{\mathrm{Q}}$ and using (19),

$$C_1 k^{-1/d_{\mathrm{int}}} \le W_k^* \le (1+\xi) C_2 k^{-1/d_{\mathrm{int}}}, \qquad k \in \{n, \eta n\}. \tag{20}$$

Let $\mathcal{E} := \mathcal{E}_{\mathrm{MC}} \cap \mathcal{E}_{\mathrm{Q}}$. Then $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$.

Control of the population log-ratio D. Define

$$D := \log W_n^* - \log W_{\eta n}^*.$$

From (20),

$$\frac{W_n^*}{W_{\eta n}^*} \in \left[\frac{C_1 n^{-1/d_{\mathrm{int}}}}{(1+\xi) C_2 (\eta n)^{-1/d_{\mathrm{int}}}},\ \frac{(1+\xi) C_2 n^{-1/d_{\mathrm{int}}}}{C_1 (\eta n)^{-1/d_{\mathrm{int}}}}\right] = \left[\frac{C_1}{(1+\xi) C_2}\eta^{1/d_{\mathrm{int}}},\ \frac{(1+\xi) C_2}{C_1}\eta^{1/d_{\mathrm{int}}}\right].$$

Taking logs and denoting $D_0 := \log \eta / d_{\mathrm{int}}$ gives

$$D \in \left[D_0 - b,\ D_0 + b\right], \qquad b := \log\left(\frac{(1+\xi) C_2}{C_1}\right) \ge 0,$$

hence $|D - D_0| \le b$. Moreover, the stated condition on η implies $b \le \frac{\gamma}{2} D_0$.

Stability of $\log \widehat{W}_k$ under Monte Carlo error. On $\mathcal{E}$, we have $|\widehat{W}_k - W_k^*| \le \kappa_N$. We first ensure positivity of $\widehat{W}_{\eta n}$ (and similarly $\widehat{W}_n$) by requiring

$$\kappa_N \le \frac{1}{2} W_{\eta n}^*.$$

Using $W_{\eta n}^* \ge C_1 (\eta n)^{-1/d_{\text{int}}}$ from (19), it is enough that

$$\kappa_N \le \frac{1}{2} C_1 (\eta n)^{-1/d_{\text{int}}},$$

which is implied by the second lower bound on N in the theorem.

Assume this condition holds. Then for $k \in \{n, \eta n\}$,

$$\left| \log \widehat{W}_k - \log W_k^* \right| = \left| \log \left(1 + \frac{\widehat{W}_k - W_k^*}{W_k^*} \right) \right|.$$

Since $|\widehat{W}_k - W_k^*| \le \kappa_N \le \frac{1}{2} W_k^*$, we have $\left| \frac{\widehat{W}_k - W_k^*}{W_k^*} \right| \le \frac{1}{2}$ and thus $|\log(1+u)| \le 2|u|$ for $|u| \le 1/2$ yields

$$\left| \log \widehat{W}_k - \log W_k^* \right| \le 2 \frac{|\widehat{W}_k - W_k^*|}{W_k^*} \le 2 \frac{\kappa_N}{W_k^*}.$$

Therefore, with $\widehat{D} := \log \widehat{W}_n - \log \widehat{W}_{\eta n}$,

$$|\widehat{D} - D| \le 2\kappa_N \left(\frac{1}{W_n^*} + \frac{1}{W_{\eta n}^*} \right). \tag{21}$$

Using $W_n^* \ge C_1 n^{-1/d_{\text{int}}}$ and $W_{\eta n}^* \ge C_1 (\eta n)^{-1/d_{\text{int}}}$,

$$2\kappa_N \left(\frac{1}{W_n^*} + \frac{1}{W_{\eta n}^*} \right) \le \frac{2\kappa_N}{C_1} n^{1/d_{\text{int}}} \left(1 + \eta^{1/d_{\text{int}}} \right) \le \frac{4\kappa_N}{C_1} \eta^{1/d_{\text{int}}} n^{1/d_{\text{int}}}.$$

Plugging $\kappa_N = \mathrm{Diam}(\rho) \sqrt{\log(8/\delta)/(2N)}$, we get

$$|\widehat{D} - D| \le \frac{4\mathrm{Diam}(\rho)}{C_1} \eta^{1/d_{\text{int}}} n^{1/d_{\text{int}}} \sqrt{\frac{\log(8/\delta)}{2N}}.$$

The first lower bound on N in the theorem ensures that the RHS is at most $\frac{\gamma}{2} D_0$.

Denominator interval and propagation to $\widehat{d}^*$. On $\mathcal{E}$ we have

$$|\widehat{D} - D_0| \le |\widehat{D} - D| + |D - D_0| \le \frac{\gamma}{2} D_0 + \frac{\gamma}{2} D_0 = \gamma D_0,$$

so $(1-\gamma) D_0 \le \widehat{D} \le (1+\gamma) D_0$. Finally, we use the relation $d_N^* = \log \eta / \widehat{D}$ to obtain

$$\frac{\log \eta}{(1+\gamma) D_0} \le \widehat{d}^* \le \frac{\log \eta}{(1-\gamma) D_0} \quad \Longleftrightarrow \quad \frac{d_{\text{int}}}{1+\gamma} \le \widehat{d}^* \le \frac{d_{\text{int}}}{1-\gamma}.$$

Since $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, this completes the proof.

Remark: Observe that, for the proof to hold, we required the condition $\eta n \le n_{\max}$ to hold. □

B.1. Intrinsic Dimension estimation under the manifold hypothesis

We briefly discuss here, how our proof stands under the manifold hypothesis assumption, as in (Block et al., 2022). We recall here their assumption.

Assumption B.1 (Geometric regularity)**.** Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact manifold of dimension $d_{\mathcal{M}}$ with diameter $\Delta = \operatorname{diam}(\mathcal{M})$ and reach $\tau > 0$. Furthermore, suppose that the data-generating distribution ρ admits a density with respect to the uniform measure on $\mathcal{M}$, bounded as $0 < w_{\min} \leq \frac{\mathrm{d}\rho}{\mathrm{dvol}_{\mathcal{M}}} \leq w_{\max} < \infty$.

The positive reach τ guarantees that $\mathcal{M}$ is free of self-intersections and has controlled curvature; informally, below the scale τ the nearest-point (normal) projection onto $\mathcal{M}$ is single-valued, so $\mathcal{M}$ behaves locally like $\mathbb{R}^{d_{\mathcal{M}}}$. Formally, the reach of $\mathcal{M}$ is

$$\operatorname{reach}(\mathcal{M}) := \sup\Big\{r > 0 \ : \ \forall z \in \mathbb{R}^D,\ \operatorname{dist}(z, \mathcal{M}) < r \ \Rightarrow \ \underset{y \in \mathcal{M}}{\operatorname{argmin}} \, \|z - y\| \text{ is unique}\Big\}.$$

This notion is standard in geometric inference, and the assumption $\tau > 0$ holds, for instance, if $\mathcal{M} \subset \mathbb{R}^d$ is a compact, embedded $C^{1,1}$ submanifold. We refer to (Block et al., 2022) for more details on the notion of reach and its role in OT-based intrinsic-dimension estimation. Here, we note $\omega_{d_{\mathcal{M}}}$ the volume of the unit d-dimensional ball.

Proposition B.2 (((Block et al., 2022))Upper and lower bound on the Wasserstein error under the manifold hypothesis)**.** *For*

$$n > \frac{d_{\mathcal{M}} \operatorname{vol} \mathcal{M}}{4\omega_{d_{\mathcal{M}}} w_{\min}} \left(\frac{\tau}{8}\right)^{-d_{\mathcal{M}}}.$$

we have, for any measure σ_n of n points

$$W_1(\sigma_n, \rho) \geq \frac{1}{32} \left(\frac{d_{\mathcal{M}} \operatorname{vol} \mathcal{M}}{4\omega_{d_{\mathcal{M}}} w_{\min}}\right)^{\frac{1}{d_{\mathcal{M}}}} n^{-\frac{1}{d_{\mathcal{M}}}}.$$

Also, for a constant $C \leq 48$, we have

$$\mathbb{E}[W_1(\rho_n, \rho)] \leq C \left(\frac{\operatorname{vol} \mathcal{M}}{n w_{\min}}\right)^{\frac{1}{d}} \sqrt{\log\left(\frac{n w_{\min} \operatorname{Diam}(\rho)^{d_{\mathcal{M}}}}{d_{\mathcal{M}} \operatorname{vol}_{\mathcal{M}}}\right)}.$$

Using Proposition B.2, we thus can apply our Theorem 4.2, with the constants: $d_{\text{int}} = d_{\mathcal{M}}$, $n_{\min} = \frac{d_{\text{int}} \operatorname{vol} \mathcal{M}}{4w\omega_{\min}} \left(\frac{\tau}{8}\right)^{-d_{\text{int}}}$, $n_{\max} = +\infty$, $C_1 = \frac{1}{32}\left(\frac{d_{\text{int}} \operatorname{vol} \mathcal{M}}{4w\omega_{\min}}\right)$, and $C_2 = C\left(\frac{\operatorname{vol} \mathcal{M}}{\omega_{\min}}\right)^{\frac{1}{d_{\text{int}}}} \sqrt{\log\left(\frac{n\omega_{\min} \operatorname{Diam}(\rho)^{d_{\text{int}}}}{d_{\text{int}} \operatorname{vol}_{\mathcal{M}}}\right)}$.

C. Proof of Proposition 5.3: Convergence of Diagonal Richardson

Proof. We analyze the bias and stochastic deviation separately.

Let the regularization parameter scale as $\varepsilon_n \asymp n^{-a}$. Based on standard expansions and our assumption that the statistical bound holds with equality, we have:

$$\text{Bias}_{\text{ent}}(n) = C_{\text{ent}} n^{-2a} + o(n^{-2a}) \,, \tag{22}$$

$$\text{Bias}_{\text{stat}}(n) = C_{\text{stat}} n^{-(1-a d_{\text{int}})/2} + o(n^{-(1-a d_{\text{int}})/2}) \,. \tag{23}$$

To achieve the optimal convergence rate, we balance the leading order terms by equating their exponents:

$$2a = \frac{1 - a d_{\text{int}}}{2} \iff 4a = 1 - a d_{\text{int}} \iff a(d_{\text{int}} + 4) = 1.$$

This yields the optimal decay choice $a = \frac{1}{d_{\text{int}} + 4}$. Under this choice, both bias terms decay at the common rate γ:

$$\gamma := 2a = \frac{1 - a d_{\text{int}}}{2} = \frac{2}{d + 4}.$$

Consequently, the total expectation admits the expansion:

$$\mathbb{E}[\widehat{S}_{\varepsilon_n, n}] = W_2^2 + \underbrace{(C_{\text{ent}} + C_{\text{stat}})}_{C_{\text{total}}} n^{-\gamma} + o(n^{-\gamma}). \tag{24}$$

The diagonal Richardson estimator is defined as $\widehat{R}_{2n}^{\text{Diag}} := w_{2n}\widehat{S}_{\varepsilon_{2n},2n} + w_n\widehat{S}_{\varepsilon_n,n}$. By linearity, its expected value is:

$$\mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}] = w_{2n}\mathbb{E}[\widehat{S}_{\varepsilon_{2n},2n}] + w_n\mathbb{E}[\widehat{S}_{\varepsilon_n,n}].$$

Substituting the expansion from (24) separately for the entropic and statistical terms:

$$\begin{aligned}\mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}] = W_2^2&(w_{2n} + w_n)\\ &+ C_{\text{ent}}\left(w_{2n}(2n)^{-\gamma} + w_n n^{-\gamma}\right) \quad \text{(Entropic First Order)}\\ &+ C_{\text{stat}}\left(w_{2n}(2n)^{-\gamma} + w_n n^{-\gamma}\right) \quad \text{(Statistical First Order)}\\ &+ o(n^{-\gamma}).\end{aligned}$$

The weights are chosen as $w_{2n} = \frac{2^\gamma}{2^\gamma - 1}$ and $w_n = \frac{-1}{2^\gamma-1}$. We verify the cancellation conditions:

1. $w_{2n} + w_n = \frac{2^\gamma - 1}{2^\gamma - 1} = 1.$
2. $w_{2n}(2n)^{-\gamma} + w_n = \frac{2^\gamma \cdot 2^{-\gamma} - 1}{2^\gamma - 1} = \frac{1-1}{2^\gamma-1} = 0.$

Because both biases scale with $n^{-\gamma}$, the *same* weights eliminate the leading terms of both C_{ent} and C_{stat} simultaneously. Thus:

$$\left|\mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}] - W_2^2\right| = o(n^{-\gamma}). \tag{25}$$

Using Proposition 4 from (Chizat et al., 2020), we have $\mathbb{P}\left[|\widehat{S}_{\varepsilon,n} - \mathbb{E}[\widehat{S}_{\varepsilon,n}]| \geq t\right] \leq 2\exp(-nt^2/D^24)$, therefore, noting $Z_n = \widehat{S}_{\varepsilon_n,n} - \mathbb{E}[\widehat{S}_{\varepsilon_n,n}]$, we have

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{R}_{2n}^{\text{Diag}} - \mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}]\right|\right] &\leq |w_{2n}|\mathbb{E}[|Z_{2n}|] + |w_n|\mathbb{E}[|Z_n|]\\ &\leq |w_{2n}|\sqrt{\frac{2\pi D^2}{n}} + |w_n|\sqrt{\frac{4\pi D^2}{n}} = O(n^{-1/2}).\end{aligned}$$

Combining the residual bias and stochastic deviation bounds, we conclude

$$\mathbb{E}\left[|\widehat{R}_{2n}^{\text{Diag}} - W_2^2|\right] \leq o(n^{-\gamma}) + O(n^{-1/2}) = o(n^{-\frac{2}{d_{\text{int}}+4}}).$$

□

Remark: Robustness to Rate Mismatch and Small Sample Efficiency. While the proof above assumes perfect balancing ($\gamma_{\text{ent}} = \gamma_{\text{stat}} = \gamma$), the diagonal Richardson estimator remains highly effective even if the statistical bias decays at a slightly faster rate $\beta > \gamma$ (making it asymptotically negligible compared to entropic bias). This is particularly relevant in the small sample regime (n small), where "asymptotically better" terms may still possess large constants.

Let the statistical bias be of the form $\text{Bias}_{\text{stat}}(n) = C_{\text{stat}} n^{-\beta}$. The statistical first order bias after extrapolation becomes:

$$w_{2n}C_{\text{stat}}(2n)^{-\beta} + w_n C_{\text{stat}} n^{-\beta} = C_{\text{stat}} n^{-\beta}\left(w_{2n}2^{-\beta} + w_n\right).$$

Substituting the weights $w_{2n} = \frac{2^\gamma}{2^\gamma-1}$ and $w_n = \frac{-1}{2^\gamma-1}$, we define the *Bias Reduction Factor* $\rho(\beta, \gamma)$:

$$\rho(\beta,\gamma) := \frac{w_{2n}2^{-\beta} + w_n}{1} = \frac{2^{\gamma-\beta} - 1}{2^\gamma - 1}. \tag{26}$$

We observe two key properties:

1. **Exact Cancellation:** If $\beta = \gamma$, then $\rho(\gamma, \gamma) = 0$, recovering the main theorem.

2. **Local Continuity (Small n Regime):** In optimal transport, the entropic rate $\gamma = 2a$ and statistical rate $\beta = (1 - ad_{\text{int}})/2$ are algebraically coupled. For any choice of a near the optimal value $\frac{1}{d_{\text{int}}+4}$, the difference $\delta = \beta - \gamma$ is small. A Taylor expansion around $\delta \approx 0$ yields:
$$\rho(\beta, \gamma) \approx \frac{-\delta \ln 2}{2^\gamma - 1}.$$
Thus, even if the rates are not identical, the estimator suppresses the statistical bias by a factor proportional to the rate mismatch δ for computationally tractable n (where $n^{-\beta}$ is not yet negligible). This further shows that this method is also robust to our estimation of d_{int}.

In the same way, we observe that Richardson is robust to small oscillations of C_{stat}.

Remark: different bias conjecture Recent results by Goldfeld et al. (2024) establish a Central Limit Theorem for the Sinkhorn Divergence, noting that the expected finite-sample error scales as $\mathbb{E}[S_{\varepsilon,n}] - S_\varepsilon = \mathcal{O}(1/\sqrt{n} + n^{-1} r_\varepsilon)$ (see their Remark 7). The factor r_ε depends on the variance of the Sinkhorn potentials. If we conjecture the scaling $r_\varepsilon \asymp \varepsilon^{-(d_{\text{int}}+2)}$, coming from Sinkhorn second derivative, and the statistical bias becomes $\mathcal{O}(n^{-1}\varepsilon^{-(d_{\text{int}}+2)})$. Balancing this against the entropic bias $O(\varepsilon^2)$ yields the optimal decay choice $\varepsilon \asymp n^{-1/(d_{\text{int}}+4)}$. Notably, this recovers the exact same exponent $a = \frac{1}{d_{\text{int}}+4}$ derived under the standard $\sqrt{n}$-rate assumption, suggesting that our diagonal Richardson strategy is structurally robust to different statistical regimes.

D. Proof of Proposition 5.4: Variance Reduction with Bagging

Proof. Let $\mathbf{D}_{2N} = \{(X_i, Y_i)\}_{i=1}^{2N}$ be a dataset of $2N$ i.i.d. samples drawn from the joint distribution of (μ, ν). We generate K bags, where each bag corresponds to a subsample index set $\mathcal{S}_k$ of size $|\mathcal{S}_k| = N$, drawn uniformly at random with replacement.

We analyze the asymptotic variance using the Law of Total Variance, conditioning on the observed dataset $\mathbf{D}_{2N}$. The total variance decomposes as:
$$\mathrm{Var}(\widehat{R}_{K,N}) = \underbrace{\mathrm{Var}\left(\mathbb{E}[\widehat{R}_{K,N} \mid \mathbf{D}_{2N}]\right)}_{\text{Term I: Dataset Variance}} + \underbrace{\mathbb{E}\left[\mathrm{Var}(\widehat{R}_{K,N} \mid \mathbf{D}_{2N})\right]}_{\text{Term II: Bagging Variance}}. \tag{27}$$

Asymptotic Linear Expansion.

We first use that the empirical Sinkhorn Divergence is second-order Hadamard differentiable (Goldfeld et al., 2024). Therefore, we can use the second order Taylor expansion. The first-order term L_ε is linear and can be rewritten such as
$$L_\varepsilon = \frac{1}{2N} \sum_{i=1}^{2N} \phi_\varepsilon(Z_i)$$
where ϕ_ε is the so-called influence function.The first-order von Mises expansion gives:
$$\widehat{S}_{\varepsilon,2N} = S_\varepsilon + \frac{1}{2N} \sum_{i=1}^{2N} \phi_\varepsilon(Z_i) + r_{2N}(\varepsilon), \tag{28}$$
where E is the population value, $Z_i = (X_i, Y_i)$, and ϕ is the canonical influence function with $\mathbb{E}[\phi(Z)] = 0$ and $\mathrm{Var}(\phi_\varepsilon(Z)) = \sigma_\varepsilon^2$.

Similarly, for any subsample $\mathcal{S}_k$ of size N, the estimator expands as:
$$\widehat{S}_{\varepsilon,N}^{(k)} = S_\varepsilon + \frac{1}{N} \sum_{j \in \mathcal{S}_k} \phi_\varepsilon(Z_j) + r_N^{(k)}(\varepsilon). \tag{29}$$

Analysis of Term I (Dataset Variance). We analyze the conditional expectation $\mathbb{E}[\widehat{R}_{K,N} \mid \mathbf{D}_{2N}]$ by substituting the von Mises expansions. First, recall that for the subsample mean term, the linearity of expectation implies:
$$\mathbb{E}\left[\frac{1}{M} \sum_{j \in \mathcal{S}_k} \phi_\varepsilon(Z_j) \,\middle|\, \mathbf{D}_{2N}\right] = \sum_{i=1}^{N} \phi_\varepsilon(Z_i) \cdot \mathbb{P}(i \in \mathcal{S}_k) \cdot \frac{1}{M} = \frac{1}{N} \sum_{i=1}^{N} \phi_\varepsilon(Z_i). \tag{30}$$

Now, substituting the explicit expansions $\widehat{S}_{\varepsilon,N} = S_\varepsilon + \bar{\phi}_{\varepsilon N} + r_N(\varepsilon)$ (where $\bar{\phi}_{\varepsilon N} = \frac{1}{N}\sum \phi_\varepsilon(Z_i)$) and $\widehat{S}^{(k)}_{\varepsilon,N} = S_\varepsilon + \bar{\phi}_{\varepsilon M,k} + r_N^{(k)}(\varepsilon)$ into the definition of the estimator:

$$\begin{aligned}
\mathbb{E}[\widehat{R}_{K,N} \mid \mathbf{D}_{2N}] &= (1+\lambda)\widehat{S}_{\varepsilon,N} - \lambda \mathbb{E}[\widehat{S}^{(1)}_{\varepsilon,N} \mid \mathbf{D}_{2N}] & (31)\\
&= (1+\lambda)(E + \bar{\phi}_{\varepsilon N} + r_N) - \lambda\left(E + \mathbb{E}[\bar{\phi}_{\varepsilon M,1} \mid \mathbf{D}_{2N}] + \mathbb{E}[r_{M,1} \mid \mathbf{D}_{2N}]\right) & (32)\\
&= (1+\lambda)(E + \bar{\phi}_{\varepsilon N} + r_N) - \lambda\left(E + \bar{\phi}_{\varepsilon N} + \mathbb{E}[r_{M,1} \mid \mathbf{D}_{2N}]\right). & (33)
\end{aligned}$$

Grouping the terms by order:

$$\begin{aligned}
\mathbb{E}[\widehat{R}_{K,2N} \mid \mathbf{D}_{2N}] &= \underbrace{[(1+\lambda)-\lambda]\,E}_{\text{Bias}} + \underbrace{[(1+\lambda)-\lambda]\,\bar{\phi}_{\varepsilon 2N}}_{\text{Linear Term}} + \underbrace{((1+\lambda)r_{2N} - \lambda\mathbb{E}[r_{N,1} \mid \mathbf{D}_{2N}])}_{\text{Pooled Remainder } \rho_{2N}} & (34)\\
&= E + \frac{1}{2N}\sum_{i=1}^{2N} \phi_\varepsilon(Z_i) + \rho_{2N}. & (35)
\end{aligned}$$

We seek the variance of this quantity. Let $V_N = \frac{1}{N}\sum \phi_\varepsilon(Z_i)$.

$$\operatorname{Var}\left(\mathbb{E}[\widehat{R}_{K,2N} \mid \mathbf{D}_{2N}]\right) = \operatorname{Var}(V_{2N}) + \operatorname{Var}(\rho_{2N}) + 2\operatorname{Cov}(V_{2N}, \rho_{2N}). \tag{36}$$

Linear Variance: Since $\phi_\varepsilon(Z_i)$ are i.i.d., $\operatorname{Var}(V_{2N}) = \frac{\sigma_\varepsilon^2}{N}$.

Remainder Variance: The term r_N is the second-order error of the von Mises expansion, satisfying $r_N = O(\varepsilon^{-cd_{\text{int}}} N^{-1})$ due to second-order Hadamard differentiability (Goldfeld et al., 2024; Stromme, 2024). By the compactness assumption, moments converge, implying $\operatorname{Var}(r_N) = O(\varepsilon^{-2cd_{\text{int}}} N^{-1} N^{-2})$. Similarly, $\operatorname{Var}(r_{M,1}) = O(\varepsilon^{-2cd_{\text{int}}} N^{-2})$. Thus, $\operatorname{Var}(r_N) = O(\varepsilon^{2cd_{\text{int}}} N^{-2})$.

3. **Covariance:** By Cauchy-Schwarz, $\operatorname{Cov}(V_N, \rho_N) \le \sqrt{\operatorname{Var}(V_N)\operatorname{Var}(\rho_N)} = O(N^{-1.5}\varepsilon^{-cd_{\text{int}}})$.

Additionning both variance terms, we have

$$\operatorname{Var}\left(\mathbb{E}[\widehat{R}_{K,N} \mid \mathbf{D}_{2N}]\right) = \frac{\sigma_\varepsilon^2}{N} + O(N^{-1.5}\varepsilon^{-cd_{\text{int}}}) + O(\varepsilon^{-2cd_{\text{int}}} N^{-2}). \tag{37}$$

Analysis of Term II (Bagging Variance). Conditioned on $\mathbf{D}_{2N}$, the full sample estimator $\widehat{S}_{\varepsilon,N}$ is constant. The variance arises solely from the K subsampled estimators. Since the K bags are drawn independently (conditional on $\mathbf{D}_{2N}$):

$$\operatorname{Var}(\widehat{R}_{K,N} \mid \mathbf{D}_{2N}) = \lambda^2 \operatorname{Var}\left(\frac{1}{K}\sum_{k=1}^{K} \widehat{E}_{M,\mathcal{S}_k} \,\middle|\, \mathbf{D}_{2N}\right) = \frac{\lambda^2}{K}\operatorname{Var}(\widehat{S}^{(1)}_{\varepsilon,N} \mid \mathbf{D}_{2N}). \tag{38}$$

The estimator on the subsample $\widehat{S}^{(1)}_{\varepsilon,N}$ behaves asymptotically as the sample mean of the influence functions drawn without replacement from the finite population $\mathbf{D}_{2N}$. Let $S^2_{\phi_\varepsilon,2N}$ be the exact sample variance of the influence function on $\mathbf{D}_{2N}$. Using the finite population correction formulata, for the variance of a sample mean under simple random sampling with replacement, the conditional variance of the linear term is:

$$\operatorname{Var}\left(\frac{1}{N}\sum_{j\in\mathcal{S}_1} \phi_\varepsilon(Z_j) \,\middle|\, \mathbf{D}_{2N}\right) = S^2_{\phi_\varepsilon,2N}\left(\frac{1}{N} - \frac{1}{2N}\right) + o\left(\frac{S^2_{\phi_\varepsilon,2N}}{N}\right). \tag{39}$$

We now take the expectation over the dataset $\mathbf{D}_{2N}$. Since the samples Z_i are i.i.d., $S^2_{\phi_\varepsilon,2N}$ is an unbiased estimator of the population variance of the influence function. Thus, $\mathbb{E}[S^2_{\phi_\varepsilon,2N}] = \operatorname{Var}(\phi_\varepsilon(Z)) = \sigma_\varepsilon^2$ exactly. Therefore,

$$\mathbb{E}\left[\operatorname{Var}(\widehat{R}_{K,N} \mid \mathbf{D}_{2N})\right] = \frac{\lambda^2}{K}\sigma_\varepsilon^2\left[\frac{1}{N}\right] + o\left(\frac{S^2_{\phi_\varepsilon,2N}}{N}\right) = \frac{\lambda^2}{K}\frac{\sigma_\varepsilon^2}{N} + o\left(\frac{S^2_{\phi_\varepsilon,2N}}{N}\right). \tag{40}$$

Conclusion. Using that, by boundedness of the cost function on bounded support measures, we have $\sigma_\varepsilon \lesssim 1$, independently of ε. finally, substituting (37) and (40) back into (27), as soon as $\varepsilon^{cd_{\text{int}}} = o(N)$, we have:

$$\begin{aligned}\operatorname{Var}(\widehat{R}_{K,N}) &= \left[\frac{\sigma_\varepsilon^2}{N}\right] + \left[\frac{\lambda^2}{K}\frac{\sigma_\varepsilon^2}{N}\right] + O(N^{-1.5}\varepsilon^{-cd_{\text{int}}}) + O(\varepsilon^{-2cd_{\text{int}}}N^{-2}) + o\left(\frac{S^2_{\phi_\varepsilon,2N}}{N}\right) \quad (41)\\ &= \frac{\sigma_\varepsilon^2}{N}\left(1+\frac{\lambda^2}{K}\right) + o(1)\ . \quad (42)\end{aligned}$$

This completes the proof. □