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

Ferdinand Genans*
Sorbonne Université, CNRS, LPSM
genans.ferdinand@gmail.com

Olivier Wintenberger
Sorbonne Université, CNRS, LPSM
olivier.wintenberger@sorbonne-universite.fr

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 the 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 [28], including generative modeling [1, 7], speeding up diffusion models [24], domain adaptation [10], and natural language processing [23].

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 [12] 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 [15], showing that

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

$W_p(\rho_n, \rho)$ typically scales as $\mathcal{O}(n^{-1/d})$ in $\mathbb{R}^d$, 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 [21]. Several strategies mitigate these effects: exploiting low intrinsic dimensional structure [33], or focusing on special settings such as semi-discrete OT [29]. We refer to the recent book [8] for a comprehensive overview.

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 [33], 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 [20], and these properties extend to the entropic setting [31].

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 [11], 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 [13, 25] and recent advances achieving $\mathcal{O}(n^2/\varepsilon)$ [4], 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* [30, 16, 14]. By explicitly defining the discrepancy through the discrete Sinkhorn divergence, one achieves a parametric estimation rate of $\mathcal{O}(\varepsilon^{-d/2}n^{-1/2})$ [17, 27, 9]. 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 [31], 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. Similarly, minimax continuous approximation rates—being governed by the approximation of continuous functions [21]—should also depend on the intrinsic dimension in order to attain the optimal rate. 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 discretization error $\mathrm{OT}_c(\rho, \widehat{\rho}_n^*)$, where $\widehat{\rho}_n^*$ 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 [33]. Crucially, the proposed method scales linearly in n in both time and memory ($\mathcal{O}(nN)$ and $\mathcal{O}(n + N)$ for N Monte Carlo samples), which makes it applicable to large-scale datasets and yields a substantial improvement over the quadratic complexity of the estimator introduced by [5].

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 [9] 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 the first-order entropic bias and, if the statistical bound is tight (Remark 1), the first-order statistical one. This scheme permits us to have a convergence of $o(n^{-2/(d_{\text{int}}+4)})$ for W_2^2, compared to $o(n^{-2/(d_{\text{ambiant}}+8)})$ for the ε-only Richardson scheme with unknown d_{int} from [9].

We validate our pipeline on synthetic manifolds, continuous measures 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 denote by $\|\cdot\|$ the Euclidean norm. 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 $\rho \in \mathcal{P}(\mathbb{R}^d)$, $\text{Supp}(\rho)$ 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

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:

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

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 $\text{OT}_c(\mu, \nu)^{1/p}$ defines the p-Wasserstein distance, denoted $W_p(\mu, \nu)$ [32].

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

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

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:

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

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.

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:

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

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

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

which, for $c = \|\cdot\|^2$, is a divergence that interpolates between OT (as $\varepsilon \to 0$) and Maximum Mean Discrepancy (as $\varepsilon \to \infty$) [16, 14].

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 [33, 20] and EOT [19, 31]. More precisely, the rate tracks the ϵ-covering number $\mathcal{N}_\epsilon$, defining an effective dimension $d_\delta \approx \min_{\rho\in\{\mu,\nu\}} \frac{\log \mathcal{N}_\delta(\rho)}{-\log \delta}$. This quantity exhibits a *multi-resolution behavior*: OT and EOT adapt to the geometry at varying scale of observation, capturing d_δ 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 $d_\delta > \lim_{\delta\to 0} d_\delta$.

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:

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

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 n points.

3.1 Wasserstein discretization error and semi-discrete OT

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

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

Solving (6) typically requires expensive stochastic gradient descent or semi-discrete solvers [16, 22], 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 1. *Let c be a measurable cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. Given $\rho \in \mathcal{P}(\mathbb{R}^d)$ and a support $X = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ such that c is bounded on $\mathrm{Supp}(\rho) \times \{x_1, \ldots, x_n\}$, let $j^*(x) := \min\{\mathrm{argmin}_j c(x, x_j)\}$ denote the index of a nearest support point, ties being broken by the smallest index, and define the weights $\mathbf{w}_n$ by $w_i = \rho\big(\{x \in \mathbb{R}^d; j^*(x) = i\}\big)$, and set $\widehat{\rho}_n^* = \sum_{i=1}^n w_i \delta_{x_i}$. Then $\mathbf{w}_n \in \Delta_n$ and $\widehat{\rho}_n^*$ provides the best approximation of ρ supported on X, that is,*

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

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

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

If ties have zero mass, i.e. $\rho(\{x : \operatorname{argmin}_j c(x, x_j) \text{ is not a singleton}\}) = 0$, $\widehat{\rho}_n^*$ *is unique; this holds, e.g., a.s. for* $c(x, y) = \|x - y\|^p$, *pairwise distinct* x_i*'s and* ρ *absolutely continuous.*

Proof. By definition of $\mathbf{w}_n$, the map $T^*(x) = x_{j^*(x)}$ pushes ρ forward to $\widehat{\rho}_n^*$. Moreover, since $j^*(x) \in \operatorname{argmin}_j c(x, x_j)$, $\int c(x, T^*(x))\, d\rho(x) = \int \min_j c(x, x_j)\, d\rho(x)$. On the other hand, for any $\mathbf{a} \in \Delta_n$ and any coupling $\pi \in \Pi(\rho, \widehat{\rho}_n^{\mathbf{a}})$, $\int c(x, y)\, d\pi(x, y) \geq \int \min_j c(x, x_j)\, d\rho(x)$. Hence $\widehat{\rho}_n^*$ is a minimizer and $\mathrm{OT}_c(\rho, \widehat{\rho}_n^*) = \int \min_j c(x, x_j)\, d\rho(x) = \int \mathbf{0}_n^c(x)\, d\rho(x)$, so $\mathbf{0}_n$ is a maximizer of the semi-dual. If ties have zero ρ-mass, any coupling attaining this lower bound must send ρ-almost every x to its unique nearest support point. Its second marginal is therefore necessarily $\widehat{\rho}_n^*$, which proves uniqueness. □

3.2 A Solver-Free Monte Carlo Estimator

The key insight of Proposition 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:

$$w_i \;\approx\; \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}\{j^*(X_k) = i\} \;=:\; \widehat{w}_{i,N},$$

with the same tie-breaking rule j^* as in Proposition 1. The following theorem establishes that these estimators concentrate exponentially fast around their true values.

Proposition 2. *Let* c *be measurable with* $C_\rho := \sup_{x,y \in \mathit{supp}(\rho)} c(x, y) < \infty$, *let* $x_1, \ldots, x_n \in \mathrm{Supp}(\rho)$ *and* $N \geq 2$. *Then for any* $\delta \in (0, 1)$, *we have with probability at least* $1 - \delta$:

$$\left| \mathrm{OT}_c(\rho, \widehat{\rho}_n^*) - \widehat{\mathrm{OT}}_N^n \right| \leq \sqrt{\tfrac{2\widehat{\sigma}^2_{\mathbf{0}^c} \log(4/\delta)}{N}} + \tfrac{7 C_\rho \log(4/\delta)}{3(N-1)}\,,$$

and $\widehat{\sigma}^2_{\mathbf{0}^c} = \dfrac{1}{N-1} \sum_{k=1}^{N} \left(\min_j c(X_k, x_j) - \widehat{\mathrm{OT}}_N^n \right)^2$ *is the usual unbiased sample variance estimator. Also, for* $i \in [\![1, n]\!]$, *with probability* $1 - \delta$: $|w_i - \widehat{w}_{i,N}| \leq \sqrt{\tfrac{2\widehat{w}_{i,N}(1-\widehat{w}_{i,N}) \log(4/\delta)}{N-1}} + \tfrac{7 \log(4/\delta)}{3(N-1)}$.

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 of [26] yields the first inequality, applied to both tails with confidence $\delta/2$ each (it is one-sided). The same theorem applies to the second inequality, since $w_i \in [0, 1]$ for all i with unbiased sample variance $\frac{N}{N-1} \widehat{w}_{i,N}(1 - \widehat{w}_{i,N})$. □

We conclude with two practical observations. First, although the bound depends on the diameter C_ρ, this term appears linearly and in the non limiting term. Second, while standard practice often assumes uniform weights ($\widehat{\rho}_n^{\text{unif}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$), our estimator $\widehat{\mathrm{OT}}_N^n$ targets the optimally weighted measure $\widehat{\rho}_n^*$. Since $\mathrm{OT}_c(\rho, \widehat{\rho}_n^*) \leq \mathrm{OT}_c(\rho, \widehat{\rho}_n^{\text{unif}})$, 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.

Optimal vs. uniform weights; cost. Equation (7) estimates the optimally reweighted error, a lower bound on the uniform-weight error of empirical OT. Evaluating $\widehat{\mathrm{OT}}_N^n$ is cheap and costs $\mathcal{O}(nN)$ time and $\mathcal{O}(n + N)$ memory (streamed), with $N = \mathcal{O}(\delta^{-2})$ for accuracy δ.

Entropic counterpart. The construction extends to entropic regularization (Proposition 8, Appendix G): with the entropy measured relative to $\rho \otimes \frac{1}{n}\sum_i \delta_{x_i}$ (as in (4) for uniform weights), the optimal weights become $w_i^\varepsilon = \int e^{-c(x,x_i)/\varepsilon} \big/ \sum_j e^{-c(x,x_j)/\varepsilon}\, d\rho(x)$, the zero potential is still optimal, and this entropic semi-discrete OT has again a solver-free Monte Carlo estimator.

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 has low-dimensional structure. The manifold hypothesis assumes data lies on a smooth submanifold $\mathcal{M} \subset \mathbb{R}^d$, but real data is typically more multi-scale: its effective complexity depends on the observation scale or target accuracy. For example, let $\rho := \frac{9}{10}\mathcal{U}([0,1]^2 \times \{0\}^8) + \frac{1}{10}\mathcal{U}([1,2]^8 \times \{0\}^2) \in \mathcal{P}(\mathbb{R}^{10})$. Its support has topological dimension 8, yet its optimal-transport discretization behavior depends on the required precision. At coarse accuracy, say 50% error, the measure may behave like a 2-dimensional problem. At finer accuracy, say below 10%, the discretization must resolve the 8-dimensional component, incurring the curse of dimensionality for $d = 8$.

Thus, for probability measures, complexity is better viewed as scale-dependent rather than as a fixed integer. It is governed by covering numbers at different radii [33]. We therefore replace rigid geometric assumptions, such as the manifold hypothesis, with a functional assumption based directly on the convergence rate of the Wasserstein distance.

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}, in the following sense.

Assumption 4.1 (Effective Intrinsic Dimension)**.** Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ have 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 on ρ (and on the window $[n_{\min}, n_{\max}]$) such that:

(Lower bound) For any discrete measure σ_n supported on at most n points, the approximation error is lower-bounded by the dimension: $W_1(\rho, \sigma_n) \geq C_1 n^{-1/s}$.

(Upper bound) The empirical measure $\widehat{\rho}_n$ constructed from n i.i.d. samples satisfies: $\mathbb{E}[W_1(\rho, \widehat{\rho}_n)] \leq C_2 n^{-1/s'}$, 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 this specific setting.

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 $\widehat{\mathrm{OT}}_N^n$ (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:

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

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 for a fixed N. The statistical guarantee of our method is given in the following theorem whose proof is deferred in Appendix D.

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, with $c(x,y) = \|x-y\|$, for $\gamma, \xi > 0$, $\delta \in (0,1)$ and $n_{\min} \leq n,\ n\eta \leq n_{\max}$, using*

$$\eta \;\geq\; \Big(\frac{(1+\xi)C_2}{C_1}\Big)^{\frac{2(1+\gamma)d_{\mathrm{int}}}{\gamma}}, \quad n \;\geq\; \left(\frac{\mathrm{Diam}(\rho)^2}{2\xi^2 C_2^2}\log(4/\delta)\right)^{\frac{d_{\mathrm{int}}}{d_{\mathrm{int}}-2}},$$
$$N \;\geq\; \left[\frac{16(1+\gamma)^2\, d_{\mathrm{int}}^2}{\gamma^2(\log\eta)^2} \;\vee\; 1\right] \frac{2\,\mathrm{Diam}(\rho)^2}{C_1^2}\eta^{2/d_{\mathrm{int}}} n^{2/d_{\mathrm{int}}} \log\Big(\frac{8}{\delta}\Big),$$

we have with probability $1-\delta$: $d_{\mathrm{int}}\frac{1+\gamma}{1+2\gamma} \leq \widehat{d}^* \leq (1+\gamma)d_{\mathrm{int}}$.

For $d_{\mathrm{int}} \leq 2$, Theorem 4.2 makes no claim (the uniform-weight rate is then parametric up to logarithmic factors). Appendix D.2 gives a cheap multiscale diagnostic indicating when $\widehat{d}^*$ should not be used for calibration.

Comparison to the literature. The concept of leveraging the multi-scale decay of Wasserstein error for dimension estimation was pioneered in [5]. 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 in n, $\mathcal{O}(nN)$, 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 [5] and standard baselines from the `scikit-dimension` package [2]: MLE, CorrInt, lPCA, TwoNN.

Experimental Setup: We generate data in $\mathbb{R}^{20}$ under three configurations designed to have an effective Wasserstein dimension of $d_{\mathrm{int}} = 10$: (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.

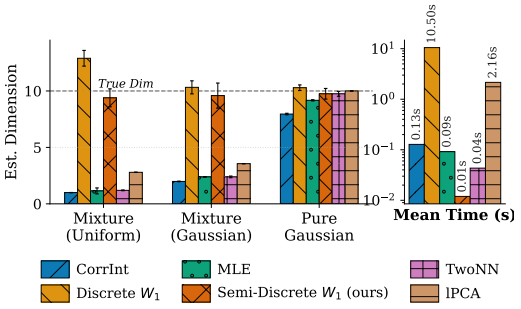

Figure 1: Intrinsic Dimension Estimation Benchmark. Comparison of our Semi-Discrete W_1 estimator against the Discrete W_1 baseline and standard geometric estimators. The dashed line represents the ground truth effective dimension $d_{\mathrm{int}} = 10$. Our estimator matches the robustness of discrete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

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 $d_{\mathrm{int}} \approx 10$. 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

The Sinkhorn divergence serves as a fast proxy for OT, supported by efficient solvers like `geomloss` [14] 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, [9] introduced a Richardson Extrapolation scheme on ε, proposing the estimator $\widehat{R}_{\varepsilon,n} = 2\widehat{S}_{\varepsilon,n} - \widehat{S}_{\sqrt{2}\varepsilon,n}$, where $\widehat{S}_{\varepsilon,n} := S_\varepsilon(\widehat{\mu}_n, \widehat{\nu}_n)$ denotes the Sinkhorn divergence between the two n-point empirical measures This approach theoretically improves the approximation error to $\mathbb{E}[|\widehat{R}_{\varepsilon,n} - W_2^2|] = \mathcal{O}(n^{-4/(d'+8)})$ (with $\varepsilon \asymp n^{-1/(d'+8)}$, $d' = 2\lfloor d/2 \rfloor$, under a second-order entropic expansion; $o(n^{-2/(d+8)})$ under the first-order one (8)).

A first improvement is to calibrate ε using our intrinsic dimension estimator $\widehat{d}^*$ from Sections 3–4. Since the relevant exponent is the *minimum* intrinsic dimension of μ and ν [31], we estimate it on both marginals and retain the smaller value (see Appendix C.3 for its connection to Assumption 4.1 and the distinction between W_1 and W_2). However, even when 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:

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

We characterize the asymptotic expansions of both terms. The entropic bias expansion (8) follows from Theorem 1 in [9] (assuming bounded Lebesgue densities with finite Fisher information; for manifold-supported measures, (8) is an assumption), while the statistical bias bound (9) is derived from Theorem 2 in [31], which establishes the lower complexity adaptation of the Sinkhorn divergence:

$$\text{Bias}_{\text{ent}}(n) = C_3\varepsilon_n^2 + o_{\text{ent}}(\varepsilon_n^2) , \tag{8}$$

$$\left|\text{Bias}_{\text{stat}}(n)\right| \le C_4\varepsilon_n^{-d_{\text{int}}/2} n^{-1/2} . \tag{9}$$

Remark 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 [31, 19]. The tightness of the rate of the statistical bias, even though discussed in these papers, was not perfectly established and remains an open question. Our cancellation argument uses slightly more than* (9)*, namely*

$$\textit{Bias}_{\textit{stat}}(n) = C_{\textit{stat}}\,\varepsilon_n^{-d_{\text{int}}/2} n^{-1/2} + o_{\textit{stat}}(\varepsilon_n^{-d_{\text{int}}/2} n^{-1/2}) \quad \textit{for some } C_{\textit{stat}} \in \mathbb{R}. \tag{H}$$

It covers the tight case $C_{\textit{stat}} \neq 0$, where Richardson cancels both first-order biases, and the non-tight case $C_{\textit{stat}} = 0$. (H) *is not proved in general; without it,* (8)–(9) *only yield $\mathcal{O}(n^{-2/(d_{\text{int}}+4)})$ below.*

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: as usual one uses two sample sizes n and $2n$. On the contrary with [9], which use a constant ε for both sample sizes, we choose ε_n and ε_{2n} accordingly:

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

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

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

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

Proposition 3. *Let μ, ν be compactly supported, $c(x,y) = \|x-y\|^2$, and assume* (8) *and* (H)*. Using the schedule $\varepsilon_n = \varepsilon_0\, n^{-1/(d_{\mathrm{int}}+4)}$, $\varepsilon_0 > 0$, the Diagonal Richardson estimator satisfies:*

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

Furthermore, if $o_{\mathit{ent}}(\varepsilon_n^2) = \mathcal{O}(\varepsilon_n^4)$ and $o_{\mathit{stat}}(\varepsilon_n^{d_{\mathrm{int}}/2} n^{-1/2}) = o_{\mathit{stat}}(n^{-(1-ad_{\mathrm{int}})}/2) = \mathcal{O}(n^{-(1-ad_{\mathrm{int}})})$, the rate improves to $\mathcal{O}(n^{-\frac{4}{d_{\mathrm{int}}+4}} + n^{-\frac{1}{2}})$, i.e. $\mathcal{O}(n^{-\frac{4}{d_{\mathrm{int}}+4}})$ if $d_{\mathrm{int}} \geq 4$.

Therefore, this rate is slightly better than the ε-only Richardson, which is $o(n^{-\frac{2}{d_{\mathrm{int}}+8}})$ using d_{int}, or at best $\mathcal{O}(n^{-\frac{2}{d_{\mathrm{int}}+4}})$ (no statistical bias is cancelled); at second order, $\mathcal{O}(n^{-\frac{4}{d_{\mathrm{int}}+4}} + n^{-\frac{1}{2}})$ improves on its $\mathcal{O}(n^{-\frac{4}{d_{\mathrm{int}}+8}})$.

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

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

The stochastic error scales as $\mathcal{O}(n^{-1/2})$, which is negligible compared to the bias terms for every $d_{\mathrm{int}} > 0$, since $\frac{2}{d_{\mathrm{int}}+4} < \frac{1}{2}$. The choice of weights cancels the leading order terms $n^{-\frac{2}{d_{\mathrm{int}}+4}}$ in the bias, leaving only the higher-order remainder. Both regimes of Remark 1 give this rate.

If $\widehat{d}^*$ replaces d_{int}, the entropic leading term is still cancelled exactly and the residual statistical first-order term is reduced by a factor $\mathcal{O}(|\widehat{d}^* - d_{\mathrm{int}}|)$ (Appendix E).

Bagged Diagonal Richardson.

While the Diagonal Richardson estimator $\widehat{R}_{2n}^{\mathrm{Diag}}$ 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:

$$\widehat{R}_{K,2n}^{\mathrm{Bagged}} := w_{2n}\widehat{S}_{\varepsilon_{2n},2n} + \frac{w_n}{K}\sum_{k=1}^{K} \widehat{S}_{\varepsilon_n,n}^{(k)}\ , \tag{10}$$

where $\widehat{S}_{\varepsilon_n,n}^{(k)}$ is computed on the k-th subsample of n pairs drawn without replacement from the $2n$ available ones, independently across bags.

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 F, builds on the second-order Hadamard differentiability of the Sinkhorn divergence established in [18].

The following proposition quantifies the variance reduction achieved by bagging, showing that the additional variance due to subsampling decreases as $1/K$. For a common regularization, this allows the estimator to asymptotically recover the variance of the full-sample Sinkhorn estimator as K increases.

Proposition 4 (Variance Reduction with Bagging)**.** *Let μ, ν be compactly supported, $c(x,y) = \|x-y\|^2$. The variance of the Bagged Richardson estimator relates to the variance of the standard Sinkhorn estimator $\widehat{S}_{\varepsilon_{2n},2n}$ as:*

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

With the common regularization ε_{2n} in the subsample terms, the remainder is $o(1/n)$.

This crude analysis is valid for any fixed regularization parameters $\varepsilon_1, \varepsilon_2 > 0$; sharpening the main reminder term in $|\varepsilon_n - \varepsilon_{2n}|^2 \log^2 n$ for decreasing regularization would require uniform control

of the von Mises remainder of the Sinkhorn divergence [18] as $\varepsilon \to 0$, which is not available to our knowledge. Nevertheless, the same stabilizing effect is observed numerically for our diagonal Richardson estimator (Appendix B).

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 $\psi_\theta(x) = \frac{1}{2}\|x\|^2 + \text{ICNN}(x)$, where μ denotes the underlying data distribution. As ψ_θ is convex, $(\text{Id}, \nabla\psi_\theta)_{\#}\mu$ is cyclically monotone, hence optimal for any μ with finite second moment [32, Thm. 5.10], providing an exact reference distance $W_2^2(\mu,\nu) = \mathbb{E}_\mu[\|X - T^*(X)\|^2]$, with $T^* = \nabla\psi_\theta$, evaluated by Monte Carlo over the dataset.

Experimental Setup. For each dataset, with $n = 1000$, we first estimate the intrinsic dimension $\widehat{d}^*$ (Sec. 4; $\widehat{d}^* \approx 13/15/24$ on MNIST/FashionMNIST/CIFAR10) 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") [9].

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.

Table 1: Benchmark of 2-Wasserstein estimation

Dataset	Method	Error ± std	Avg Time(s)
CIFAR	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

Additional experiments. Appendix B studies the robustness to k in $\varepsilon_n \propto n^{-1/(d+k)}$, to n, η, the ambient dimension and scale (up to 10^7 points in a minute) for $\widehat{d}^*$, and the role of K (variance only). We also give a practical protocol, in Appendix D.2, to detect when the sizes lie outside a stable-exponent regime, validated on a multiscale mixture.

6 Conclusion and Future Work

In this work, we introduced a solver-free estimator of the intrinsic dimension of probability measures, avoiding the two costly discrete OT solves required in the approach of [5]. Building on this estimate, we showed that explicitly accounting for intrinsic dimension is essential for overcoming the statistical limitations of discrete Optimal Transport. In particular, we demonstrated that our Diagonal Richardson estimator, calibrated using this solver-free intrinsic-dimension estimate, substantially improves the accuracy of Wasserstein distance estimation on high-dimensional manifolds.

Beyond these results, our framework points toward several directions for future research. First, dimension-aware solvers could be designed to exploit intrinsic geometry and approximate OT costs more accurately at lower computational cost. Such solver-level improvements would also provide a natural bridge to a second direction: extending geometry-aware debiasing, together with its theoretical analysis, beyond scalar transport costs to Kantorovich potentials and gradients, with the goal of improving training stability in generative modeling.

Acknowledgements

The authors are grateful to Austin Stromme for valuable discussions and to the anonymous reviewers for their helpful comments and suggestions.

References

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

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

[3] B. Bercu and J. Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *The Annals of Statistics*, 49 (2):968–987, 2021.

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

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

[6] J.-D. Boissonnat, A. Lieutier, and M. Wintraecken. The reach, metric distortion, geodesic convexity and the variation of tangent spaces: J.-d. boissonnat et al. *Journal of applied and computational topology*, 3(1):29–58, 2019.

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

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

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

[10] N. Courty, R. Flamary, and D. Tuia. 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*, pages 274–289. Springer, 2014.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contents

1 Introduction **1**

2 Preliminaries **3**

3 A Simple Estimator of the Wasserstein Discretization Error **4**

3.1 Wasserstein discretization error and semi-discrete OT 4

3.2 A Solver-Free Monte Carlo Estimator 5

4 Discretization Error and Intrinsic Dimension Estimation **6**

4.1 Geometric setting: Multi-scale behavior and Covering Numbers 6

4.2 Leveraging our discretization estimator to estimate the intrinsic dimension. 6

4.3 Experiments 7

5 Coupled Bias Reduction: The Diagonal Richardson Estimator **7**

5.1 From OT Geometry to Wasserstein Distance Estimation 10

6 Conclusion and Future Work **10**

A Experiments Details **14**

B Additional Experiments **14**

B.1 Sensitivity of the intrinsic-dimension estimator 14

B.2 Large-scale applicability 15

B.3 Sensitivity of the Diagonal Richardson estimator 16

B.4 Wasserstein estimation on continuous measures with exact ground truth 17

B.5 Estimation on finite datasets 18

C Intrinsic Dimension and Wasserstein Distances **18**

C.1 Notions. 18

C.2 Properties 18

C.3 Minimum Intrinsic Dimension scaling 19

D Proof of Theorem 4.2: Dimension Estimation Control **20**

D.1 Intrinsic Dimension estimation under the manifold hypothesis 22

D.2 A multiscale diagnostic for Assumption 4.1 23

E Proof of Proposition 3: Convergence of Diagonal Richardson **24**

F Proof of Proposition 4: Variance Reduction with Bagging **26**

G Complements to Section 3: Entropic Counterpart **28**

A Experiments Details

Environment: All experiments were executed on Google Colab using the freely available GPU runtime, except the large-scale experiment of Table 6, run on a single A100 GPU.

In all synthetic experiments, samples are represented in an ambient space $\mathbb{R}^{d_{\mathrm{amb}}}$ (with $d_{\mathrm{amb}} = 20$ unless otherwise stated). To construct distributions with intrinsic dimension $d_{\mathrm{int}} < d_{\mathrm{amb}}$, we generate samples in $\mathbb{R}^{d_{\mathrm{int}}}$ and embed them into the ambient space via a canonical (or random orthogonal) inclusion. We use three building blocks.

Uniform hypercube. For a component with intrinsic dimension d_{int}, we sample uniformly from

$$\mathrm{Unif}\big([0,1]^{d_{\mathrm{int}}} \times \{0\}^{d_{\mathrm{amb}}-d_{\mathrm{int}}}\big) \subset \mathbb{R}^{d_{\mathrm{amb}}}.$$

The samples sit in $\mathbb{R}^{d_{\mathrm{amb}}}$, but their support is exactly a d_{int}-dimensional axis-aligned cube.

Low-rank Gaussian. For Gaussian components, we sample

$$x \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = U\,\Lambda\,U^{\top},$$

where $U \in \mathbb{R}^{d_{\mathrm{amb}} \times d_{\mathrm{int}}}$ has orthonormal columns and $\Lambda \in \mathbb{R}^{d_{\mathrm{int}} \times d_{\mathrm{int}}}$ is positive definite. The covariance Σ has rank d_{int}, so the Gaussian is supported on the d_{int}-dimensional subspace spanned by U and is degenerate in the orthogonal directions.

Mixtures. For mixtures, we draw a component label from a Bernoulli (or categorical) distribution with the prescribed weights, then sample from the corresponding component. In particular, our running mixture in $\mathbb{R}^{20}$ combines a high-mass low-dimensional component (e.g., 80% on a 2-dimensional support) with a low-mass high-dimensional one (e.g., 20% on a 10-dimensional support); this produces a measure whose effective Wasserstein dimension at moderate sample sizes is governed by the high-dimensional component, while standard local geometric estimators are biased by the low-dimensional bulk.

Wasserstein benchmark on images (Section 5). The target $\nu = (\nabla\psi_\theta)_{\#}\mu$ is built from a convex potential $\psi_\theta = \frac{1}{2}\|\cdot\|^2 + \mathrm{ICNN}_\theta$, and the reference $W_2^2(\mu,\nu) = \mathbb{E}_\mu\|X - \nabla\psi_\theta(X)\|^2$ is exact for the population measure μ whatever the parameters θ, since $\nabla\psi_\theta$ is an optimal map for any convex ψ_θ (Section 5). The quality of the ICNN fit therefore does not affect the reliability of the reference: θ only shapes the target measure. The only error of the reference is the Monte Carlo evaluation of this expectation by the average over the dataset; this ambiguity is removed by the continuous benchmark of Appendix B.4, where fresh samples are drawn at will. For $n = 1000$, the estimated intrinsic dimensions are on average $\widehat{d}^* \approx 13$ (MNIST), 15 (Fashion-MNIST) and 24 (CIFAR-10); they set both the schedule $\varepsilon_n \propto n^{-1/(\widehat{d}^*+4)}$ and the Richardson weights. The number of bags $K = 12$ was fixed without tuning from Figure 3, and its influence is studied in Appendix B.3.

B Additional Experiments

B.1 Sensitivity of the intrinsic-dimension estimator

We study how $\widehat{d}^*$ behaves as we vary (i) the sample size n and the scale ratio η, and (ii) the ambient dimension d_{amb}. The data follow the three configurations of Section 4: a hypercube mixture (Config 1), a low-rank Gaussian mixture (Config 2), and a single low-rank Gaussian (Config 3), each with $d_{\mathrm{int}} = 10$ and $d_{\mathrm{amb}} = 20$ unless otherwise stated.

Sample size and scale ratio. Tables 2 and 3 report $\widehat{d}^*$ on the mixture of Config 2 as we sweep, respectively, n at $\eta = 2$ and η at $n = 2000$. Two patterns emerge. First, the estimate is centred close to the ground truth $d_{\mathrm{int}} = 10$ across the entire grid, with no detectable bias. Second, the standard deviation decreases monotonically in both n and η: this is the empirical counterpart of Theorem 4.2, where the precision of $\widehat{d}^*$ scales as $1/\log\eta$ in η and as a power of n governed by d_{int}. In our remaining experiments we use $\eta = 2$ as a default; this value is large enough to keep the slope estimator stable, while remaining within the regime where Assumption 4.1 is plausible.

Table 2: Intrinsic-dimension estimation as a function of the sample size n, with $\eta = 2$. Mean and standard deviation over 20 runs.

n	Mean $\widehat{d}^*$	Std
100	9.34	$\pm$ 2.31
300	9.31	$\pm$ 0.99
700	9.66	$\pm$ 0.76
1200	9.62	$\pm$ 0.51
2000	9.71	$\pm$ 0.40
3500	9.81	$\pm$ 0.35

Table 3: Intrinsic-dimension estimation as a function of the scale ratio η, with $n = 2000$. Mean and standard deviation over 20 runs.

η	Mean $\widehat{d}^*$	Std
1.2	9.75	$\pm$ 1.13
1.4	9.73	$\pm$ 0.48
1.6	9.76	$\pm$ 0.40
1.9	9.69	$\pm$ 0.24
2.2	9.72	$\pm$ 0.21
2.5	9.70	$\pm$ 0.19

Ambient dimension. We next isolate the effect of the ambient dimension d_{amb}. In Table 4 we keep $d_{\mathrm{int}} = 10$ fixed and grow d_{amb} from 15 to 100. Across this entire range the estimate stays within ± 0.3 of the ground truth on all three configurations, with no systematic drift. This is the practical instantiation of the multi-scale theory of Section 4: $\widehat{d}^*$ depends on the geometry of the support, not on the ambient embedding, exactly as predicted by the minimum-intrinsic-dimension scaling.

Table 4: Intrinsic-dimension estimation under increasing ambient dimension d_{amb}, with the intrinsic dimension fixed at $d_{\mathrm{int}} = 10$. Mean and standard deviation over 20 runs at $n = 2000$, $\eta = 2$.

d_{amb}	Config 1	Config 2	Config 3
15	9.63 ± 0.16	9.89 ± 0.18	9.94 ± 0.09
25	9.56 ± 0.21	9.98 ± 0.22	9.98 ± 0.12
50	9.69 ± 0.14	10.02 ± 0.28	9.89 ± 0.12
100	9.61 ± 0.13	9.82 ± 0.21	10.13 ± 0.11

Table 5 stresses the estimator in the opposite regime: the support nearly fills the ambient space, with $d_{\mathrm{int}} = d_{\mathrm{amb}} - 2$ for $(d_{\mathrm{amb}}, d_{\mathrm{int}}) \in \{(5,3), (10,8), (20,18), (40,38)\}$. This experiment guards against the trivial failure mode in which a method underestimates the ambient dimension: a procedure that does so would still appear accurate in Table 4. Up to $(20, 18)$, $\widehat{d}^*$ tracks the true intrinsic dimension closely. The $(40, 38)$ regime, however, shows a noticeable underestimation, especially for the mixture configurations. This is expected: the discretization rate $n^{-1/d_{\mathrm{int}}}$ flattens for large d_{int}, so the log-ratio that defines $\widehat{d}^*$ shrinks while its noise stays roughly constant, and recovering the inverse becomes intrinsically hard. In this regime substantially larger sample sizes are required.

Table 5: Intrinsic-dimension estimation when the intrinsic dimension is close to the ambient dimension, $d_{\mathrm{int}} = d_{\mathrm{amb}} - 2$. Mean and standard deviation over 20 runs at $n = 2000$, $\eta = 2$.

$(d_{\mathrm{amb}}, d_{\mathrm{int}})$	Config 1	Config 2	Config 3
$(5, 3)$	2.91 ± 0.04	3.04 ± 0.03	3.07 ± 0.06
$(10, 8)$	7.46 ± 0.19	7.93 ± 0.19	7.99 ± 0.05
$(20, 18)$	17.10 ± 0.60	17.47 ± 0.39	18.05 ± 0.31
$(40, 38)$	26.02 ± 0.92	33.20 ± 1.84	34.01 ± 0.55

B.2 Large-scale applicability

The support sizes of Section 4 ($n = 2000$, $\eta n = 3000$) are dictated by the Discrete W_1 baseline, which requires exact OT solves and becomes impractical beyond a few thousand points; all methods were run at the same sizes for a fair comparison. Our estimator has no such limitation as reported in Table 6: 10^7 support points are processed in under a minute, a regime out of reach of discrete-OT-based estimators.

Table 6: Large-scale intrinsic-dimension estimation on a rank-10 Gaussian ($\eta = 10$, $N = 10^6$, mean $\pm$ standard deviation over 10 repetitions, single A100 GPU).

n	10^3	10^4	10^5	10^6	10^7
$\widehat{d}^*$ ($n \to 10n$)	9.82 ± 0.05	10.03 ± 0.01	10.04 ± 0.004	10.01 ± 0.002	10.03 ± 0.002
Runtime (s)	0.05	0.14	0.62	5.10	49.8

B.3 Sensitivity of the Diagonal Richardson estimator

We now turn to the Diagonal Richardson estimator. The three studies below probe its calibration in d_{int}, its calibration in the regularisation exponent, and the variance trade-off it pays.

Calibration to the intrinsic dimension. Figure 2 sweeps the dimension parameter d used in the Diagonal Richardson weights $w_{2n} = 2^\gamma/(2^\gamma - 1)$, $w_n = -1/(2^\gamma - 1)$ with $\gamma = 2/(d+4)$. The data are a mixture in $\mathbb{R}^{10}$ (90% on a 5-dimensional Gaussian, 10% on a 1-dimensional Gaussian), the target is a 10-dimensional Gaussian, and we fix $2n = 2000$. The estimation error $|\widehat{R}^{\text{Diag}}_{2n} - W_2^2|$ is sharply minimised at $d = 5$, the dominant intrinsic dimension of the source mixture. Two observations follow: (i) calibrating the Diagonal Richardson weights to the *intrinsic* (rather than ambient) dimension is essential, and (ii) the relevant quantity is the effective dimension of the source, the lower-dimensional of the two measures ($5 < 10$), which is that of its dominant 5-dimensional component and not of its 1-dimensional one, in line with the minimum-intrinsic-dimension scaling recalled in Section 2. This empirically validates the use of $\widehat{d}^*$ as the calibration parameter in our pipeline.

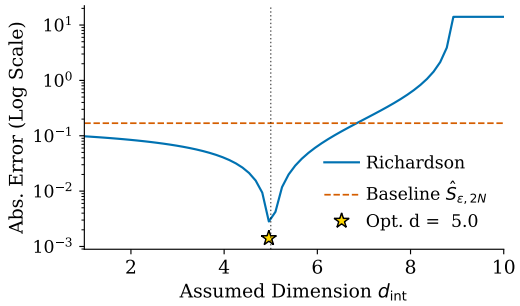

Figure 2: Sensitivity of the Diagonal Richardson estimator to the dimension parameter d used in its weights, on a $\mathbb{R}^{10}$ mixture with dominant component of intrinsic dimension 5 and a 10-dimensional Gaussian target. The error is minimised exactly at $d = 5$.

Regularisation exponent. Table 7 reports the estimation error of Diagonal Richardson with regularisation schedule $\varepsilon_n = n^{-a}$ and $a = 1/(d_{\text{int}} + k)$ for several values of k. The theoretical optimum derived in Section 5 corresponds to $k = 4$. Across all sample sizes $n \in \{200, 300, 500, 650\}$, every choice of k improves the standard estimator by at least an order of magnitude. The empirical optimum varies slightly with n—e.g. $k = 4$ wins at $n = 200$ and $n = 650$, while $k = 8$ wins at $n = 300$ and $k = 2$ at $n = 500$—but the spread between values is small in absolute terms. This mirrors the bias-reduction-factor analysis of Appendix E: small mismatches in the rate exponent produce only small residual leading-order biases, with magnitude proportional to the mismatch. In practice, the theory-recommended $k = 4$ is a safe default and does not require finetuning.

Variance reduction via bagging. Figure 3 measures the variance of the Bagged Diagonal Richardson estimator on the same source/target pair as Figure 2, as a function of the bag count K. The variance is estimated over 20 runs and averaged across 10 repetitions. As predicted by Proposition 4 for a common regularization, and although only an additive control is proved for (10) itself, the curve decays as $1/K$ and asymptotes to the variance floor of the standard Sinkhorn estimator $\widehat{S}_{\varepsilon_{2n},2n}$ (dashed line). The variance overhead from extrapolation therefore vanishes in the limit $K \to \infty$, and even moderate values of K recover most of the gain: at $K = 12$, which we use throughout the main

Table 7: Robustness of the Diagonal Richardson estimator to the regularisation exponent. We report the estimation error for $\varepsilon_n = n^{-a}$ with $a = 1/(d_{\text{int}} + k)$. The best result for each sample size is in bold. Standard refers to the unextrapolated Sinkhorn estimator with the same ε_n.

n	Standard	$k = 0$	$k = 2$	$k = 4$	$k = 6$	$k = 8$
200	0.4839	0.0383	0.0229	**0.0159**	0.0217	0.0231
300	0.4207	0.0063	0.0165	0.0163	0.0090	**0.0040**
500	0.3272	0.0087	**0.0018**	0.0237	0.0085	**0.0018**
650	0.2964	0.0025	0.0078	**0.0010**	0.0040	0.0087

experiments, the residual overhead is already within a few percent of the floor. The bagged estimator thus inherits the bias reduction of Diagonal Richardson at essentially no asymptotic variance cost.

Choice of K. The value $K = 12$ used in Section 5 was fixed without tuning from Figure 3. A dedicated sweep in the setting of Table 1, with $K \in \{1, 2, 4, 8, 12, 24\}$ and 20 repetitions per value, confirms the picture: the empirical variance decreases by a factor 8 to 33 from $K = 1$ to $K = 24$, while the mean error is approximately constant for $K \geq 8$. This is consistent with K entering only the variance and not the bias: K is a computation–variance trade-off, and when the OT losses are already averaged over mini-batches, the variance overhead may be negligible without any bagging.

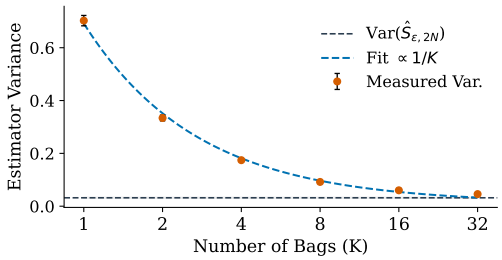

Figure 3: Variance of the Bagged Diagonal Richardson estimator as a function of the number of bags K, on the mixture of Figure 2. The variance decays as $1/K$ towards the standard-Sinkhorn variance floor (dashed line), matching Proposition 4.

B.4 Wasserstein estimation on continuous measures with exact ground truth

To remove any ambiguity due to finite datasets, we repeat the protocol of Table 1 on continuous measures, from which fresh samples are drawn at will: a mixture of low-rank Gaussians, a single rank-10 Gaussian, and four rank-10 Gaussian clusters. The target is $\nu = (\nabla\psi)_{\#}\mu$ for a convex potential ψ as in Section 5, so that the reference $W_2^2(\mu, \nu) = \mathbb{E}_\mu\|X - \nabla\psi(X)\|^2$ is exact; it is evaluated on 10^5 independent samples. With $n = 10^5$, each estimator is compared with the base Sinkhorn estimator on the same samples, through the paired ratio of their errors (a ratio r means an error r times smaller); Table 8 reports its mean, minimum and maximum over repetitions.

Table 8: Error reduction of the Diagonal Richardson estimator over the base Sinkhorn estimator on continuous measures ($n = 10^5$, exact Brenier ground truth): paired ratio of errors, mean [min, max] over repetitions.

	Mixture of low-rank Gaussians	Rank-10 Gaussian	Rank-10 clusters
Diag-Rich / Base	11.7× [2.1, 62.5]	9.5× [3.9, 35.0]	13.4× [2.5, 67.8]

The conclusions of Table 1 are unchanged in this fully controlled setting. The minimum ratios are well above 1: the Diagonal Richardson estimator improves on the base estimator in every single repetition, so the gains are systematic and not within the noise.

B.5 Estimation on finite datasets

For real datasets the population ρ is replaced by a finite sample of M points, which we partition into a support set of size n (used to define $\widehat{\rho}_n^*$) and a Monte Carlo set of size $N_{\mathrm{MC}} = M - n$ (used to estimate $\widehat{\mathrm{OT}}_N^n$). Both sets are drawn without replacement from the dataset and are disjoint by construction. The estimator and its analysis are unchanged: the only operational requirement is that N_{MC} be large enough to make the Monte Carlo concentration of Proposition 2 tight, since the residual MC error decays as $1/\sqrt{N_{\mathrm{MC}}}$. This is the procedure we use for MNIST, FashionMNIST, and CIFAR-10 in Section 5, with $n = 1000$ and N_{MC} equal to the rest of the training set. In settings where the dataset is too small to afford this split (a few thousand samples or fewer), one can fall back to bootstrap-style averaging over multiple support/MC partitions, at a small additional computational cost.

C 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 [33, 31].

C.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 $S \subseteq \bigcup_{1 \le i \le m} B_i$. 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 [12].

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

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

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

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

C.2 Properties

Proposition 5 (Prop. 5 and Corollary 1 of [33])**.** *Let $p \in [1,\infty)$. Suppose there exist $\delta' \le 1$ and $s > 2p$ such that $d_\delta(\mu, \delta^{\frac{sp}{s-2p}}) \le s$ for all $\delta \le \delta'$. Then there exist constants C_1, C_2 such that:*

$$\mathbb{E}[W_p^p(\mu,\widehat{\mu}_n)] \le C_1 n^{-p/s} + C_2 n^{-1/2}. \tag{11}$$

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

We now derive a slightly different proposition, based on Proposition 6 of [33], with a similar proof.

Proposition 6 (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*

$$\mathcal{N}'_\delta(\mu, \tau) \geq A\delta^{-s}, \quad \forall \delta \in [a, b]. \tag{12}$$

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:

$$W_p^p(\mu, \nu) \geq \left(\tau 4^{-p} A^{p/s}\right) n^{-p/s}. \tag{13}$$

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

$$N \geq A\varepsilon^{-s} > A\varepsilon_n^{-s} = n. \tag{14}$$

Let $\{B_1, \ldots, B_N\}$ 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 and Bach [33] (Proposition 6), the aggregated transport cost is lower bounded by:

$$W_p^p(\mu, \nu) \geq \tau \left(\frac{\varepsilon}{4}\right)^p = \tau 4^{-p} \varepsilon^p. \tag{15}$$

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:

$$W_p^p(\mu, \nu) \geq \tau 4^{-p} \left[(n/A)^{-1/s}\right]^p = \tau 4^{-p} A^{p/s} n^{-p/s}. \tag{16}$$

□

C.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 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 5 and 6 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 theorem above.

W_1 versus W_2. We use W_1 in Section 4 only for convenience: its cost is 1-Lipschitz and bounded by $\mathrm{Diam}(\rho)$, which yields better Monte Carlo constants. The estimator and its analysis carry over to $c = \|\cdot\|^2$ under the analogous assumption on W_2^2, the error then decaying as $n^{-2/d_{\mathrm{int}}}$ (the same proof applies up to constants, the condition $d_{\mathrm{int}} > 2$ of the concentration step becoming $d_{\mathrm{int}} > 4$) (the estimator becomes $2\log\eta/(\log \widehat{\mathrm{OT}}_N^{n} - \log \widehat{\mathrm{OT}}_N^{\eta n})$). In both cases the exponent is driven by the covering numbers of the support (Proposition 5), so that, e.g., under the manifold hypothesis and for n large enough, both give the manifold dimension.

D 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

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

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

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

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

Define the Monte Carlo accuracy level

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

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

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

satisfies

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

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

$$W_k^* \geq C_1 k^{-1/d_{\mathrm{int}}} \qquad \text{a.s.} \tag{17}$$

and for any $t > 0$, using $W_k^* \leq W_1(\rho, \widehat{\rho}_k^{\mathrm{unif}})$ with $\widehat{\rho}_k^{\mathrm{unif}} := \frac{1}{k}\sum_{j=1}^{k} \delta_{x_j}$ (Proposition 1) and $\mathbb{E}[W_1(\rho, \widehat{\rho}_k^{\mathrm{unif}})] \leq C_2 k^{-1/d_{\mathrm{int}}}$ (Assumption 4.1),

$$\mathbb{P}\left(W_k^* \geq C_2 k^{-1/d_{\mathrm{int}}} + t\right) \leq \mathbb{P}\left(W_1(\rho, \widehat{\rho}_k^{\mathrm{unif}}) \geq \mathbb{E}[W_1(\rho, \widehat{\rho}_k^{\mathrm{unif}})] + t\right) \leq \exp\left(-\frac{2kt^2}{\mathrm{Diam}(\rho)^2}\right),$$

where $\mathrm{Diam}(\rho)$ is the diameter of $\mathrm{Supp}(\rho)$. The last inequality is McDiarmid's bounded-differences inequality [33, Prop. 20]: replacing one support point x_j by $x_j' \in \mathrm{Supp}(\rho)$ changes $W_1(\rho, \widehat{\rho}_k^{\mathrm{unif}})$ by at most

$$|W_1(\rho, \widehat{\rho}_k^{\mathrm{unif}}) - W_1(\rho, \widehat{\rho}_k^{\mathrm{unif}\prime})| \leq W_1(\widehat{\rho}_k^{\mathrm{unif}}, \widehat{\rho}_k^{\mathrm{unif}\prime}) \leq \|x_j - x_j'\|/k \leq \mathrm{Diam}(\rho)/k.$$

Taking $t = \xi C_2 k^{-1/d_{\mathrm{int}}}$ for $\xi > 0$ yields

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

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

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

satisfies $\mathbb{P}(\mathcal{E}_{\mathrm{Q}}) \geq 1 - \delta/2$. On $\mathcal{E}_{\mathrm{Q}}$ and using (17),

$$C_1 k^{-1/d_{\mathrm{int}}} \leq W_k^* \leq (1+\xi)C_2 k^{-1/d_{\mathrm{int}}}, \qquad k \in \{n, \eta n\}. \tag{18}$$

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

Control of the population log-ratio D. Define

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

From (18),

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

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

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

hence $|D - D_0| \leq b$. Moreover, the stated condition on η implies

$$b \leq \frac{\gamma}{2(1+\gamma)} D_0.$$

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

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

Using $W_{\eta n}^* \geq C_1(\eta n)^{-1/d_{\mathrm{int}}}$ from (17), it is enough that

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

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

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

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

$$|\widehat{D} - D| \leq 2\kappa_N \left(\frac{1}{W_n^*} + \frac{1}{W_{\eta n}^*}\right). \tag{19}$$

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

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

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

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

The first lower bound on N in the theorem ensures that the RHS is at most

$$\frac{\gamma}{2(1+\gamma)} D_0.$$

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

$$|\widehat{D} - D_0| \le |\widehat{D} - D| + |D - D_0| \le \frac{\gamma}{2(1+\gamma)} D_0 + \frac{\gamma}{2(1+\gamma)} D_0 = \frac{\gamma}{1+\gamma} D_0.$$

Hence

$$\frac{1}{1+\gamma} D_0 \le \widehat{D} \le \frac{1+2\gamma}{1+\gamma} D_0.$$

In particular,

$$\widehat{D} \ge \frac{D_0}{1+\gamma}.$$

Using $\widehat{d}^* = \log \eta / \widehat{D}$, we obtain

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

Moreover, the upper bound on $\widehat{D}$ gives

$$\widehat{d}^* \ge \frac{\log \eta}{\frac{1+2\gamma}{1+\gamma} D_0} = d_{\text{int}} \frac{1+\gamma}{1+2\gamma}.$$

Therefore,

$$d_{\text{int}} \frac{1+\gamma}{1+2\gamma} \le \widehat{d}^* \le d_{\text{int}}(1+\gamma).$$

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. □

D.1 Intrinsic Dimension estimation under the manifold hypothesis

We briefly discuss here, how our proof stands under the manifold hypothesis assumption, as in [5]. We recall here their assumption.

Assumption D.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 $0 < w_{\min} \le \frac{\mathrm{d}\rho}{\mathrm{d}U_{\mathcal{M}}} \le w_{\max} < \infty$, where $U_{\mathcal{M}} := \operatorname{vol}_{\mathcal{M}} / \operatorname{vol} \mathcal{M}$ is the normalized volume measure.

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

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

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 [5] 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 7 (Upper and lower bounds on the Wasserstein error under the manifold hypothesis)**.** *There exist $A \ge 1$ and $\kappa \in (0, 1]$ depending only on $d_{\mathcal{M}}$ such that, for*

$$n \ge n_{\min} := \frac{\operatorname{vol} \mathcal{M}}{2A\, \omega_{d_{\mathcal{M}}}\, w_{\max}} (\kappa \tau)^{-d_{\mathcal{M}}},$$

any measure σ_n supported on at most n points satisfies

$$W_1(\sigma_n, \rho) \ge \frac{1}{4} \left(\frac{\operatorname{vol} \mathcal{M}}{2A\, \omega_{d_{\mathcal{M}}} w_{\max}} \right)^{\frac{1}{d_{\mathcal{M}}}} n^{-\frac{1}{d_{\mathcal{M}}}}.$$

Also [5], for a constant $C \le 48$, we have

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

Proof of the lower bound. By Lemma 3 in [6], denoting B the Euclidean ball, if $s \le \tau$,

$$B(z,s) \cap \mathcal{M} \subseteq B_{\mathcal{M}}\left(z,\, 2\tau \arcsin\left(\frac{s}{2\tau}\right)\right) \subseteq B_{\mathcal{M}}(z, 2s),$$

where the second inclusion follows from $\arcsin(u) \le 2u$ for $u \in [0, 1/2]$. Moreover, the volume comparison argument in the proof of Proposition 9 of Block et al. [5] shows that small geodesic balls satisfy

$$\mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(z,r)) \le C_{d_{\mathcal{M}}} r^{d_{\mathcal{M}}}.$$

Consequently, there exist $A \ge 1$ and $\kappa \in (0,1]$, depending only on $d_{\mathcal{M}}$, such that

$$\mathrm{vol}_{\mathcal{M}}(B(z,s) \cap \mathcal{M}) \le A\, \omega_{d_{\mathcal{M}}} s^{d_{\mathcal{M}}}, \qquad s \le \kappa\tau.$$

Let $y_1, \dots, y_n$ denote the support points of σ_n, and let $r > 0$ satisfy $2r \le \kappa\tau$. If $B(y_i, r) \cap \mathcal{M} \neq \varnothing$, choose $z_i \in B(y_i, r) \cap \mathcal{M}$. Then, by the triangular inequality

$$B(y_i, r) \cap \mathcal{M} \subseteq B(z_i, 2r) \cap \mathcal{M}.$$

Using the upper density bound,

$$\rho(B(y_i,r)) \le \frac{w_{\max}}{\mathrm{vol}\,\mathcal{M}} \mathrm{vol}_{\mathcal{M}}\big(B(z_i,2r) \cap \mathcal{M}\big) \le \frac{w_{\max} A \omega_{d_{\mathcal{M}}} (2r)^{d_{\mathcal{M}}}}{\mathrm{vol}\,\mathcal{M}}.$$

Hence

$$\rho\left(\bigcup_{i=1}^{n} B(y_i, r)\right) \le \frac{n w_{\max} A \omega_{d_{\mathcal{M}}} (2r)^{d_{\mathcal{M}}}}{\mathrm{vol}\,\mathcal{M}}.$$

Taking

$$r_n = \frac{1}{2}\left(\frac{\mathrm{vol}\,\mathcal{M}}{2A\omega_{d_{\mathcal{M}}} w_{\max} n}\right)^{1/d_{\mathcal{M}}}$$

gives

$$\rho\left(\bigcup_{i=1}^{n} B(y_i, r_n)\right) \le \frac{1}{2}.$$

The assumption $n \ge n_{\min}$ ensures that $2r_n \le \kappa\tau$. Therefore at least half of the mass of ρ lies at distance at least r_n from the support of σ_n. Any coupling between ρ and σ_n must transport this mass by at least r_n, and thus

$$W_1(\sigma_n, \rho) \ge \frac{r_n}{2} = \frac{1}{4}\left(\frac{\mathrm{vol}\,\mathcal{M}}{2A\omega_{d_{\mathcal{M}}} w_{\max}}\right)^{1/d_{\mathcal{M}}} n^{-1/d_{\mathcal{M}}}.$$

□

Using Proposition 7, we can therefore apply our Theorem 4.2 with the constants: $d_{\mathrm{int}} = d_{\mathcal{M}}$, $n_{\min}$ as in Proposition 7, any finite $n_{\max} \ge n_{\min}$, $C_1 = \frac{1}{4}\left(\frac{\mathrm{vol}\,\mathcal{M}}{2A\,\omega_{d_{\mathcal{M}}}\,w_{\max}}\right)^{\frac{1}{d_{\mathrm{int}}}}$, and $C_2 = C\left(\frac{\mathrm{vol}\,\mathcal{M}}{w_{\min}}\right)^{\frac{1}{d_{\mathrm{int}}}} \sqrt{\log\left(\frac{n_{\max}\, w_{\min}\, \mathrm{Diam}(\rho)^{d_{\mathrm{int}}}}{d_{\mathrm{int}}\ \mathrm{vol}_{\mathcal{M}}}\right)}$, the logarithmic factor of Proposition 7 being bounded by its value at $n_{\max}$ over the window. The constants thus depend on the geometry of $\mathcal{M}$, on the density of ρ, and mildly on $n_{\max}$; in particular $C_2/C_1 \propto (w_{\max}/w_{\min})^{1/d_{\mathrm{int}}}\sqrt{\log n_{\max}}$ up to factors depending on d_{int} only.

D.2 A multiscale diagnostic for Assumption 4.1

Since $\widehat{\mathrm{OT}}_N^m$ is cheap, the stability of the exponent around the chosen sizes can be checked at the same order of cost $\mathcal{O}(nN)$: evaluate $\widehat{\mathrm{OT}}_N^m$ on the geometric grid $m \in \{n/4, n/2, n, 2n, 4n\}$, compute the local slopes $\widehat{d}^*_m := \log 2/(\log \widehat{\mathrm{OT}}_N^m - \log \widehat{\mathrm{OT}}_N^{2m})$ between consecutive sizes and a global log-log slope, and report the maximal relative spread of the $\widehat{d}_m$'s. If this spread exceeds a chosen tolerance, Assumption 4.1 is violated at these scales (e.g., in the transition regime of a multiscale mixture) and $\widehat{d}^*$ should not be used for downstream calibration before increasing the sample size. The coefficient of determination of the global fit alone is not a reliable criterion, as it can remain high in a transition regime (see Table 9 below).

Numerical illustration. We run this diagnostic on a mixture of uniform measures on a 2-dimensional and an 8-dimensional cube as in Section 4, with weights 0.1 and 0.9, $N = 10^6$, 20 repetitions, a tolerance of 15%, and sizes n around the transition between the two regimes (Table 9). The local estimate is ≈ 4 at the coarsest scales ($m \approx 50$–800), then increases through the transition and stabilizes at the dimension 8 of the dominant component. The diagnostic behaves as intended: it flags every run inside the transition ($n \leq 4000$, spreads of 50–171%), and switches off at $n = 8000$ (spread 8.6%, 5% of runs flagged), where $\widehat{d}^* = 7.93$. Note that the coefficient of determination R^2 of the global fit stays above 0.88 in the middle of the transition, which is why the protocol relies on the spread of the local slopes rather than on R^2.

Table 9: Multiscale diagnostic on a 2D/8D mixture (weights $0.1/0.9$): global slope $\widehat{d}^*$ (mean $\pm$ std over 20 runs), R^2 of the global log-log fit, maximal relative spread of the local slopes, and fraction of runs flagged (tolerance 15%).

n	global $\widehat{d}^*$	R^2	slope spread	flagged runs
200	3.90 ± 0.06	0.990	50%	100%
1000	4.28 ± 0.08	0.904	144%	100%
2000	5.71 ± 0.22	0.880	171%	100%
4000	6.87 ± 0.38	0.985	56%	100%
8000	7.93 ± 0.16	0.9996	8.6%	5%

E Proof of Proposition 3: Convergence of Diagonal Richardson

Proof. We analyze the bias and stochastic deviation separately.

Let the regularization parameter scale as $\varepsilon_n = \varepsilon_0 n^{-a}$. Based on the expansions (8) and (H) (with $C_{\text{stat}} \in \mathbb{R}$ possibly zero), we have:

$$\text{Bias}_{\text{ent}}(n) = C_{\text{ent}} n^{-2a} + o(n^{-2a}) \,, \tag{20}$$

$$\text{Bias}_{\text{stat}}(n) = C_{\text{stat}} n^{-(1-ad_{\text{int}})/2} + o(n^{-(1-ad_{\text{int}})/2}) \,. \tag{21}$$

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

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

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

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

Consequently, the total expectation admits the expansion:

$$\mathbb{E}[\widehat{S}_{\varepsilon_n, n}] = W_2^2 + \underbrace{(C_{\text{ent}} + C_{\text{stat}})}_{C_{\text{total}}} n^{-\gamma} + o(n^{-\gamma}). \tag{22}$$

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

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

$$\left|\mathbb{E}[\widehat{R}_{2n}^{\text{Diag}}] - W_2^2\right| = o(n^{-\gamma}). \tag{23}$$

Using Proposition 4 from [9], we have $\mathbb{P}\left[|\widehat{S}_{\varepsilon,n} - \mathbb{E}[\widehat{S}_{\varepsilon,n}]| \geq t\right] \leq 2\exp(-nt^2/(4D^2))$, where $D := \text{Diam}(\text{Supp}(\mu) \cup \text{Supp}(\nu))^2$ (their bound is stated for supports of diameter 1 and is rescaled here). Therefore, denoting $Z_n = \widehat{S}_{\varepsilon_n,n} - \mathbb{E}[\widehat{S}_{\varepsilon_n,n}]$, 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* $\rho(\beta,\gamma)$:

$$\rho(\beta,\gamma) := \frac{w_{2n}2^{-\beta} + w_n}{1} = \frac{2^{\gamma-\beta} - 1}{2^\gamma - 1}. \tag{24}$$

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: propagation of the error on $\widehat{d}^*$. Treat $\widehat{d}^*$ as fixed and use $\widehat{a} := 1/(\widehat{d}^* + 4)$ and $\widehat{\gamma} := 2\widehat{a}$ in both the schedule and the weights. The entropic exponent $2\widehat{a}$ equals $\widehat{\gamma}$, so the entropic leading term is still cancelled exactly. The statistical exponent becomes

$$\beta = \frac{1 - \widehat{a}d_{\text{int}}}{2} = \widehat{\gamma} + \frac{1 - \widehat{a}(d_{\text{int}} + 4)}{2} = \widehat{\gamma} + \frac{\widehat{d}^* - d_{\text{int}}}{2(\widehat{d}^* + 4)}.$$

Under (H) for this schedule, the first-order statistical term after extrapolation is $C_{\text{stat}}\,\rho(\beta,\widehat{\gamma})\,n^{-\beta}$, with $\rho(\beta,\widehat{\gamma}) = \mathcal{O}(|\widehat{d}^* - d_{\text{int}}|)$ by the computation above. Hence, if $\widehat{d}^* \geq d_{\text{int}}$, then $\beta \geq \widehat{\gamma}$ and the error remains $o(n^{-\widehat{\gamma}})$, at the slightly slower rate $\widehat{\gamma} \leq \gamma$; if $\widehat{d}^* < d_{\text{int}}$ (and $\beta > 0$, i.e. $d_{\text{int}} < \widehat{d}^* + 4$), the error is $C_{\text{stat}}\rho(\beta,\widehat{\gamma})n^{-\beta} + o(n^{-\beta})$, with a leading constant proportional to $|\widehat{d}^* - d_{\text{int}}|$, which is controlled by Theorem 4.2.Moreover, when d_{int} is assumed to be an integer, as under the manifold hypothesis, one can round $\widehat{d}^*$. Thus, if $|d_{\text{int}} - \widehat{d}^*| < \frac{1}{2}$, then $\text{round}(\widehat{d}^*) = d_{\text{int}}$, and hence no estimation error occurs after rounding.

Remark: different bias conjecture. Recent results by Goldfeld et al. [18] establish a Central Limit Theorem for the Sinkhorn Divergence, noting that the expected finite-sample error scales as $\mathbb{E}[\widehat{S}_{\varepsilon,n}] - S_\varepsilon = \mathcal{O}(1/\sqrt{n} + n^{-1}r_\varepsilon)$ (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.

F Proof of Proposition 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 without replacement from $\{1, \ldots, 2n\}$. Throughout this proof, $\lambda := |w_n| = 1/(2^\gamma - 1)$ denotes the (positive) absolute value of the low-resolution weight in the Diagonal Richardson estimator, so that $w_{2n} = 1 + \lambda$ and $w_n = -\lambda$.

Common regularization. We first consider the auxiliary estimator

$$\widehat{R}_{K,2n} := (1+\lambda)\widehat{S}_{\varepsilon,2n} - \frac{\lambda}{K}\sum_{k=1}^{K}\widehat{S}^{(k)}_{\varepsilon,n},$$

in which all terms use the same fixed regularization $\varepsilon > 0$.

We analyze the asymptotic variance using the Law of Total Variance, conditioning on the observed dataset $\mathbf{D}_{2n}$. The total variance decomposes as:

$$\text{Var}(\widehat{R}_{K,n}) = \underbrace{\text{Var}\left(\mathbb{E}[\widehat{R}_{K,n} \mid \mathbf{D}_{2n}]\right)}_{\text{Term I: Dataset Variance}} + \underbrace{\mathbb{E}\left[\text{Var}(\widehat{R}_{K,n} \mid \mathbf{D}_{2n})\right]}_{\text{Term II: Bagging Variance}}. \tag{25}$$

Asymptotic Linear Expansion.

We first use that the empirical Sinkhorn Divergence is second-order Hadamard differentiable [18]. Therefore, we can use the second order Taylor expansion. The first-order term L_ε is linear and can be rewritten 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:

$$\widehat{S}_{\varepsilon,2n} = S_\varepsilon + \frac{1}{2n}\sum_{i=1}^{2n}\phi_\varepsilon(Z_i) + r_{2n}(\varepsilon), \tag{26}$$

where S_ε is the population value, $Z_i = (X_i, Y_i)$, and ϕ_ε is the canonical influence function with $\mathbb{E}[\phi_\varepsilon(Z)] = 0$ and $\text{Var}(\phi_\varepsilon(Z)) = \sigma_\varepsilon^2$.

Similarly, for any subsample $\mathcal{S}_k$ of size n, the estimator expands as:

$$\widehat{S}^{(k)}_{\varepsilon,n} = S_\varepsilon + \frac{1}{n}\sum_{j\in\mathcal{S}_k}\phi_\varepsilon(Z_j) + r^{(k)}_n(\varepsilon). \tag{27}$$

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:

$$\mathbb{E}\left[\frac{1}{n}\sum_{j\in\mathcal{S}_k}\phi_\varepsilon(Z_j)\,\middle|\,\mathbf{D}_{2n}\right] = \sum_{i=1}^{2n}\phi_\varepsilon(Z_i)\cdot\mathbb{P}(i\in\mathcal{S}_k)\cdot\frac{1}{n} = \frac{1}{2n}\sum_{i=1}^{2n}\phi_\varepsilon(Z_i). \tag{28}$$

Now, substituting the explicit expansions $\widehat{S}_{\varepsilon,2n} = S_\varepsilon + \bar{\phi}_{\varepsilon,2n} + r_{2n}(\varepsilon)$ (where $\bar{\phi}_{\varepsilon,2n} = \frac{1}{2n}\sum_{i=1}^{2n}\phi_\varepsilon(Z_i)$) and $\widehat{S}^{(k)}_{\varepsilon,n} = S_\varepsilon + \bar{\phi}_{\varepsilon,n,k} + r^{(k)}_n(\varepsilon)$ into the definition of the estimator:

$$\mathbb{E}[\widehat{R}_{K,n} \mid \mathbf{D}_{2n}] = (1+\lambda)\widehat{S}_{\varepsilon,2n} - \lambda\,\mathbb{E}[\widehat{S}^{(1)}_{\varepsilon,n} \mid \mathbf{D}_{2n}] \tag{29}$$

$$= (1+\lambda)(S_\varepsilon + \bar{\phi}_{\varepsilon,2n} + r_{2n}) - \lambda\left(S_\varepsilon + \mathbb{E}[\bar{\phi}_{\varepsilon,n,1} \mid \mathbf{D}_{2n}] + \mathbb{E}[r^{(1)}_n \mid \mathbf{D}_{2n}]\right) \tag{30}$$

$$= (1+\lambda)(S_\varepsilon + \bar{\phi}_{\varepsilon,2n} + r_{2n}) - \lambda\left(S_\varepsilon + \bar{\phi}_{\varepsilon,2n} + \mathbb{E}[r^{(1)}_n \mid \mathbf{D}_{2n}]\right). \tag{31}$$

Grouping the terms by order:

$$\mathbb{E}[\widehat{R}_{K,2n} \mid \mathbf{D}_{2n}] = \underbrace{[(1+\lambda)-\lambda]\,S_\varepsilon}_{\text{Bias}} + \underbrace{[(1+\lambda)-\lambda]\,\bar{\phi}_{\varepsilon,2n}}_{\text{Linear Term}} + \underbrace{\left((1+\lambda)r_{2n} - \lambda\,\mathbb{E}[r^{(1)}_n \mid \mathbf{D}_{2n}]\right)}_{\text{Pooled Remainder } \rho_{2n}} \tag{32}$$

$$= S_\varepsilon + \frac{1}{2n}\sum_{i=1}^{2n}\phi_\varepsilon(Z_i) + \rho_{2n}. \tag{33}$$

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

$$\mathrm{Var}\left(\mathbb{E}[\widehat{R}_{K,2n} \mid \mathbf{D}_{2n}]\right) = \mathrm{Var}(V_{2n}) + \mathrm{Var}(\rho_{2n}) + 2\mathrm{Cov}(V_{2n},\rho_{2n}). \tag{34}$$

(i) Linear variance. Since $\phi_\varepsilon(Z_i)$ are i.i.d., $\mathrm{Var}(V_{2n}) = \frac{\sigma^2_\varepsilon}{2n}$.

(ii) Remainder variance. For any fixed $\varepsilon > 0$, Remark 7 of Goldfeld et al. [18] establishes that $r_n(\varepsilon) = O_{\mathbb{P}}(n^{-1})$. Combined with the concentration inequality of Chizat et al. [9, Proposition 4] and the boundedness of the influence functions on compact supports, this yields $\mathbb{E}[r_n(\varepsilon)^2] = o(n^{-1})$ by uniform integrability. Consequently, for fixed $\varepsilon > 0$,

$$\mathrm{Var}(r_{2n}(\varepsilon)) = \mathrm{Var}(r_{2n}(\varepsilon)) = o(n^{-1})\,.$$

By Jensen's inequality and the definition of ρ_{2n}, we obtain

$$\mathrm{Var}(\rho_{2n}) \le \mathbb{E}[\rho^2_{2n}] \le 2(1+\lambda)^2\mathbb{E}[r_{2n}(\varepsilon)^2] + 2\lambda^2\mathbb{E}[r^{(1)}_n(\varepsilon)^2] = o(n^{-1}).$$

(iii) Covariance. By Cauchy-Schwarz, $\mathrm{Cov}(V_{2n},\rho_{2n}) \le \sqrt{\mathrm{Var}(V_{2n})\mathrm{Var}(\rho_{2n})} = O(n^{-1.5})$.

Summing both variance terms, we have

$$\mathrm{Var}\left(\mathbb{E}[\widehat{R}_{K,n} \mid \mathbf{D}_{2n}]\right) = \frac{\sigma^2_\varepsilon}{2n} + O(n^{-1.5}) + O(n^{-2}). \tag{35}$$

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

$$\mathrm{Var}(\widehat{R}_{K,n} \mid \mathbf{D}_{2n}) = \lambda^2\mathrm{Var}\left(\frac{1}{K}\sum_{k=1}^{K}\widehat{S}^{(k)}_{\varepsilon,n}\,\middle|\,\mathbf{D}_{2n}\right) = \frac{\lambda^2}{K}\mathrm{Var}(\widehat{S}^{(1)}_{\varepsilon,n} \mid \mathbf{D}_{2n}). \tag{36}$$

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 formula for the

variance of a sample mean under simple random sampling without replacement, the conditional variance of the linear term is:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{j\in\mathcal{S}_1}\phi_\varepsilon(Z_j)\,\middle|\,\mathbf{D}_{2n}\right)=S^2_{\phi_\varepsilon,2n}\left(\frac{1}{n}-\frac{1}{2n}\right)+o\left(\frac{S^2_{\phi_\varepsilon,2n}}{n}\right). \tag{37}$$

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, $\mathbb{E}[S^2_{\phi_\varepsilon,2n}]=\mathrm{Var}(\phi_\varepsilon(Z))=\sigma^2_\varepsilon$ exactly. Therefore,

$$\mathbb{E}\left[\mathrm{Var}(\widehat{R}_{K,n}\mid\mathbf{D}_{2n})\right]=\frac{\lambda^2}{K}\sigma^2_\varepsilon\left[\frac{1}{2n}\right]+o\left(\frac{\sigma^2_\varepsilon}{n}\right)=\frac{\lambda^2}{K}\frac{\sigma^2_\varepsilon}{2n}+o\left(\frac{\sigma^2_\varepsilon}{n}\right). \tag{38}$$

Using that, by boundedness of the cost function on bounded support measures, we have $\sigma_\varepsilon\lesssim 1$, independently of ε.

Recalling that $\mathrm{Var}(\widehat{S}_{\varepsilon_{2n},2n})=\sigma^2_\varepsilon/(2n)+o(1/n)$ and $\lambda^2=1/(2^\gamma-1)^2$, this proves

$$\mathrm{Var}(\widetilde{R}_{K,2n})=\Big(1+\frac{1}{(2^\gamma-1)^2K}\Big)\mathrm{Var}(\widehat{S}_{\varepsilon_{2n},2n})+o\Big(\frac{1}{n}\Big). \tag{39}$$

.

Taking ε and ε'. For two probability measures α,β supported on at most n points, every $\pi\in\Pi(\alpha,\beta)$ satisfies $0\le\mathrm{KL}(\pi\,|\,\alpha\otimes\beta)\le\log n$, since this quantity is the mutual information of the coupling, bounded by the entropy of α. Hence, for $\varepsilon'\le\varepsilon$, writing $\pi_{\varepsilon'}$ for the optimal plan at level ε',

$$\mathrm{OT}_{c,\varepsilon'}(\alpha,\beta)\le\mathrm{OT}_{c,\varepsilon}(\alpha,\beta)\le\int c\,d\pi_{\varepsilon'}+\varepsilon\mathrm{KL}(\pi_{\varepsilon'}\,|\,\alpha\otimes\beta)\le\mathrm{OT}_{c,\varepsilon'}(\alpha,\beta)+(\varepsilon-\varepsilon')\log n.$$

Applying this to the cross term and to the two self-terms of the Sinkhorn divergence gives, almost surely,

$$\big|\widehat{S}_{\varepsilon,n}-\widehat{S}_{\varepsilon',n}\big|\le 2|\varepsilon-\varepsilon'|\log n. \tag{40}$$

Consequently, the actual estimator (10) satisfies, almost surely,

$$\big|\widehat{R}^{\mathrm{Bagged}}_{K,2n}-\widetilde{R}_{K,2n}\big|\le\frac{\lambda}{K}\sum_{k=1}^{K}\big|\widehat{S}^{(k)}_{\varepsilon_n,n}-\widehat{S}^{(k)}_{\varepsilon_{2n},n}\big|\le 2\lambda|\varepsilon_n-\varepsilon_{2n}|\log n.$$

For square-integrable U,V, the standard deviation being a seminorm, $|\sqrt{\mathrm{Var}(U)}-\sqrt{\mathrm{Var}(V)}|\le\|U-V\|_{L^2}$, hence

$$\big|\mathrm{Var}(U)-\mathrm{Var}(V)\big|\le 2\sqrt{\mathrm{Var}(V)}\,\|U-V\|_{L^2}+\|U-V\|^2_{L^2}.$$

We apply it with $U=\widehat{R}^{\mathrm{Bagged}}_{K,2n}$ and $V=\widetilde{R}_{K,2n}$, for which $\mathrm{Var}(V)=O(1/n)$ by (39) and $\sigma_\varepsilon\lesssim 1$. We get

$$\mathrm{Var}\big(\widehat{R}^{\mathrm{Bagged}}_{K,2n}\big)=\Big(1+\frac{1}{(2^\gamma-1)^2K}\Big)\mathrm{Var}(\widehat{S}_{\varepsilon_{2n},2n})+O\big(|\varepsilon_1-\varepsilon_2|n^{-\frac{1}{2}}\log n+|\varepsilon_1-\varepsilon_2|^2\log^2 n\big)+o\Big(\frac{1}{n}\Big).$$

The additional term is a crude pointwise control of the change of regularization and only yields an additive guarantee. A multiplicative one would require, e.g., a quantitative comparison of the Sinkhorn influence functions at the two regularization levels, which we are not aware of. □

G Complements to Section 3: Entropic Counterpart

For $\varepsilon>0$, $\mathbf{a}\in\Delta_n$ and $u_n:=\frac{1}{n}\sum_i\delta_{x_i}$, let

$$\mathrm{OT}^u_{c,\varepsilon}(\rho,\widehat{\rho}^{\mathbf{a}}_n):=\min_{\pi\in\Pi(\rho,\widehat{\rho}^{\mathbf{a}}_n)}\left\{\int c\,d\pi+\varepsilon\mathrm{KL}(\pi\,|\,\rho\otimes u_n)\right\},$$

which coincides with (4) for uniform weights $\mathbf{a}=(1/n,\dots,1/n)$, and define the soft c-transform

$$\mathbf{g}^{c,\varepsilon}(x):=-\varepsilon\log\left(\frac{1}{n}\sum_j e^{(g_j-c(x,x_j))/\varepsilon}\right).$$

The corresponding entropic semi-dual is

$$\mathrm{OT}^u_{c,\varepsilon}(\rho, \widehat{\rho}^{\mathbf{a}}_n) = \sup_{\mathbf{g}\in\mathbb{R}^n} \left\{ \int \mathbf{g}^{c,\varepsilon}(x)\, d\rho(x) + \sum_i a_i g_i \right\}.$$

Its objective is invariant under addition of a constant to $\mathbf{g}$, strictly concave on $\mathbf{1}_n^\perp$ [3], and its gradient is given by

$$a_i - \int \frac{e^{(g_i - c(x,x_i))/\varepsilon}}{\sum_j e^{(g_j - c(x,x_j))/\varepsilon}}\, d\rho(x), \qquad i \in [\![1, n]\!].$$

Proposition 8. *Under the assumptions of Proposition 1, with pairwise distinct x_i's, let $\varepsilon > 0$,*

$$w_i^\varepsilon := \int \frac{e^{-c(x,x_i)/\varepsilon}}{\sum_j e^{-c(x,x_j)/\varepsilon}}\, d\rho(x), \qquad \widehat{\rho}^\varepsilon_n := \sum_i w_i^\varepsilon \delta_{x_i}.$$

Then $\widehat{\rho}^\varepsilon_n$ is the unique minimizer of $\mathbf{a} \mapsto \mathrm{OT}^u_{c,\varepsilon}(\rho, \widehat{\rho}^{\mathbf{a}}_n)$ over Δ_n, $\mathbf{0}_n$ maximizes the semi-dual

$$\mathbf{g} \mapsto \int \mathbf{g}^{c,\varepsilon}\, d\rho + \sum_i w_i^\varepsilon g_i,$$

and

$$\mathrm{OT}^u_{c,\varepsilon}(\rho, \widehat{\rho}^\varepsilon_n) = \int \mathbf{0}_n^{c,\varepsilon}\, d\rho.$$

Remark. As in the unregularized case, $\frac{1}{N}\sum_k \mathbf{0}_n^{c,\varepsilon}(X_k)$ is then a solver-free Monte Carlo estimator of $\mathrm{OT}^u_{c,\varepsilon}(\rho, \widehat{\rho}^\varepsilon_n)$, to which Proposition 2 applies.

Proof. By definition of $\mathbf{w}^\varepsilon_n$, the gradient of the semi-dual objective at $\mathbf{0}_n$ vanishes. Hence, by its concavity, $\mathbf{0}_n$ is a maximizer, and

$$\mathrm{OT}^u_{c,\varepsilon}(\rho, \widehat{\rho}^\varepsilon_n) = \int \mathbf{0}_n^{c,\varepsilon}\, d\rho.$$

Moreover, for every $\mathbf{a} \in \Delta_n$, evaluating its semi-dual at $\mathbf{0}_n$ gives

$$\mathrm{OT}^u_{c,\varepsilon}(\rho, \widehat{\rho}^{\mathbf{a}}_n) \geq \int \mathbf{0}_n^{c,\varepsilon}\, d\rho = \mathrm{OT}^u_{c,\varepsilon}(\rho, \widehat{\rho}^\varepsilon_n),$$

so $\mathbf{w}^\varepsilon_n$ is a minimizer. If equality holds for some $\mathbf{a} \in \Delta_n$, then $\mathbf{0}_n$ also maximizes its semi-dual. Its gradient at $\mathbf{0}_n$ must therefore vanish, which, by the gradient formula above, gives $\mathbf{a} = \mathbf{w}^\varepsilon_n$. Hence the minimizer is unique. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Proofs and numerical experiments are given to support our claims.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations, such as assumption, or limits of our theory are given in the main text.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions and proofs are given for all the results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details on the experiments are given in the appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Details on the experiments are sufficent to reproduce similar results. Moreover, a code repository will be given once the paper is no more under anonymous review.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All parameters, like the number of points or the regularization are stated.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiments are reproduce several time, and the averaged results are displayed, along their standard deviations.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As stated in the appendix, all experiments were executed on Google Colab using the freely available GPU runtime.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It is methodological, uses synthetic data only, and does not involve human subjects, personal data, sensitive attributes, or deployed systems.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is theoretical and methodological, uses only synthetic data, and does not propose a deployed system or application-specific decision-making pipeline. Therefore, a discussion of broader societal impacts is not applicable.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Non applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All paper and libraries used are cited.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Non applicable.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Non applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]
Justification: Non applicable.