arXiv:2602.04335v1 [stat.ML] 4 Feb 2026

几何感知最优传输：快速内在维数与 Wasserstein 距离估计

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

摘要

在机器学习中求解大规模最优传输（OT），通常依赖对测度进行采样，以获得易于处理的离散问题。虽然离散求解器的精度可控，但离散化误差的收敛率取决于数据的内在维数。因此，真正的瓶颈在于了解并控制采样误差。本文针对这一问题，提出了用于估计采样误差和内在维数的新方法。我们的关键发现是一个简单且无需调参的 $\mathrm{OT}_c(\rho, \widehat{\rho})$ 估计量，它利用最优传输半对偶泛函，而且无需最优传输求解器。此外，我们根据采样误差估计量的多尺度衰减，推导出一种快速内在维数估计量。该框架在实践中带来显著的计算和统计优势，使我们能够：(i) 量化离散化误差的收敛率；(ii) 根据数据的内在几何校准 Sinkhorn 散度的熵正则化；以及 (iii) 引入一种基于内在维数的全新 Richardson 外推估计量，大幅消除 Wasserstein 距离估计中的偏差。数值实验表明，我们的几何感知流程能够有效缓解离散化误差瓶颈，同时保持计算效率。

1. 引言

许多现代机器学习对象最自然的建模方式是将其视为概率分布：例如图像（像素强度的分布）、点云、词嵌入，或通过其采样测度来观察的经验数据集。最优传输（OT）提供了一种有原则的方法来比较和操控这类对象，它旨在寻找以最低成本将质量从一个分布转移到另一个分布的方案。通过这一表述，我们可以得到一系列 Wasserstein（推土机）距离及其相关的传输计划；它们已成为复杂机器学习流程中的有效构建模块（Peyr´等人，2019年），应用包括生成建模（An等人，2019年；Chen等人，2019年）、加速扩散模型（Li等人，2023年）、领域自适应（Courty等人，2014）以及自然语言处理（Kusner等人，2015年）。

这些应用通常面向大规模场景：数据可能来自连续分布，也可能来自支撑集包含数量极其庞大点的离散分布。在实践中，人们因此会使用通过采样获得的离散化近似。这个简单步骤以经验测度替代未知测度，由此推动了两个互补的研究方向：（i）统计最优传输，用于量化这种离散化所引入的误差；（ii）计算最优传输，旨在设计能够处理由此产生的大规模离散问题的求解器。

理论中的统计最优传输。一个核心问题是，基于样本计算的最优传输量能在多大程度上近似其总体对应量。由（Dudley，1969年）开创的一系列研究分析了1-Wasserstein差异 $W_1(\rho_n, \rho)$：当连续测度 ρ 被由 n 个带均匀权重的独立同分布样本构造的经验测度 ρ_n 取代时，这一差异如何变化。随后，（Fournier与Guillin，2015年）将这一结果推广，表明对于所有 $p \geq 1$，$W_p(\rho_n, \rho)$通常按 $\mathcal{O}(n^{-1/d})$的速率随 $\mathbb{R}^d$ 变化，明确揭示了维数灾难：要准确近似最优传输量，所需样本数会随维数呈指数增长。正如（H¨utter与Rigollet，2021年）所证明的，估计传输映射本身也受到这一灾难的影响。若干策略可减轻这些影响：利用低内在维数结构（Weed & Bach，2017年），或聚焦于半离散最优传输等特殊情形（Pooladian等人，2023年）。关于这一主题的全面综述，可参阅近期出版的著作（Chewi等人，2024年）。

[1]LPSM，索邦大学，巴黎，法国 [2]沃尔夫冈·泡利研究所，维也纳，奥地利。联系邮箱：Ferdinand Genans< genans.ferdinand@gmail.com>。

预印本。*2026*年*2*月*5*日。

最优传输与内在维数。缓解维数灾难的一个关键洞见是，现实世界中的数据很少会完全占据环境空间 $\mathbb{R}^d$。正如（Weed & Bach，2017年）所证明的，若测度 ρ支撑在一个 $\mathcal{M}$ 内在维数较低的流形 $d_{\mathcal{M}} \ll d$上，估计误差便会显著降低至 $\mathcal{O}(n^{-1/d_{\mathcal{M}}})$。此外，最优传输呈现出良好的多尺度行为：近似误差并非由单一的全局维数决定，而是随支撑集在不同半径下的覆盖数而变化，因此往往能取得更优的非渐近表现。至关重要的是，使用最优传输比较两个不同测度时，其复杂度会适应两个测度中较小的内在维数（Hundrieser 等人，2024年）；这些性质也适用于熵正则化情形（Stromme，2024年）。

计算最优传输。在算法方面，离散最优传输可以表述为线性规划，但通用线性规划求解器在大规模问题上代价高昂。最优传输在机器学习中的实际影响，源于利用问题结构的专用算法。一个转折点是（Cuturi，2013年），其提出了熵正则化最优传输（EOT），可通过 Sinkhorn 迭代高效求解，复杂度为 $\mathcal{O}(n^2/\varepsilon^2)$，从而以 ε 精度逼近离散最优传输。后续研究改进了常数和收敛保证，包括原始-对偶方法（Dvurechensky 等人，2018年；Lin 等人，2019年），以及近期实现$\mathcal{O}(n^2/\varepsilon)$（Blanchet 等人，2023年）的进展；这似乎已是此类问题的最优结果。

Sinkhorn 散度：统计与计算之间的桥梁及其局限。弥合统计最优传输与计算最优传输之间差距的一种合理尝试，是研究 Sinkhorn 散度（Ramdas 等，2017年；Genevay 等，2018年；Feydy等人，2019年）。通过明确地以离散 Sinkhorn 散度定义差异，可以达到参数化估计速率 $\mathcal{O}(\varepsilon^{-d/2}n^{-1/2})$（Genevay等，2019年；Mena与Niles-Weed，2019年；Chizat等人，2020）。因此，维数灾难并未消除，而是转移到了正则化参数 ε 上。关键在于，d 这一控制权衡的指数是数据流形的内在维数 $d_{\mathcal{M}}$（Stromme，2024），并且经过适当校准的速度使 $\varepsilon \to 0$ 几乎能够达到估计 Wasserstein 距离的极小极大速率。

尚缺的一环。尽管取得了这些进展，统计理论与计算实践之间仍缺乏一条切实可行的纽带。虽然 Sinkhorn 散度提供了一种理论上的权衡，但要对其进行恰当校准，就需要知道通常未知的底层内在维数 $d_{\mathcal{M}}$，以便选取最优的 ε。因此，实践者缺乏具体的指导：他们往往将数值求解器运行到极高精度，暗中假定采样误差可以忽略；或者凭启发式方法调整 ε，却没有处理偏差-方差权衡。若要以经过校准且依赖于具体问题的精度应用最优传输，就需要一种实用方法，能够直接根据现有样本估计离散化误差和内在维数。

贡献。为应对这些挑战，我们提出了一种几何感知框架，可直接从样本估计离散化误差和内在维数，而不依赖昂贵的真实值计算。我们的具体贡献如下：

- **无求解器离散化误差估计器**。我们提出了一种新颖的估计量，用于估计量化误差 $\mathrm{OT}_c(\rho,\ \rho_n^*)$，其中 ρ_n^* 表示权重经过最优调整、以最小化传输代价的经验测度。我们的关键理论洞见是，该问题的半对偶形式存在闭式解。因此，我们可以通过简单的蒙特卡洛积分估计误差，完全无需使用最优传输求解器。所得估计量无需调参，可在图形处理器上高度并行化，且计算开销低。
- **一种可扩展的内在维数估计量**。利用误差估计量的收敛特性，我们推导出数据的经验内在维数估计量 d_{int}。通过分析误差随样本量变化的衰减情况，我们的估计量能够捕捉分布的多尺度几何结构，这与（Weed & Bach，2017）中的理论发现一致。关键在于，该方法在时间和空间上的复杂度均呈线性增长（即 $\mathcal{O}(n)$）$\mathcal{O}(n)$，因此适用于大规模数据集。
- **通过对角 Richardson 外推进行偏差校正的 Wasserstein 估计**。我们提出一种新的外推方案，将内在维数估计与 Sinkhorn 散度相结合。以往的工作（Chizat 等人，2020年）仅对正则化参数 ε使用 Richardson 外推，而我们引入了一种对角方法，将 ε 与样本量 n 和内在维数 d_{int}关联起来。这种联合去偏策略有效消除了一级熵偏差和统计离散化误差。该方案使我们能够在 W_2^2下达到$o(n^{-2/(d_{\mathrm{int}}+4)})$的收敛速度；相比之下，仅针对 ε的 Richardson 方案在 o下的收敛速度为$n^{-2/(d+8)}$。

我们在合成流形和真实世界数据集（MNIST、CIFAR）上验证了我们的流程，结果表明，我们的维数估计量准确，且与标准基线相比，我们的外推方案所得的 Wasserstein 估计值偏差显著降低。

记号。我们用 $\|\cdot\|$ 表示欧几里得范数，用 ω_d 表示单位 d维球的体积。对于 $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$ 和 $a \wedge b := \min\{a, b\}$. $\mathcal{P}(\mathbb{R}^d)$ 表示 $\mathbb{R}^d$中的概率集合；对于 $\rho \in \mathcal{P}(\mathbb{R}^d)$，Supp($\rho$) 表示其支撑集。 Δ_d表示 $\mathbb{R}^d$中的概率单纯形： $\Delta_d := \{w \in \mathbb{R}_+^d : \sum_{i=1}^d w_i = 1\}$。 $\mathcal{O}(\cdot)$ 和 $o(\cdot)$ 表示通常的近似阶。若存在一个通用常数 $C > 0$，使得 $f(\cdot) \leq Cg(\cdot)$，则记作 $f \lesssim g$。若 $a \lesssim b$ 和$b \lesssim a$均成立，则记作 $a \asymp b$。

2. 预备知识

2.1. 最优传输：原始问题、对偶问题与半对偶

我们考虑支撑于紧致度量空间 $\mathcal{X} \subset \mathbb{R}^d$ 上的概率测度。给定源测度和目标测度μ， $\nu \in \mathcal{P}(\mathcal{X})$，以及代价函数 c: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$，Kantorovich 最优传输问题定义如下:

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

其中， $\Pi(\mu, \nu)$ 表示边缘分布为 μ和 ν 的联合分布集合。当对于 $p \geq 1$， $c(x, y) = \|x - y\|^p$ 时，$\mathrm{OT}_c(\mu, \nu)^{1/p}$ 定义了 p-Wasserstein 距离，记作 $W_p(\mu, \nu)$（Villani，2009）。

对偶与半对偶形式　　化。Kantorovich问题具有一个涉及连续势的对偶形式
函数 f， $g \in C(\mathcal{X})$:

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

通过将势 g的 c变换定义为 $g^c(x)$: $= \inf_y[c(x, y) - g(y)]$,，便可通过设定 $f = g^c$来消除约束 $f(x) \leq c(x, y) - g(y)$。由此得到**半对偶形式**，其仅依赖于单个势 g:

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

这种半对偶形式对我们的分析至关重要，因为即使 μ是连续的，只要 ν 是离散的，就能将受约束的连续问题转化为有限维问题；我们将利用这一性质来构造误差估计量。

2.2. 熵正则化与Sinkhorn散度

为克服精确最优传输的计算代价，通常采用熵正则化。我们通过加入Kullback–Leibler惩罚，以其原始问题形式定义正则化目标:

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

其中 $\varepsilon > 0$ 是正则化强度。为校正熵引入的偏差（此时 $\mathrm{OT}_{c,\varepsilon}(\mu, \mu) \neq 0$），我们可以使用 **Sinkhorn 散度**:

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

对于 $c = \|\cdot\|^2$，它是一种在 OT（当 $\varepsilon \to 0$ 时）与最大均值差异（当 $\varepsilon \to \infty$ 时）之间插值的散度（Genevay等人，2018年；Feydy等人，2019年）。

2.3. 统计速率与内在维数

通过经验测度 $\widehat{\mu}_n$ 和 $\widehat{\nu}_n$ 逼近 μ 和 ν 时，经典结果表明，离散化误差的尺度取决于环境维度 d。粗略地说，对于 p-Wasserstein 代价和 Sinkhorn 散度:

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

关键的是，后续研究证实 d 并非环境维度，而是这两个测度的**最小内在维数**，这一结论对最优传输（Weed & Bach, 2017；Hundrieser 等，2024年）和熵正则化最优传输（Groppe & Hundrieser，2024年；Stromme，2024年）均成立。更准确地说，收敛速率由 ϵ-覆盖数 $\mathcal{N}_\epsilon$ 决定，其定义的有效维数为 $d_\epsilon \approx \min_{\rho\in\mu,\ \nu} \log \mathcal{N}_\epsilon(\rho) - \log \epsilon$。该量呈现出多分辨率行为：最优传输和熵正则化最优传输会适应不同观测尺度下的几何结构，在离散化误差仍然显著的非渐近区间内捕捉 d_ε。因此，在 $d_\epsilon > \lim_{\varepsilon\to 0} d_\epsilon$ 的情形下，估计量能够利用粗尺度下的有效几何结构。

3. Wasserstein 离散化误差的简单估计量

如第 2 节所述，机器学习中最优传输的可靠性取决于能否控制以离散样本替代连续测度时产生的误差。Wasserstein 距离的三角不等式为理解这一误差提供了一个基本见解。对于任意两个测度 μ，ν 及其离散近似 $\widehat{\mu}$，$\widehat{\nu}$，估计误差有如下界：

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

这一分解表明，总误差取决于离散样本捕捉底层分布几何结构的能力。渐近理论预言其收敛速率为 $n^{-1/d}$，但这些界限对于有限数据集的实际指导作用有限。直接根据数据估计 $\rho=\mu$，ν 的 $\mathrm{OT}_c(\rho,\ \widehat{\rho})$ 有两个重要作用：(i) 它对近似质量进行具体的非渐近量化；(ii) 该误差随样本量增加而衰减的速率，可作为测度内在维数 的指标。

在本节中，我们将介绍一种快速、无需求解器的估计量，用于估计单样本离散化误差 $\mathrm{OT}_c(\rho,\ \widehat{\rho}_n^*)$，其中 $\widehat{\rho}_n^*$表示在给定一个包含 n 个点的固定支撑集时，ρ 的最优离散近似。

3.1. Wasserstein离散化误差与半离散最优传输

考虑一个测度 ρ和一个固定支撑集 $X=(X_1,\ \ldots,\ X_n)$，从 ρ中独立同分布地采样得到。对于任意权重向量 $\mathbf{a}\in\Delta_n$，令$\widehat{\rho}_{m,\ \mathbf{a}}=\sum_{i=1}^n a_i\delta_{X_i}$ 为对应的离散测度。计算 最优传输$_c(\rho,\ \widehat{\rho}_{m,\ \mathbf{a}})$ 构成一个半离散最优传输问题。即使 ρ是连续的，该问题仍具有有限维半对偶形式（3）：

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

求解（6）通常需要耗费大量计算的随机梯度下降或半离散求解器（Genevay等人，2018年；Kitagawa等人，2016），因此，如果目的仅仅是在使用最优传输$_c(\widehat{\mu}_n,\ \widehat{\nu}_n)$ 作为最优传输$_c(\mu,\ \nu)$ 的计算可行替代方案时评估离散化质量，那么这种方法并不实用。值得注意的是，以下命题表明，我们可以完全绕过优化。通过为固定支撑集选取最优权重，对偶势函数将退化为零向量，从而得到闭式解。

命题 3.1. 设 c 为连续代价函数。给定 $\rho\in\mathcal{P}(\mathbb{R}^d)$ 和一个支撑集 $X=(x_1,\ldots,x_n)\in(\mathbb{R}^d)^n$，定义权重 $\mathbf{w}_n=(w_1,\ldots,w_n)^\top$ 为

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

a并令集合 $\widehat{\rho}_n^*=\sum_{i=1}^n w_i\delta_{x_i}$。那么，$\widehat{\rho}_n^*$ 给出了支撑集位于 X 上的 ρ 的最佳逼近，也就是说 $s,$

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

且极小化器是唯一的。此外，零向量 $\mathbf{0}_n$ 是半对偶 (6) 在 $\widehat{\rho}_n^*$处的极大化器。也就是说，

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

证明。 根据$\mathbf{w}_n$的定义，我们有向量 $\mathbf{0}_n$ 是 ρ 与 ρ_n之间半离散最优传输问题的Kantorovich势。事实上，一阶最优性条件指出，$\mathbf{g}$ 最优当且仅当 $\nabla H(\mathbf{g}) = \mathbf{0}_n$，即$\rho(\{\mathbf{g}^c(x) = c(x,X_i) - g_i\}) = w_i$，这正是我们在固定 $g_i=0$、$i\in[\![1,n]\!]$时定义权重 w_i的方式。

由于 $\mathbf{0}_n$ 是其Kantorovich势，对所有 $x \in \mathbb{R}^d$，最优传输为： $x \mapsto X_i = \text{argmin}_{X_i} c(x, X_i)$。也就是说，每个点都会被送到 X中的最近邻。对于任何其他概率权重向量，另一个$\mathbf{g}' \in \mathbb{R}^n$ 满足$\mathbf{g}' \notin \{\lambda \mathbf{1}_n, \lambda \in \mathbb{R}\}$，也会是一个最优向量；这将导致一个最优传输方案，其中 ρ 有严格为正的质量未被送到其最近邻，因此增加了最优传输代价。

□

3.2. 无需求解器的蒙特卡洛估计器

命题 3.1 的关键见解是，对于最优加权测度 $\widehat{\rho}_n^*$，半对偶目标函数在 $\mathbf{0}_n$ 处达到最大值。这使最优传输代价简化为零向量的 c-变换的积分，而该积分恰好就是支撑集 X 中到最近邻的距离。因此，我们可以使用直接的蒙特卡洛（MC）估计器 $\widehat{\ }$，通过使用 N 个额外样本 $X_1, \ldots\ldots, X_N \overset{\text{独立同分布}}{\sim} \rho$ 来估计最优传输 $_c(\rho, \widehat{\rho}_n^*)$:

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

该估计器计算效率高，可在图形处理器上完全并行化，且无需最优传输求解器。此外，如果需要显式权重，也可以用自然的蒙特卡洛估计器来估计:

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

以下定理证明，这些估计量以指数速度集中于其真实值附近。

命题3.2。 假设 ρ 具有有界支撑集，且代价函数连续。那么对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，有:

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

其中 $C_\rho =$ 上确界$_{x,\ y\in supp(\rho)}\ c(x,y)$ 而 $\widehat{\sigma}^2_{\mathbf{0}^c}$ 是常用的有偏样本方差估计量

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

此外，对于任意 $i \in [\![1, n]\!]$，以概率 $1-\delta$，有

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

证明。 在我们的假设下，可以几乎必然地将 $\mathbf{0}_n^c(\cdot) =$ 最小值$_{j\in[\![1,\ n]\!]}\ c(\cdot,\ x_j)$ 界定为 C_ρ。因此，(Maurer & Pontil, 2009) 中定理4的经验伯恩斯坦界给出了第一个不等式。由于对所有 i， $w_i \in [0,1]$，同一定理也适用于第二个不等式。

□

最后，我们总结两点实践观察。第一，尽管该界取决于直径 C_ρ，但这一项呈线性出现，不会影响收敛率。第二，虽然标准做法通常假设权重均匀分布（$\widehat{\rho}_n^{\text{unif}} = \frac{1}{m}\sum \delta_{x_i}$），但我们的估计量 $\widehat{\ }\mathrm{OT}_N^n$ 针对的是最优加权测度 $\widehat{\rho}_n^*$。由于 $\mathrm{OT}_c(\rho,\ \widehat{\rho}_n^*) \le \mathrm{OT}_c(\rho,\ \widehat{\rho}_n^{\text{unif}})$，我们的方法为支撑集 X上可达到的离散化误差提供了紧下界。更重要的是，由于 $\widehat{\rho}_n^*$ 会根据支撑集的几何特性自适应地调整质量，$\mathrm{OT}_c(\rho,\ \widehat{\rho}_n^*)$ 随着 n 增大而呈现的衰减率，完全由 ρ的内在几何决定。这一性质使 OT_N 成为我们快速内在维数估计量的关键构件，详见下一节。

4. 离散化误差与内在维度估计

4.1. 几何设定：多尺度行为与覆盖数

机器学习中一个反复出现的主题是，高维数据往往呈现出低维结构。尽管流形假设提出数据集中在光滑子流形 $\mathcal{M} \subset \mathbb{R}^d$上，为我们提供了有用的基线，但现实世界的数据很少会严格遵循这类理想化假设。相反，复杂分布往往表现出多分辨率行为，其有效几何复杂度取决于观测尺度或所需的近似精度。

为说明这一点，考虑由以下混合物构造的 $\rho \in \mathcal{P}(\mathbb{R}^{10})$： $\rho := \frac{1}{2}\mathcal{U}([0,1]^2 \times \{0\}^8) + \frac{1}{2}\mathcal{U}([1,2]^8 \times \{0\}^2)$。严格来说，其支撑集的拓扑维数为 8。然而，最优传输离散化误差的表现揭示了更为微妙的现实。如果目标是进行粗略近似（例如，离散化误差为 50%），该问题的实际表现可能类似于一个二维问题。相反，若要求更高的精度（例如误差小于 10%），则必须使离散化覆盖整个八维体积，从而使误差率受到与 $d = 8$ 相关的维数灾难影响。

这个例子凸显了，对于概率测度，相关的复杂度概念并非一个静态整数，而是一个动态量，它与支撑集在不同半径下的覆盖数有关（Weed & Bach，2017）。因此，我们不再采用僵化的几何定义（例如流形假设），而是改用基于 Wasserstein 距离本身收敛率的函数性假设。

我们假设，对于特定的样本量范围 $n \in [n_{\min}, n_{\max}]$，离散化误差由有效内在维度 d_{int} 控制。

假设4.1（有效内在维度）。设 $\rho \in \mathcal{P}(\mathbb{R}^d)$ 是一个具有有限直径的概率测度。我们假设，对于 $n \in [n_{\min}, n_{\max}]$，1-Wasserstein距离离散化误差的收敛率由维度参数 $s = d_{\text{int}}$ 控制。具体而言，存在常数 C_1， $C_2 > 0$ 其本质上取决于 Diam(ρ)，使得 $C_2 \le cC_1$对于某个普适常数 $c \ge 1$，并且：

下界（量化极限）： 对于任意至多由 n 个点支撑的离散测度 σ_n，其近似误差的下界由维数决定：

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

上界（统 计性能）： 由 n 个独立同分布样本构造的经验测度 $\widehat{\rho}_n$ 满足 ：

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

对所有 $s' \ge s$。我们将指数 s 称为 ρ 在 $[n_{\min}, n_{\max}]$ 区间内的**内在维数** d_{int}。

这一假设认为， d_{int} 刻画了用有限个点近似 ρ 的 Wasserstein 离散化误差的“难度”。在渐近极限（$n_{\min} \to \infty$）下，如果 ρ 相对于一个维数为 $d_{\mathcal{M}}$ 的流形具有密度，那么 s 与 $d_{\mathcal{M}}$ 重合。关于这一点的进一步讨论，以及特定情形下常数的界，请参阅附录。

4.2. 利用我们的离散化估计器估计内在维数。

为估计内在维数，我们利用假设4.1中确立的单样本离散化误差的缩放行为。使用第 ⌒节中定义的蒙特卡洛估计器$^{n}_{N}$最优传输（ $c(x,y) = \|x - y\|$，我们在两个不同的支撑规模3处评估误差，分别为 n 和 ηn （其中 $\eta > 1$）。这些估计值的比率揭示了决定收敛率的指数，从而得到以下估计量：

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

⌒其中， N 表示用于蒙特卡洛近似的样本数。计算 d^* 只需进行两次蒙特卡洛估计，因此其复杂度关于 n为线性。我们方法的统计保证由以下定理给出。

定理4.2。假设 *th* 在假设 *4.1* 成立且 $d_{\text{int}} > 2$ 并令 *Diam*(ρ) 表示 ρ 的支撑集直径 *t.*

那么，对于 γ、$\xi > 0$ 和 $n\eta \le n_{\max}$，利用

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

我们以概率 $1-\delta$ 有

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

与文献的比较。 利用 Wasserstein 误差的多尺度衰减进行维数估计这一概念由（Block 等人，2022年）率先提出。然而，其估计量依赖于双样本经验距离：

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

其中 $\widehat{\rho}_n$ 与 $\widehat{\rho}'_n$ 是 n 个点的独立经验测度。该方法基于如下事实：$\mathbb{E}[W_1(\widehat{\rho}_n, \widehat{\rho}'_n)] \approx \mathbb{E}[W_1(\widehat{\rho}_n, \rho)]$,，但它面临严重的计算瓶颈：计算 $W_1(\widehat{\rho}_n, \widehat{\rho}'_n)$ 需要求解一个完整的离散最优传输问题。标准线性规划求解器的复杂度按 $\mathcal{O}(n^3)$ 增长，而近似 Sinkhorn 求解器的复杂度按 $\mathcal{O}(n^2/\varepsilon)$ 增长。这一过高的成本将此类估计限制于小型数据集，或需要大量近似，从而可能使维数估计产生偏差。

$\widehat{\ }$相比之下，我们的估计量 d^* 依据到最优支撑集的单样本误差，并通过蒙特卡洛方法近似计算，无需进行任何优化。由此得到的复杂度为线性 $\mathcal{O}(n)$，因此可在大规模数据集上进行维数估计，而标准最优传输求解器在这类数据集上难以处理。

4.3. 实验

我们将自己的估计量（记为半离散 W_1）与基于（Block 等人，2022年）的离散 W_1 估计量，以及scikit-dimension 软件包（Bac 等人，2021年）中的标准基线进行比较：最大似然估计、CorrInt、局部主成分分析、TwoNN。

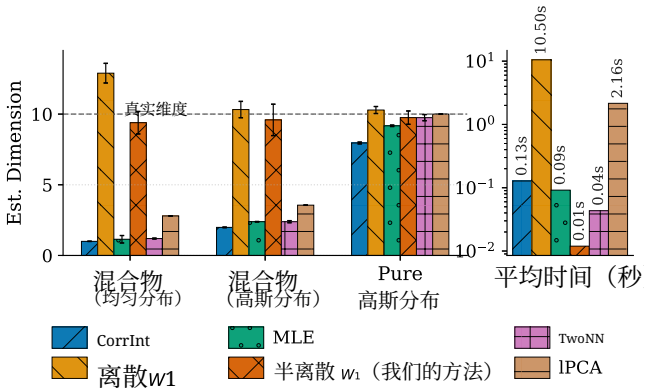


图1。 内在维数估计基准。比较我们的半离散 W_1 估计量与离散 W_1 基线及标准几何估计量。虚线表示真实的有效维数 $d_{\text{int}} = 10$。在混合分布上，我们的估计量具有与离散最优传输相当的稳健性，但速度快了几个数量级（< 0.1秒，而非 ~ 10秒）。

实验设置： 我们在 $\mathbb{R}^{20}$ 中生成数据，采用三种配置，使其有效 Wasserstein 维数为 $d_{\text{int}} = 10$：(1) 超立方体混合分布（80% 为 2 维，20% 为 10 维）；(2) 低秩高斯分布混合（80% 秩为 2，20% 秩为 10）；(3) 单个秩为 10 的高斯分布。基于最优传输的估计量比较大小为 $n = 2000$ 和 $\eta n = 3000$的支撑集，而基线采用固定大小 3000。实验重复进行了 20 次。

图 1 展示了性能差异。虽然所有方法在简单流形（配置3）上都表现良好，但标准几何估计量无法捕捉混合分布的全局传输复杂性（配置1和2）。相比之下，两种基于最优传输的方法都能稳健地恢复 $d_{\text{int}} \approx 10$。尤为重要的是，我们的半离散估计量效率远高于基于离散Wasserstein距离的估计量：前者在 0.1 秒内即可计算出估计值，而后者在图形处理器上需要 $\simeq 10$ 秒。离散 W_1 基线需要求解精确最优传输问题。作为参考，文中还提供了 `scikit-dimension` 基线在中央处理器上的运行时间。

5. 耦合偏差消减：对角Richardson估计量

Sinkhorn 散度是最优传输的快速替代指标，得益于 geomloss（Feydy等人，2019年）等高效求解器，这些求解器具有线性内存占用和出色的图形处理器性能，前提是 ε 并非小到可以忽略。然而，使用 Sinkhorn 散度根据有限样本估计未经正则化的最优传输代价面临双重挑战：估计量同时受到正则化参数 ε （熵偏差）和有限样本量 n （统计偏差）的影响。为进一步减少熵偏差，（Chizat 等人，2020年）在 ε, 上引入了 Richardson外推方案，并提出估计量 $R_{\varepsilon,n} = 2S_{\varepsilon,n} - S_{\sqrt{\varepsilon},n}$。尽管理论上，这种方法能将近似误差改进至 $\mathbb{E}[|\widehat{R}_{\varepsilon,n} - W_2^2|] = o(n^{-2/(d+8)})$ （其中 $\varepsilon \asymp n^{-1/(d+8)}$）。第一步改进自然是在 ε 中用我们针对 d_{int}的估计量 d^* 替换 d。接下来，即使样本近似并非渐近限制因素，其较大的常数前因子仍可能使熵偏差校正在实践中失效，因为该方法针对的并非误差的主要来源。

为更好地处理熵偏差和统计偏差，我们提出一种耦合策略——**对角 Richardson 估计量**——将正则化方案 ε 直接与样本量 n关联。通过同时处理这两种偏差，该方法利用我们的内在维数估计量，对偏差校正进行最优校准。

总偏差分解。统计最优传输领域早已众所周知，熵正则化应随样本数量增加而衰减；即，对于某个指数 $a > 0$，有 $\varepsilon_n \asymp n^{-a}$。总期望误差可分解为：

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

我们刻画了这两项的渐近展开。熵偏差的渐近展开（10）由（Chizat 等人，2020年）中的命题4推得（假设测度具有有限的Fisher信息）；而统计偏差的渐近展开（11）则由（Stromme，2024年）中的定理2推得，该定理确立了Sinkhorn散度的低复杂度自适应：

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

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

注 5.1。统计偏差的控制常数在形式上为 $c(\varepsilon)n^{-1/2}$，取决于Sinkhorn散度的二阶导数。上界 $c(\varepsilon) \lesssim \varepsilon^{-d_{\text{int}}/2}$ 由（Stromme，2024；Groppe & Hundrieser，2024）独立地通过例 5 中关于标准误差的速率 $n^{-(1-ad_\epsilon)/2}$ （当 $\varepsilon \asymp n^{-a}$ 时）得到。该速率 完全由统计偏差决定，因为根据（Chizat 等人，2020）中的命题4，$S_{\varepsilon,n} - \mathbb{E}[S_{\varepsilon,n}] = \mathcal{O}_{\mathcal{P}}(1/\sqrt{n})$在 $\varepsilon \ge 0$ 上均匀成立。统计偏差速率的尖锐性尚未得到完全确立，不过相关讨论针对我们关注的情形提供了结果：$\varepsilon = n^{-a}$ 和 d_ϵ 替换为 d_{int}。因此，为确保严谨，我们将（11）中的等式作为假设，因为Richardson 方法要有效消去偏差项，就必须满足这一条件。

假设 5.2 （统计偏差的尖锐性）. 我们假设 Sinkhorn 散度的统计估计误差具有尖锐的渐近展开。也就是说，（11）中等式成立：

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

其中 $C_4 \neq 0$ 与 d_{int} 为数据的内在维数。

最优调度与对角估计量。为构造高效的估计量，我们寻求一种速率 $\varepsilon \asymp n^{-a}$，使熵偏差与统计偏差的收敛率达到平衡。令指数 $4a = 1 - ad_{\text{int}}$ 相等，可得最优衰减率 $a = \frac{1}{d_{\text{int}}+4}$。在此调度下，对角Richardson外推可以同时消除两种偏差的一阶项。

我们使用样本量 n 和 $2n$ 定义对角Richardson估计量：

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

其中，权重由偏差项的共同衰减率 $\gamma = \frac{2}{d_{\mathrm{int}}+4}$ 决定：

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

以下命题确立了这一耦合估计量改进后的收敛率 r.

命题 5.3。 在假设5.2下，并采用调度 $\varepsilon_n \asymp n^{-1/(d_{\mathrm{int}}+4)}$，对角*Richardson*估计量满足：

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

此外，如果高阶项满足 $o_{ent}(n^{-2a}) = \mathcal{O}(n^{-4a})$ 且 $o_{stat}(n^{-(1-ad_{\mathrm{int}})}) = \mathcal{O}(n^{-2+2ad_{\mathrm{int}}})$，则速率提高为 $\mathcal{O}(n^{-\frac{4}{d_{\mathrm{int}}+4}})$。

T因此，该速率略优于仅使用 ε 的 Richardson 估计量；后者在使用 d_{int} 时为 $(n^{-\frac{2}{d_{\mathrm{int}}+8}})$。

证明概要。 令 $R^{\mathrm{Diag}} = \lim_{n\to\infty}\mathbb{E}[\widehat{R}_{2n}^{\mathrm{Diag}}]$。误差分解为：

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

随机误差的量级为 $\mathcal{O}(n^{-1/2})$；对于 $d_{\mathrm{int}} > 2$，与偏差项相比可以忽略不计。权重的选择抵消了偏差中的主导阶项 $n^{-\frac{2}{d_{\mathrm{int}}+4}}$，仅留下高阶余项。需要注意的是，如果不满足假设5.2，我们仍然可以获得渐近地（关于 d_{int}）的极小极大速率 $\mathcal{O}(n^{-2/(d_{\mathrm{int}}+4)})$。

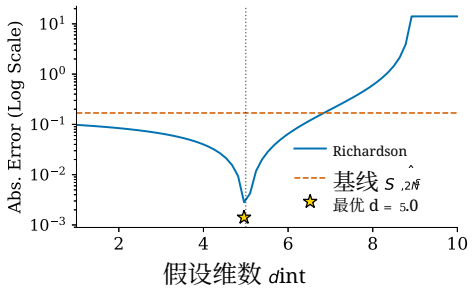


图 *2.* **对内在维数的敏感性**。 我们对 $\mathbb{R}^{10}$ 中的源合成混合分布（90%的质量位于5维高斯分布上，10%位于1维高斯分布上）执行对角 Richardson 外推 $2n = 2000$，并外推至10维高斯分布。通过改变外推权重中所使用的维度参数 d，我们观察到估计误差恰好在主导内在维度 $d = 5$处最小。这证实了将调度正确校准至有效几何结构对于实现最优偏差校正至关重要。

袋装对角 Richardson。 虽然对角Richardson估计量 $\widehat{R}_{2n}^{\mathrm{Diag}}$ 成功降低了偏差，但这一改进是以方差增大为代价的。方差增大源于外推过程中对系数 w_{2n} 和 w_n进行加权差分 ^。因此，子采样项 $S_{\varepsilon_n,n}$ 中的噪声会在最终估计中被放大。此外，仅依赖一个大小为 n 的随机子集在统计上效率低下，因为这会丢弃未使用样本中对低分辨率分量有用的信息。为缓解方差增大，我们采用一种袋装策略，对 K 个独立子样本上的低分辨率估计量取平均，同时将完整数据集用于高分辨率项。袋装 *Richardson* 估计量 定义如下：

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

这种方法通过降低相减项的方差来稳定估计量。我们以相对于标准 Sinkhorn 估计量的方差膨胀程度来量化这种稳定性。附录中的证明基于 (Goldfeld等人，2024年) 中确立的 Sinkhorn 散度的二阶Hadamard可微性。

命题5.4 (方差稳定性)。 袋装 *Richardson* 估计量的方差与 ^标准 *Sinkhorn* 估计量 $S_{\varepsilon_{2n},\ 2n}$ 的方差关系如下：

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

关键在于，随着袋数 $K \to \infty$ 增加，方差开销会消失。因此，袋装估计量既能达到 Richardson 方案更优的偏差率，又能收敛至由完整样本量决定的最优方差下限 $2n$。

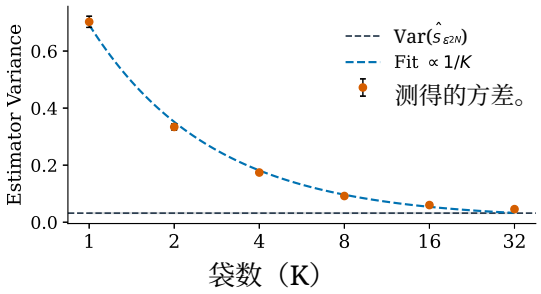


图*3*。 **通过装袋降低方差**。 使用与图2相同的混合设置，我们测量袋装 Richardson 估计量的方差随装袋数量 K 的变化。方差是在 20 次运行中计算，并重复 10 次。我们观察到方差开销明显 $1/K$ 下降，并渐近地收敛到标准 Sinkhorn估计量的方差下限$\widehat{S}_{\varepsilon_{2n},\ 2n}$ （虚线）。这证实了装袋法消除了外推带来的稳定性成本。

5.1. 从最优传输几何到 Wasserstein 距离估计

我们通过在图像流形（MNIST、FashionMNIST、CIFAR10）上估计 W_2^2，验证了完整的“几何感知”流程。为克服经验测度缺乏真实值的问题，我们使用严格凸势函数 $\psi_\theta(x) = \frac{1}{2}\|x\|^2 +$ 输入凸神经网络(x) 构造合成目标测度 $\nu = (T^*)_{\#}\mu$。根据 Brenier 定理， $T^* = \nabla\psi_\theta$ 是唯一的最优映射，可提供精确的参考距离 $W_2^2(\mu,\ \nu) = \mathbb{E}[\|x - T^*(x)\|^2]$。

^**实验设置**。对于每个数据集，给定 $n = 1000$，我们首先估计内在维数 d^* （第4节），并将其代入我们的袋装对角 Richardson 估计器（$K = 12$）以校准偏差消除权重。我们将该方法与标准 Sinkhorn 散度（“基线”）以及标准的 ε-Richardson 外推（“Eps-Rich”）（Chizat 等人，2020年）进行比较。

表*1*。**Wasserstein 估计基准测试** W_2^2。

数据集	方法	误差 ± 标准差	平均耗时（秒）
CIFAR10	Base	1.210 ± 0.02	0.94
	n-Diag-Rich	**0.089** ± 0.03	2.85
	$\frac{n}{4}$-Diag-Rich	0.201 ± 0.03	0.89
	Eps-Rich	1.007 ± 0.02	1.56
时尚 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
	$\frac{n}{4}$-对角丰富型	0.420 ± 0.04	0.95
	Eps-Rich	0.706 ± 0.01	1.40

结果。 表 1 证实，纳入内在几何对于准确估计至关重要。标准 ε-外推无法解决高维中占主导地位的统计偏差，即使使用

$\widehat{\ }$ $\asymp d^*$。相比之下，我们通过 d^* 校准的对角估计量将误差降低了一个数量级（例如，在 CIFAR10 上为 $1.21 \to 0.09$ ）。这验证了将正则化方案与内在维数（$\varepsilon \asymp n^{-1/(\widehat{d}^*+4)}$）相结合能够有效缓解离散化瓶颈。此外，即使在预算较少的情况下（样本量减少 $\frac{n}{4}$ - 装袋法 $K = 12$），我们的方法仍显著优于基线，同时计算成本与其相当。

6. 结论与未来工作

本文证明，明确估计数据的内在维数是突破离散最优传输统计局限的关键。我们表明，通过无需求解器的内在维数估计进行校准后，我们的对角Richardson估计量显著提高了高维流形上Wasserstein距离估计的准确性。该框架为未来研究开辟了多个方向。首先，我们的偏差抵消的有效性依赖于统计误差展开的精确性（假设 5.2），这一性质值得进一步进行理论分析。其次，尽管本文聚焦于标量传输代价，将这种“几何感知”的偏差校正策略及其分析扩展至Kantorovich势函数和梯度，仍是提升生成建模训练稳定性的一个有前景的方向。

参考文献

An, D., Guo, Y., Lei, N., Luo, Z., Yau, S.-T., and Gu, X. Ae-ot：一种基于扩展半离散最优传输的新型生成模型。*ICLR2020*, 2019. Bac, J., Mirkes, E. M., Gorban, A. N., Tyukin, I., and Zinovyev, A. Scikit-dimension：一个用于内在维数估计的 Python 软件包。《*Entropy*》, 23(10):1368, 2021. Blanchet, J., Jambulapati, A., Kent, C., and Sidford, A. 朝着最优传输的最优运行时间迈进。《*Operations Research Letters*》, 2023. Block, A., Jia, Z., Polyanskiy, Y., and Rakhlin, A. 使用 Wasserstein 距离进行内在维数估计。《机器学习研究杂志》, 23(313):1–37, 2022. Chen, Y., Telgarsky, M., Zhang, C., Bailey, B., Hsu, D., and Peng, J. 通过显式 Wasserstein 最小化，以渐进的半离散方法训练生成网络。载于国际机器学习会议，第1071–1080页。P MLR, 2019. Chewi, S., Niles-Weed, J., and Rigollet, P. 统计最优传输。*arXiv*预印本 *arXiv:2407.18163*, 3, 2024. Chizat, L., Roussillon, P., L´eger, F., Vialard, F.-X., and Peyr´e, G. 利用 Sinkhorn 散度加快 Wasserstein 距离估计。神经信息处理系统进展, 33:2257–2269, 2020. Courty, N., Flamary, R., and Tuia, D. 基于正则化最优传输的领域自适应。载于《机器学习与数据库知识发现：欧洲会议 *ECML PKDD 2014*，法国南锡，*2014*年*9*月*15*日至*19*日。论文集，第一部分 *14*》，第274–289页。Springer, 2014. Cuturi, M. Sinkhorn 距离：光速计算最优传输。神经信息处理系统进展, 26, 2013. Dudley, R. M. Glivenko-Cantelli 收敛的均值速度。《数学统计年刊》, 40(1):40–50, 1969. Dvurechensky, P., Gasnikov, A., and Kroshnin, A. 计算最优传输：加速梯度下降算法的复杂度优于 Sinkhorn 算法。载于国际机器学习会议，第1367–1376页，2018. Feydy, J., S´ ejourn´e, T., Vialard, F.-X., Amari, S.-i., Trouv´e, A., and Peyr´e, G. 使用 Sinkhorn 散度在最优传输与最大均值差异之间插值。载于第*22*届国际人工智能与统计会议，第2681–2690页。PMLR, 2019. Fournier, N. and Guillin, A. 关于经验测度在 Wasserstein 距离下的收敛率。《概率论及相关领域》, 162(3):707–738, 2015. Genevay, A., Peyr´e, G., and Cuturi, M. 使用 Sinkhorn 散度学习生成模型。载于国际人工智能与统计会议，第1608–1617页。PMLR, 2018.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyr´e, G. Sinkhorn 散度的样本复杂度。载于第*22*届国际人工智能与统计会议，第1574–1583页。PMLR, 2019. Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. 熵正则最优传输映射与 Sinkhorn 散度的极限定理。《电子统计学杂志》, 18(1):980–1041, 2024. Groppe, M. and Hundrieser, S. 经验熵正则最优传输的低复杂度自适应。《机器学习研究杂志》, 25(344):1–55, 2024. Hundrieser, S., Staudt, T., and Munk, A. 不同测度之间的经验最优传输可适应较低复杂度。载于《亨利·庞加莱研究所年刊（*B*辑）：概率与统计》，第60卷，第824–846页。Institut Henri Poincar´e, 2024. H¨utter, J.-C. and Rigollet, P. 光滑最优传输映射的极小极大估计。《统计年刊》, 49(2), 2021. Kitagawa, J., M´erigot, Q., and Thibert, B. 半离散最优传输的牛顿算法。《欧洲数学学会杂志》, 21, 03 2016. doi: 10.4171/JEMS/889. Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. 从词嵌入到文档距离。载于国际机器学习会议，第957–966页。PMLR, 2015. Li, Z., Li, S., Wang, Z., Lei, N., Luo, Z., and Gu, D. X. Dpm-ot：一种基于最优传输的新型扩散概率模型。载于*IEEE/CVF*国际计算机视觉会议论文集，第22624–22633页，2023. Lin, T., Ho, N., and Jordan, M. 关于高效最优传输：贪心算法与加速镜像下降算法分析。载于国际机器学习会议，第3982–3991页。PMLR, 2019. Maurer, A. and Pontil, M. 经验伯恩斯坦界与样本方差惩罚。载于第*22*届学习理论年会（*COLT*），2009. Mena, G. and Niles-Weed, J. 熵正则最优传输的统计界：样本复杂度与中心极限定理。神经信息处理系统进展, 32, 2019. Peyr´e, G., Cuturi, M., et al. 计算最优传输：数据科学中的应用。《机器学习基础与趋势》, 11(5-6):355–607, 2019. Pooladian, A.-A., Divol, V., and Niles-Weed, J. 不连续最优传输映射的极小极大估计：半离散情形。载于国际机器学习会议，第28128–28150页。PMLR, 2023. Ramdas, A., Garc´ıa Trillos, N., and Cuturi, M. 关于 Wasserstein 双样本检验及相关非参数检验。《*Entropy*》, 19(2):47, 2017. Stromme, A. J. 熵正则最优传输的最小内在维度标度。载于概率与统计软方法国际会议，第491–499页。Springer, 2024. Villani, C. 最优传输：旧与新，第338卷。Springer, 2009. Weed, J. and Bach, F. 经验测度在 Wasserstein 距离下收敛的精确渐近收敛率与有限样本收敛率。《伯努利》, 25, 06 2017. doi: 10.3150/18-BEJ1065.

A. 内在维数与Wasserstein距离

我们回顾一些与内在维数以及 Wasserstein 维数和 Sinkhorn 散度的多分辨率行为相关的概念和性质，这些内容见于（Weed & Bach，2017；Stromme，2024年）。

A.1. 概念。

定义（δ-覆盖数）。给定集合 $S \subseteq X$， S 的 δ-覆盖数记为 $\mathcal{N}_\delta(S)$，它是满足以下条件的最小值m：存在 m 个直径为 δ 的闭球 B_1，……， B_m，使得 $S \subseteq \bigcup_{1\le i\le m} B_i$。 S 的 δ-维数定义为

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

以下定义在处理测度而非集合时尤其有用，因为它允许忽略一小部分质量。该定义最早见于（Dudley，1969年）。

定义（(δ τ)-测度覆盖数）。给定定义在 X 上的测度 μ，(δ,τ)-覆盖数为

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

(δ,τ)-维数为

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

极限定义为

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

定义（Wasserstein 维数）。上、下 Wasserstein 维数分别为

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

A.2. 性质

命题 A.1 (命题 5 和推论 1（Weed & Bach，2017））。设 $p \in [1,\infty)$。假设存在 $\delta' \le 1$ 和 $s > 2p$使得 $d_\delta(\mu, \delta^{\frac{sp}{s-2p}}) \le s$ 对所有 $\varepsilon \le \delta'$ 都成立。那么存在常数 C_1, C_2 使得：

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

特别地， $\mathbb{E}[W_p(\mu,\widehat{\mu}_n)] \lesssim n^{-1/s}$。作为推论，若 $s > d_p^*(\mu)$，则有 $\mathbb{E}[W_p(\mu,\widehat{\mu}_n)] \lesssim n^{-1/s}$。

现在我们推导一个略 有不同的命题，该命题基于（Weed & Bach，2017）中的命题 6，证明类似。

命题 A.2 (Wasserstein 收敛速率的局部下界)。设 μ 是 $\mathbb{R}^D$ 上的一个概率测度，并设 $p \in [1,\infty)$。假设存在常数 $\tau \in (0、1)$、 $A > 0$、 $s > 0$，以及一个尺度区间 $[a,b] \subset (0,1]$，使得(δ,τ)-装填数满足

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

如果样本量 n 满足有效性条件 $Ab^{-s} < n < Aa^{-s}$，那么对于任意支撑集至多包含 n 个点的离散测度 ν（包括经验测度 $\widehat{\mu}_n$），我们有如下下界：

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

证明。设 n 满足有效性条件。我们定义临界半径 $\varepsilon_n := (n/A)^{-1/s}$。由条件 $Ab^{-s} < n < Aa^{-s}$，可明确得出 $a < \varepsilon_n < b$。

任取一个满足 $\varepsilon \in [a, b]$ 的 $\varepsilon < \varepsilon_n$。根据装填假设，可装入 μ 的支撑集（捕获质量 $1-\tau$）中的半径为 ε 的互不相交球的数量满足 $N := \mathcal{N}'_\varepsilon(\mu, \tau) \geq A\varepsilon^{-s}$。由于 $\varepsilon < \varepsilon_n$，我们严格有

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

令 $\{B_1, \ldots, B_N\}$ 表示这些互不相交球的装填。由于 ν 至多支撑在 n 个点上，根据鸽巢原理，这些球中至少有 $N - n$ 个不包含 ν 的任何支撑点。令 $\mathcal{I}_{\text{empty}}$ 表示这些空球的索引。

Wasserstein距离 $W_p^p(\mu, \nu)$ 涉及将这些空球中的质量传输到 ν 的支撑集。对于任意球B_i，若满足 $i \in \mathcal{I}_{\text{empty}}$，则质量 $\mu(B_i)$ 至少必须移动 ε 的距离，才能到达 $\text{supp}(\nu)$ 中的任意点（该点位于B_i之外）。遵循 Weed & Bach (2017)（命题 6）中的精确推导，汇总传输代价的下界为：

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

由于此　不等式对任意 $\varepsilon < \varepsilon_n$都成立，我们取极限 $\varepsilon \to \varepsilon_n$。代入 $\varepsilon_n = (n/A)^{-1/s}$ 可得　s:

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

□

A.3. 最小内在维度标度

这里回顾 Sinkhorn 散度的最小内在维度缩放，我们将利用它进行 Wasserstein 距离估计。

定理（Sinkhorn 散度的 MID 缩放）。对于与所有问题参数无关的数值常数，

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

该估计中的维数量由尺度为 ε的最小覆盖数所控制，体现了 MID缩放现象。此处，常数 L 指代代价函数的利普希茨界；对于 $\|\cdot\|^p$代价，在有界情形下，该界可确保为有限值。

与内在维数的联系。 在本文中，我们通过 $\mathbb{E}[W_1(\mu, \widehat{\mu}_n)]$ 的收敛率及其对应的下界，刻画固定尺度下的内在维数。现在，我们将这一刻画与前文给出的 Sinkhorn 散度的 MID缩放下界联系起来。

考虑代价函数 $c = \|\cdot\|$，其利普希茨常数为 $L = 1$。假设在尺度区间 $\varepsilon \in [\varepsilon_a, \varepsilon_b]$ 上，测度具有足够的正则性，使得维数在不同质量阈值下保持稳定；具体而言，$d_\varepsilon(\mu, \tau) = d_\varepsilon(\mu, 0)$对所有 $\tau \in (0, 1/2]$均成立。在这些条件下，命题 A.1 和 A.2 为 Wasserstein距离的收敛率提供了相匹配的上下界，其中 $d_{\text{int}} = d_\varepsilon$ 适用于一段范围 $n \in [N_{\min}, N_{\max}]$。因此，我们可以将内在维数 d_{int} 直接等同于尺度依赖覆盖维数 d_ε，而这正是定理2中 MID缩放所支配的量。

B. 定理 4.2 的证明：维数估计控制

证明。 固定 $n \geq n_{\min}$ 与 $\eta > 1$。对每个 $k \in \{n, \eta n\}$，我们引入记号

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

其中 $x_1, \ldots, x_k \overset{\text{i.i.}}{\sim} \rho^{\text{d}}$定义了 $\widehat{\rho}_k^*$ 的支撑集，而 $X_1, \ldots, X_N \overset{\text{i.i.}}{\sim} \rho^{\text{d}}$ 是独立的蒙特卡洛样本（与 x_j 无关）。令 $c(\cdot) = \|\cdot\|$，并注意到 $\text{Diam}(\rho) := \sup_{x,y \in \text{Supp}(\rho)} c(x, y)$。

蒙特卡洛集中 W_k。 在给定 $(x_1, \dots, x_k)$ 的条件下，随机变量 $\min_{j\le k} c(X_i, x_j)$ 独立同分布，且取值于 $[0, \mathrm{Diam}(\rho)]$。因此，根据霍夫丁不等式，对于任意 $t > 0$,

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

定义蒙特卡洛精度水平

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

于是，对每个 k， $\mathbb{P}(|W_k - W_k^*| > \kappa_N) \le \delta/4$。对 $k \in \{n, \eta n\}$ 使用联合界，事件

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

满足

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

对 W_k^* 在尺度 n 和 ηn 上的双侧控制。 根据假设 4.1 给出的界，对于所有 $k \in [n_{\min}, n_{\max}/\eta]$,

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

并且对于任意 $t > 0$，利用半对偶最优解的有界性 $\mathbf{0}_c \le \mathrm{Diam}(\rho)$,

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

其中，$\mathrm{Diam}(\rho)$ 是 $\mathrm{Supp}(\rho)$ 的直径。取 $t = \xi C_2 k^{-1/d_{\mathrm{int}}}$，令 $\xi > 0$，可得

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

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

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

满足 $\mathbb{P}(\mathcal{E}_{\mathrm{Q}}) \ge 1 - \delta/2$。在 $\mathcal{E}_{\mathrm{Q}}$ 并利用（19），

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

设 $\mathcal{E} := \mathcal{E}_{\mathrm{MC}} \cap \mathcal{E}_{\mathrm{Q}}$。则 $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$。

总体对数比率 D的控制。定义

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

由（20），

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

取对数并记作 $D_0 := \log \eta / d_{\mathrm{int}}$，可得

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

因此 $|D - D_0| \le b$。此外，关于 η 的所述条件蕴含 $b \le \frac{\gamma}{2} D_0$。 –

稳定性：log W_k 在蒙特卡洛误差下。在 $\mathcal{E}$ 上，我们有 $|W_k - W_k^*| \le \kappa_N$。我们首先通过要求来确保 $W_{\eta n}$（以及类似地 W_n）的正性

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

根据 $W_{\eta n}^* \ge C_1(\eta n)^{-1/d_{\mathrm{int}}}$ 由 (19)，只需满足

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

而定理中关于 N 的第二个下界蕴含了这一点。

假设该条件成立。那么对于 $k \in \{n, \eta n\}$,

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

由于 $|\widehat{W}_k - W_k^*| \le \kappa_N \le \frac{1}{2} W_k^*$，我们有 $\left| \frac{\widehat{W}_k - W_k^*}{W_k^*} \right| \le \frac{1}{2}$。因此 $|\log(1+u)| \le 2|u|$ 对于 $|u| \le 1/2$，得到

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

因此，对于 $D := \log W_n - \log W_{\eta n}$,

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

利用 $W_n^* \ge C_1 n^{-1/d_{\mathrm{int}}}$ 和 $W_{\eta n}^* \ge C_1 (\eta n)^{-1/d_{\mathrm{int}}}$,

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

代入 $\kappa_N = \mathrm{Diam}(\rho) \sqrt{\log(8/\delta)/(2N)}$，得到

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

定理中关于 N 的第一个下界保证右端不超过 $\frac{\gamma}{2} D_0$。

分母区间及向 d^* 的传播。在 $\mathcal{E}$ 我们有

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

因此 $(1-\gamma)D_0 \le D \le (1+\gamma)D_0$。最后，我们利用关系 $d_N^* = \log \eta / D$ 得到

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

由于 $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$，证明至此完成。

备注： 注意，为使证明成立，我们要求条件 $\eta n \le n_{\max}$ 成立。 □

B.1. 流形假设下的内在维度估计

我们在此简要讨论，在流形假设下，我们的证明如何成立，正如（Block 等人，2022年）所述。我们在此回顾其假设。

假设 B.1 （几何正则性）。设 $\mathcal{M} \subset \mathbb{R}^d$ 是一个维数为 $d_{\mathcal{M}}$、直径为 Δ =diam($\mathcal{M}$) 且 reach 为 $\tau > 0$ 的紧流形。此外，假设数据生成分布 ρ 关于 $\mathcal{M}$ 上的均匀测度存在一个密度，且该密度有界为 $0 < w_{\min} \leq \mathrm{d}\rho\ \mathrm{dvol}_{\mathcal{M}} \leq w_{\max} < \infty$。

正 reach τ保证 $\mathcal{M}$ 没有自交且曲率受控；直观地说，在尺度τ以下，投影到 $\mathcal{M}$ 上的最近点（法向）投影是单值的，因此 $\mathcal{M}$ 在局部表现得像 $\mathbb{R}^{d_{\mathcal{M}}}$。形式上， $\mathcal{M}$的 reach 为

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

这一概念在几何推断中是标准的；例如，若 $\tau > 0$ 成立，且 $\mathcal{M} \subset \mathbb{R}^d$ 是一个紧致嵌入 $C^{1,\ 1}$ 子流形。关于 reach 概念及其在基于最优传输的内在维数估计中的作用，详见（Block 等人，2022年）。这里我们指出， $\omega_{d_{\mathcal{M}}}$ 是单位 d维球的体积。

命题B.2(((Block 等人，2022年))流形假设下 Wasserstein 误差的上界和下界)。对于

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

对于任意由 n 个点构成的测度 σ_n，我们有

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

此外，对于常数 $C \leq 48$，我们有

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

利用命题B.2，因此我们可以应用定理4.2，其中常数为：$d_{\mathrm{int}} = d_{\mathcal{M}}$, $n_{\min} = {}^{d_{\mathrm{int}}}$体积 $_{4w\omega_{\min}}{}^{\mathcal{M}}\left(\frac{\tau}{8}\right)^{-d_{\mathrm{int}}}$, $n_{\max} = +\infty$, $C_1 = \frac{1}{32}\left({}^{d_{\mathrm{int}}}$体积 $_{4w\omega_{\min}}{}^{\mathcal{M}}\right)^{\frac{1}{d_{\mathrm{int}}}}$, 以及 $C_2 = C\left(\text{体积}_{\omega_{\min}}{}^{\mathcal{M}}\right)^{\frac{1}{d}}$ 积分 $\sqrt{\text{对数}\left(\frac{n w_{\min}\mathrm{Diam}(}{\rho\ {}^{d_{\mathrm{int}}})\ \text{体积}_{\mathcal{M}}}\right)}$。

C. 命题 5.3的证明：对角Richardson 的收敛性

证明。 我们分别分析偏差和随机偏差。

设正则化参数按 $\varepsilon_n \asymp n^{-a}$ 缩放。基于标准展开，并假设统计界等号成立，我们有：

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

$$\mathrm{Bias}_{\mathrm{stat}}(n) = C_{\mathrm{stat}} n^{-(1-ad_{\mathrm{int}})/2} + o(n^{-(1-ad_{\mathrm{int}})/2}). \tag{23}$$

为了达到最优收敛速率，我们通过令各项的指数相等来平衡主导阶项：

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

由此得到最优衰减选择 $a = \frac{1}{d_{\mathrm{int}}+4}$。在此选择下，两个偏差项都以相同的速率 γ 衰减：

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

因此，总期望具有如下展开：

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

对角 Richardson on估计量定义为 $\widehat{R}_{2n}^{\text{Diag}} := w_{2n}\widehat{S}_{\varepsilon_{2n},\,2n} + w_n\widehat{S}_{\varepsilon_n,\,n}$。由线性性，其期望值为：

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

分别将 (24中的展开式代入熵项和统计项：

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

权重选为 $w_{2n} = \frac{2^\gamma}{2^\gamma-1}$ 和 $w_n = \frac{-1}{2^\gamma-1}$。我们验证消去条件：

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

由于两个偏差都按 $n^{-\gamma}$缩放，相同的 权重可以同时消除 C_{ent} 和 C_{stat} 的首项。因此：

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

利用（Chizat 等人，2020年）中的命题4，我们有 $\mathbb{P}\left[|\widehat{S}_{\varepsilon,n} - \mathbb{E}[\widehat{S}_{\varepsilon,n}]| \geq t\right] \leq 2\exp(-nt^2/D^24)$，因此，注意到 $Z_n = S_{\varepsilon_n,n} - \mathbb{E}[S_{\varepsilon_n,n}]$，我们有

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

结合剩余偏差界与随机偏差界，我们得出

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

□

备注：对速率失配的稳健性与小样本效率。 尽管上述证明假设完美平衡（$\gamma_{\text{ent}} = \gamma_{\text{stat}} = \gamma$），对角Richardson估计量即使在统计偏差以稍快的速率 $\beta > \gamma$ 衰减时（使其相对于熵偏差渐近地可忽略）仍然高度有效。这在小样本情形（n 较小）中尤为相关，此时"渐近地更优"的项仍可能具有较大的常数。

设统计偏差的形式为$\text{Bias}_{\text{stat}}(n) = C_{\text{stat}}n^{-\beta}$。外推后的统计一阶偏差变为：

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

代入权重 $w_{2n} = \frac{2^\gamma}{2^\gamma-1}$ 和 $w_n = \frac{-1}{2^\gamma-1}$，我们定义偏差缩减因子 $\rho(\beta, \gamma)$:

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

我们观察到两个关键性质：

1. **精确抵消：** 如果 $\beta = \gamma$，那么 $\rho(\gamma, \gamma) = 0$，从而恢复主定理。

2. 局部连续性（小 n 情形）： 在最优传输中，熵正则速率 $\gamma = 2a$ 和统计速率 $\beta = (1 - ad_{\text{int}})/2$ 在代数上相互耦合。对于接近最优值 $a = \frac{1}{d_{\text{int}}+4}$ 的任意 a，差值 $\delta = \beta - \gamma$ 都很小。在 $\delta \approx 0$ 附近进行泰勒展开可得：

$$\rho(\beta, \gamma) \approx \frac{-\delta \ln 2}{2^{\gamma} - 1}.$$

因此，即使各项速率并不相同，对于计算上可行的 n，估计量仍能按与速率不匹配程度成正比的因子 δ 抑制统计偏差（此时 $n^{-\beta}$ 尚未可忽略）。这进一步表明，该方法对我们对 d_{int} 的估计也具有稳健性。

同样，我们观察到 Richardson 对 C_{stat} 的小幅振荡具有稳健性。

备注：关于不同偏差的猜想 Goldfeld等人Goldfeld等人（2024年）的最新研究为 Sinkhorn 散度建立了中心极限定理，并指出预期的有限样本误差按 $\mathbb{E}[S_{\varepsilon,n}] - S_{\varepsilon} = \mathcal{O}(1/\sqrt{n} + n^{-1}r_{\varepsilon})$ 缩放（见其备注 7）。因子 r_{ε} 取决于 Sinkhorn 势函数的方差。若我们猜想缩放为 $r_{\varepsilon} \asymp \varepsilon^{-(d_{\text{int}}+2)}$（源自 Sinkhorn 二阶导数），则统计偏差为 $\mathcal{O}(n^{-1}\varepsilon^{-(d_{\text{int}}+2)})$。将其与熵偏差 $O(\varepsilon^2)$ 平衡，可得最优衰减选择 $\varepsilon \asymp n^{-1/(d_{\text{int}}+4)}$。值得注意的是，这得到了与标准 $\sqrt{n}$ 速率假设下完全相同的指数 $a = \frac{1}{d_{\text{int}}+4}$，表明我们的对角 Richardson 策略在结构上对不同统计机制具有稳健性。

命题的证明5.4：装袋法的方差缩减

证明。 设$\mathbf{D}_{2N} = \{(X_i, Y_i)\}_{i=1}^{2N}$ 是一个数据集，包含从 (μ, ν) 的联合分布中抽取的 $2N$ 个独立同分布样本。我们生成K 个袋，每个袋对应一个**子样本索引集** $\mathcal{S}_k$，其大小为 $|\mathcal{S}_k| = N$，通过有放回的均匀随机抽样得到。

我们利用全方差公式分析渐近方差，并以观测到的数据集 $\mathbf{D}_{2N}$ 为条件。总方差可分解为：

$$\text{Var}(\widehat{R}_{K,N}) = \underbrace{\text{Var}\left(\mathbb{E}[\widehat{R}_{K,N} \mid \mathbf{D}_{2N}]\right)}_{\text{Term I: Dataset Variance}} + \underbrace{\mathbb{E}\left[\text{Var}(\widehat{R}_{K,N} \mid \mathbf{D}_{2N})\right]}_{\text{Term II: Bagging Variance}}. \tag{27}$$

渐近线性展开。

我们首先利用经验 Sinkhorn 散度是二阶 Hadamard 可微的这一性质（Goldfeld等人，2024年）。因此，我们可以使用二阶泰勒展开。一阶项 L_{ε} 是线性的，可改写为

$$L_{\varepsilon} = \frac{1}{2N} \sum_{i=1}^{2N} \phi_{\varepsilon}(Z_i)$$

其中 ϕ_{ε} 是所谓的影响函数。一阶 von Mises 展开给出：

$$\widehat{S}_{\varepsilon,2N} = S_{\varepsilon} + \frac{1}{2N} \sum_{i=1}^{2N} \phi_{\varepsilon}(Z_i) + r_{2N}(\varepsilon), \tag{28}$$

其中，E 是总体值，$Z_i = (X_i, Y_i)$，而 ϕ 是典范影响函数，且 $\mathbb{E}[\phi(Z)] = 0$ 以及 $\text{Var}(\phi_{\varepsilon}(Z)) = \sigma_{\varepsilon}^2$。

类似地，对于任意大小为 N的子样本 $\mathcal{S}_k$，估计量可展开为：

$$\widehat{S}_{\varepsilon,N}^{(k)} = S_{\varepsilon} + \frac{1}{N} \sum_{j \in \mathcal{S}_k} \phi_{\varepsilon}(Z_j) + r_N^{(k)}(\varepsilon). \tag{29}$$

项 I 的分析（数据集方差）。 我们通过代入冯·米塞斯展开来分析条件期望 $\mathbb{E}[R_{K,N} \mid \mathbf{D}_{2N}]$。首先，回顾子样本均值项，期望的线性性意味着：

$$\mathbb{E}\left[\frac{1}{M} \sum_{j \in \mathcal{S}_k} \phi_{\varepsilon}(Z_j) \,\middle|\, \mathbf{D}_{2N}\right] = \sum_{i=1}^{N} \phi_{\varepsilon}(Z_i) \cdot \mathbb{P}(i \in \mathcal{S}_k) \cdot \frac{1}{M} = \frac{1}{N} \sum_{i=1}^{N} \phi_{\varepsilon}(Z_i). \tag{30}$$

现在，将显式展开$\widehat{S}_{\varepsilon,N} = S_\varepsilon + \bar{\phi}_{\varepsilon N} + r_N(\varepsilon)$（其中 $\bar{\phi}_{\varepsilon N} = \frac{1}{N}\sum \phi_\varepsilon(Z_i)$）和 $\widehat{S}_\varepsilon^{(k)}{}_{,N} = S_\varepsilon + \bar{\phi}_{\varepsilon M,k} + r_N^{(k)}(\varepsilon)$代入估计量的定义：

$$\mathbb{E}[\widehat{R}_{K,N} \mid \mathbf{D}_{2N}] = (1+\lambda)\widehat{S}_{\varepsilon,N} - \lambda\mathbb{E}[\widehat{S}_{\varepsilon,N}^{(1)} \mid \mathbf{D}_{2N}] \tag{31}$$

$$= (1+\lambda)(E + \bar{\phi}_{\varepsilon N} + r_N) - \lambda\left(E + \mathbb{E}[\bar{\phi}_{\varepsilon M,1} \mid \mathbf{D}_{2N}] + \mathbb{E}[r_{M,1} \mid \mathbf{D}_{2N}]\right) \tag{32}$$

$$= (1+\lambda)(E + \bar{\phi}_{\varepsilon N} + r_N) - \lambda\left(E + \bar{\phi}_{\varepsilon N} + \mathbb{E}[r_{M,1} \mid \mathbf{D}_{2N}]\right). \tag{33}$$

按阶将各项分组：

$$\mathbb{E}[\widehat{R}_{K,2N} \mid \mathbf{D}_{2N}] = \underbrace{[(1+\lambda)-\lambda]\,E}_{\text{Bias}} + \underbrace{[(1+\lambda)-\lambda]\,\bar{\phi}_{\varepsilon 2N}}_{\text{Linear Term}} + \underbrace{((1+\lambda)r_{2N} - \lambda\mathbb{E}[r_{N,1} \mid \mathbf{D}_{2N}])}_{\text{Pooled Remainder } \rho_{2N}} \tag{34}$$

$$= E + \frac{1}{2N}\sum_{i=1}^{2N}\phi_\varepsilon(Z_i) + \rho_{2N}. \tag{35}$$

我们求该量的方差。令 $V_N = \frac{1}{N}\sum \phi_\varepsilon(Z_i)$。

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

线性方差：由于 $\phi_\varepsilon(Z_i)$是独立同分布的，$\mathrm{Var}(V_{2N}) = \frac{\sigma_\varepsilon^2}{N}$。

余项方差：项 r_N 是von Mises展开的二阶误差，满足 $r_N = O$（$\varepsilon^{-cd_{\text{int}}}N^{-1}$）由于二阶Hadamard可微性（Goldfeld等人，2024年；Stromme，2024年）。根据紧性假设，矩收敛，从而$\mathrm{Var}(r_N) = O(\varepsilon^{-2cd_{\text{int}}}N^{-1}N^{-2})$。类似地，$\mathrm{Var}(r_{M,1}) = O(\varepsilon^{-2cd_{\text{int}}}N^{-2})$。因此，$\mathrm{Var}(r_N) = O(\varepsilon^{2cd_{\text{int}}}N^{-2})$。

3. **协方差：** 根据柯西-施瓦茨不等式，$\mathrm{Cov}(V_N, \rho_N) \le \sqrt{\mathrm{Var}(V_N)\mathrm{Var}(\rho_N)} = O(N^{-1.5}\varepsilon^{-cd_{\text{int}}})$.

将两个方差项相加，我们得到

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

项 II（Bagging 方差）的分析。 给定 $\mathbf{D}_{2N}$ 后，全样本估计量 $S_{\varepsilon,N}$ 是常数。方差完全来自 K 子采样估计量。由于 K 个袋是独立抽取的（给定 $\mathbf{D}_{2N}$）：

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

子样本上的估计量 $\widehat{S}_\varepsilon^{(1)}{}_{,N}$ 渐近地表现为从有限总体 $\mathbf{D}_{2N}$中不放回抽取的影响函数的样本均值。令 $S^2_{\phi_\varepsilon,2N}$为影响函数在 $\mathbf{D}_{2N}$上的精确样本方差。根据有限总体修正公式，对于有放回简单随机抽样下样本均值的方差，线性项的条件方差为：

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

我们现在对数据集$\mathbf{D}_{2N}$取期望。由于样本 Z_i 是独立同分布的，$S^2_{\phi_\varepsilon,\ 2N}$ 是影响函数总体方差的无偏估计量。因此，$\mathbb{E}[S^2_{\phi_\varepsilon,2N}] = \mathrm{Var}(\phi_\varepsilon(Z)) = \sigma_\varepsilon^2$ 精确成立。因此，

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

结论。利用这一点，由代价函数在有界支撑测度上的有界性，我们有 $\sigma_\varepsilon \lesssim 1$，与 ε无关。最后，将（37）和（40）代回（27），一旦 $\varepsilon^{cd_{\text{int}}} = o(N)$，我们便有：

$$\text{Var}(\widehat{R}_{K,N}) = \left[\frac{\sigma_\varepsilon^2}{N}\right] + \left[\frac{\lambda^2}{K}\frac{\sigma_\varepsilon^2}{N}\right] + O(N^{-1.5}\varepsilon^{-cd_{\text{int}}}) + O(\varepsilon^{-2cd_{\text{int}}}N^{-2}) + o\left(\frac{S^2_{\phi_\varepsilon,2N}}{N}\right) \tag{41}$$

$$= \frac{\sigma_\varepsilon^2}{N}\left(1 + \frac{\lambda^2}{K}\right) + o(1)\,. \tag{42}$$

至此证明完毕。 □