arXiv:2602.03566v1 [cs.LG] 3 Feb 2026

黎曼神经最优传输

Alessandro Micheli [1] **Yueqi Cao** [2] **Anthea Monod** [1] **Samir Bhatt** [1 3]

摘要

计算最优传输（OT）为生成建模提供了一个有理论依据的框架。神经最优传输方法利用神经网络以摊销方式从数据中学习最优传输映射（或势函数），训练完成后可在样本外进行评估，但现有方法均针对欧几里得几何。如何将神经最优传输扩展到高维黎曼流形，仍是一项尚待解决的挑战。本文证明，任何通过生成传输映射的离散近似来处理流形上最优传输的方法，都不可避免地会遭遇维数灾难：若要达到固定精度，所需参数数量会随流形维度呈指数增长。受此局限启发，我们提出黎曼神经最优传输（RNOT）映射，这是一种对流形上最优传输映射进行连续神经网络参数化的方法，避免了离散化，并在构造中融入几何结构。在温和的正则性假设下，我们证明，RNOT 映射能够以关于维度的次指数复杂度近似黎曼最优传输映射。在合成数据集和真实数据集上的实验表明，与基于离散化的基线方法相比，该方法具有更好的可扩展性和有竞争力的性能。

1. 引言

最优传输（OT）将生成建模表述为一个传输问题：目标是寻找一个映射（或耦合），将简单的参考分布（例如高斯分布）推送至数据分布，同时最小化期望成本（Kantorovitch，1958年；Villani，2016年）。在采用二次代价的欧几里得空间中，最优蒙日映射具有一种特别易于处理的表示形式：在温和条件下，它是某个凸势函数的梯度（Brenier，1991年）。神经最优传输方法利用这一结构，通过神经网络对映射或其势函数进行参数化，并通过优化最优传输目标函数对其进行训练（Makkuva 等人，2020年；Korotin 等人，2023年；Geuter 等人，2025年）。训练完成后，采样过程得以摊销：从参考分布中抽取样本，并通过学习到的传输映射（或由学习到的势函数诱导的映射）将其向前推送，从而生成一个数据样本。

将计算最优传输扩展到欧几里得空间之外仍面临挑战。许多数据集天然定义在黎曼流形上，例如球面和环面，尤其是在观测值表示角度、轴或方向时（Mardia & Jupp，1999年），其应用包括蛋白质建模（Hamelryck 等人，2006年；Mardia 等人，2006年；Boomsma 等人，2008年）、地质学（Peel 等人，2001）和机器人学（Feiten 等人，2013年；Senanayake & Ramos，2018年）。目前已有一些有前景的尝试，旨在将基于最优传输的生成模型引入流形；例如，黎曼凸势映射（RCPM）通过以下方式构建流形传输层：c-凹势函数，这些势函数表示为有限多个平方距离模板的最小值，并锚定于有限个点位上（Cohen 等人，2021年）。此类方法属于广义的基于离散化的范式：学习到的传输由有限离散表示中介，因此得到的映射具有内在的离散复杂度。在本文中，我们证明这一范式面临着根本性障碍：任何在紧黎曼流形上构造最优传输映射离散近似的方法，都必然遭受维数灾难（CoD）；若要达到固定的近似精度，所需参数数量将随流形维数呈指数增长。

这一负面结果并不排除这样一种可能性：存在一种真正连续的流形神经最优传输框架，能够摆脱维数灾难。恰恰相反，几何深度学习领域近期的逼近理论表明，在适当条件下，神经网络架构在学习流形之间的结构化函数时，可以避免维度导致的爆炸式增长(Kratsios &Papon，2022年)。然而，尽管欧几里得空间中已有基于理论的神经最优传输框架（Makkuva 等人，2020年；Korotin 等人，2023年；Geuter 等人，2025年），据我们所知，目前尚无同样严谨的神经最优传输框架，能够在一般黎曼流形上以内蕴方式学习传输势函数和映射，遑论还能提供保证，确保其内蕴地摆脱维数灾难。

[1]伦敦帝国理工学院，伦敦，英国 [2]瑞典皇家理工学院，斯德哥尔摩，瑞典 [3]哥本哈根大学，哥本哈根，丹麦。联系作者：Alessandro Micheli<a.micheli19@imperial.ac.uk>。

预印本。*2026*年*2*月*4*日。

摆脱维数灾难。

我们的主要贡献是通过引入黎曼神经最优传输（RNOT）来弥合这一差距。RNOT是一个具有理论依据的神经最优传输框架，可直接在黎曼流形上学习摊销传输映射，并具备可证明的多项式复杂度保证。RNOT避免对流形进行离散化。相反，它将传输势表示为 $\mathcal{M}$ 上的连续函数，并借助 c-变换，通过构造强制满足二次代价下 c-凹性这一结构性最优传输约束。由此，通过指数公式 $T(x) = \exp_x(-\nabla\phi(x))$，得到一族实用的流形值传输映射，从而能够将参考分布的样本前推，实现样本外生成。我们建立了一项普适性结果，表明在我们的隐式 c-凹函数类中对势函数进行逼近，所诱导的传输映射会收敛到真实的黎曼最优传输映射。对于神经网络实例化，我们推导出为以规定的逐点精度逼近最优传输映射所需的、关于神经网络参数数量与深度的显式多项式界。最后，在合成数据集与真实流形值数据集上的实验表明，相较于基于离散化的基线方法，该方法提升了可扩展性，并具有竞争力的性能。

2. 背景

全文中，$(\mathcal{M}, g)$ 表示一个连通、紧致、光滑且无边界的p维黎曼流形，其测地距离为 $d(\cdot,\cdot)$。令 $\mathcal{P}(\mathcal{M})$ 表示 $\mathcal{M}$ 上的博雷尔概率测度集合，并记$\mathrm{vol}_{\mathcal{M}}$ 为黎曼体积测度。附录A和B中简要回顾了黎曼几何和最优传输所需的概念，并介绍了相关符号。

2.1. 最优传输背景

最优传输中的一个核心对象是 c-变换，它将勒让德–芬切尔变换推广到了黎曼流形上。设 $c\colon \mathcal{M}\times\mathcal{M}\to\mathbb{R}$ 为一个代价函数。

定义 2.1 (引自Cordero-Erausquin 等人(2001))。集合 $\Psi_c(\mathcal{M})$ 由 c-凹函数构成，即所有满足以下条件的函数 $\phi\colon \mathcal{M}\to\mathbb{R}\cup\{-\infty\}$不恒等于 $-\infty$，且存在函数 $\psi\colon \mathcal{M}\to\mathbb{R}\cup\{-\infty\}$ 使得

$$\phi(x) = \inf_{y\in\mathcal{M}} \big(c(x,y) - \psi(y)\big) \quad \forall x\in\mathcal{M}. \tag{1}$$

我们将 ϕ 称为 c的变换，其中 ψ，并将式（1）简写为 $\phi = \psi^c$。类似地，给定 $\phi\in\Psi_c$（$\mathcal{M}$），我们定义其c变换 $\phi^c\in\Psi_c$（$\mathcal{M}$）为

$$\phi^c(y) := \inf_{x\in\mathcal{M}} \big(c(x,y) - \phi(x)\big), \quad \forall y\in\mathcal{M}. \tag{2}$$

对于 $\phi\in\Psi_c(\mathcal{M})$，由（2）可以直接证明 $\phi^{cc} = \phi$（3）。如 McCann (2001) 所述，$\mathcal{M}$ 以及 $c(x,y)$ 的局部 Lipschitz 正则性意味着 ϕ^c是利普希茨连续的，无论 $\phi\colon \mathcal{M}\to\mathbb{R}\cup\{-\infty\}$是否连续。因此，利用（3），不失一般性，我们可以假设（ψ 和 ϕ）中的函数（1）属于 $C(\mathcal{M},\mathbb{R})$。

最优传输研究如何高效地将质量从源测度 $\mu\in\mathcal{P}(\mathcal{M})$ 转移到目标测度 $\nu\in\mathcal{P}(\mathcal{M})$。在蒙日问题（MP）的表述中，问题是寻找一个可测映射 $T\colon \mathcal{M}\to\mathcal{M}$，使得 $T_{\#}\mu = \nu$ 并使运输代价最小

$$\inf_{T_{\#}\mu=\nu} \int_{\mathcal{M}} c\big(x, T(x)\big)\,\mathrm{d}\mu(x). \tag{MP}$$

在本文中，我们关注平方距离代价

$$c(x,y) := \tfrac{1}{2}d(x,y)^2.$$

当 μ 相对于体积测度$_{\mathcal{M}}$绝对连续时，McCann（McCann，2001，定理 9）证明存在一个最优传输映射 $T\colon \mathcal{M}\to\mathcal{M}$，它在 μ 几乎处处唯一，将 μ 向前推送至 ν（即 $T_{\#}\mu=\nu$），并使（MP）达到最小值。此外，最优映射由一个 c凹势函数诱导：存在一个与最优传输问题相关的 c凹函数 ϕ，使得

$$T(x) = \exp_x\big(-\nabla\phi(x)\big),$$

其中 $\exp_x\colon T_x\mathcal{M}\to\mathcal{M}$ 是黎曼指数映射，∇ 表示黎曼梯度。

2.2. 几何深度学习背景

几何深度学习中一种常见的策略是借助适当的特征映射（或嵌入），将流形上的学习转化为欧几里得空间中的学习。由于我们的目标是在黎曼流形上建立最优传输势函数和映射的通用逼近结果，因此我们回顾标准欧几里得逼近定理中的一致收敛概念，并概述它如何适用于紧域。

对于连续函数空间，我们采用紧集上一致收敛（ucc）拓扑。对于拓扑空间 X，若且唯若序列 $(f_k)_{k\in\mathbb{N}}\subset C(X,\mathbb{R})$ 在 ucc 拓扑下收敛到 $f\in C(X,\mathbb{R})$，则

$$\forall K\subset X \text{ compact}, \quad \sup_{x\in K}|f_k(x) - f(x)| \xrightarrow[k\to\infty]{} 0. \tag{4}$$

当 $X=\mathbb{R}^n$，这是在非紧域上逼近所用的标准收敛模式；等价地，它是由通常的紧集上一致收敛度量所诱导的拓扑（见Kratsios & Bilokopytov (2020年，式 (2)，第 2.2 节)）。当X 为紧致时，（4）退化为通常的一致收敛，因为可以取 $K=X$。特别地，由于 $\mathcal{M}$

在我们的设定中是紧致的， $C(\mathcal{M}, \mathbb{R})$ 上的ucc拓扑与由一致范数诱导的拓扑相重合

$$\|g\|_\infty := \sup_{x\in\mathcal{M}} |g(x)|. \tag{5}$$

Following Kratsios & Bilokopytov（2020），我们引入一个连续特征映射 $\varphi: \mathcal{M} \to \mathbb{R}^n$，并考虑由复合诱导的 $\mathcal{M}$ 上的函数类。本文始终用 $\mathcal{F}$ 表示在 ucc 下 $C(\mathbb{R}^n, \mathbb{R})$ 的一个稠密子集，例如 Leshno 等（1993）、Lu 等（2017）和 Zhou （ 2020 ）研究的神经网络架构，或如 Micchelli 等人（2006）所述，具有通用核的高斯过程的后验均值。随后，我们定义以下 φ 拉回（或特征诱导）类:

$$\varphi^*\mathcal{F} := \{f \circ \varphi : f \in \mathcal{F}\}.$$

复合类 $\varphi^*\mathcal{F}$ 的逼近能力不仅取决于 $\mathcal{F}$，还关键地取决于特征映射 φ 的几何性质。具体而言， φ 必须保留 $\mathcal{M}$ 上各点的足够信息，以使不同的点在嵌入后仍可区分。

假设 2.2（特征映射正则性）。 特征映射 $\varphi: \mathcal{M} \to \mathbb{R}^n$ 连续且单射。

假设 2.2恰好是欧几里得近似能够迁移到流形上的条件。非正式地说， Kratsios& Bilokopytov(2020年)表明:

- 若 $\mathcal{F}$ 在 $\mathbb{R}^n$ 上具有通用性（在 $C(\mathbb{R}^n, \mathbb{R})$ 中关于 ucc 稠密），且 φ 是连续且单射的，那么 $\varphi^*\mathcal{F}$ 在 $\mathcal{M}$ 上具有通用性（在 $C(\mathcal{M}, \mathbb{R})$ 中关于 $\|\cdot\|_\infty$稠密）；参见Kratsios & Bilokopytov (2020年，定理3.3）。
- 反过来，单射性实际上是必要的：若 φ不是单射的，则存在定义在 $\mathcal{M}$ 上的连续函数，无法通过复合 $f \circ \varphi$表示（或一致近似）；参见 Kratsios & Bilokopytov (2020年，定理 3.4）。

就我们的目的而言，一类特别相关的特征映射，是通过编码点到一组地标点的距离而得到的。在紧致的黎曼情形中，Gromov (1983)的经典构造表明，可以选择这类映射使其具有单射性。

命题2.3（到地标点距离嵌入（Gro-mov，1983年））。对于充分小的 $\delta > 0$，令 $\{x_i\}_{i\in I}$ 为 δ- 分离子集 $\mathcal{M}$（由紧致性可知其为有限集），并定义

$$\varphi(x) := (d(x, x_i))_{i\in I} \in \mathbb{R}^I.$$

那么 φ 是连续且单射的（因此满足假设 2.2）。

2.3. 深层 ReLU 网络的（非）维数灾难逼近率背景

在逼近理论中，形式化描述维数灾难的一种便捷方式，是考察模型规模（此处以参数数量 W衡量）必须如何增长，才能在一个 ε 维区域上保证给定的一致精度 n。作为典型基准，可考虑 H¨older 单位球 $\mathcal{H}_{r,\ n}$ 即立方体 r上的 $[0,1]^n$-光滑函数（见（16））。在这种情况下，目标是得到如下形式的估计:

$$\sup_{f\in\mathcal{H}_{r,n}} \|f - \widehat{f}_W\|_\infty \lesssim W^{-\beta},$$

其中， $\widehat{f}_W$ 表示由具有 W 个参数的模型产生的逼近， $\beta > 0$ 表示逼近指数，我们将其称为速率。

Yarotsky & Zhevnerchuk（2020）关于使用深层 ReLU 网络对 H¨lder 球 $\mathcal{H}_r$、$_N$ 进行一致逼近的相图，为可达到的速率提供了清晰的视角。在其框架中，如果存在具有 W 个权重的网络架构及权重分配映射，使得在 $\mathcal{H}_r$、$_n$上的最坏情形误差按 $\mathcal{O}(W^{-\beta})$衰减（当 $W \to \infty$ 时），则称速率 β 是可达到的。对于 H¨lder 类，经典参考速率为 r/n：在连续区间内，权重必须连续地依赖于目标函数，Yarotsky & Zhevnerchuk（2020，定理 3.1 和定理 3.2）证明， r/n是精确阈值——该速率可以达到，并且在连续性约束下无法改进。允许不连续权重分配并使用足够深的网络后，可以进入更快的深度不连续阶段：对于任意 $r > 0$ 和任意指数 $\frac{r}{n} < \beta < \frac{2r}{n}$，都存在能够实现速率 β 的深度ReLU构造（Yarotsky & Zhevnerchuk，2020，定理 3.3 和图 3）。

Kratsios&Papon (2022年) 利用深度—不连续区间，获得了流形上的逼近保证，避免了对维度的指数依赖。他们并不追求在 $\mathcal{M}$ 上的一致逼近；相反，他们只在固定数据集 $\mathcal{D} \subset \mathcal{M}$ 上逼近目标分布，而非在 $\mathcal{M}$ 上一致逼近。具体而言，对于 $f \in C^{kp,\ 1}(\mathcal{M})$，当 $k \in \mathbb{N}$ (等价地，H¨older 光滑性 $r = kp + 1$) 时，他们证明 f 可以由深度神经网络在 ε 上以精度 $\mathcal{D}$ 进行一致逼近，且网络在 ε^{-1} (Kratsios &Papon，2022年，第 3.3 节）的复杂度是次指数级的（事实上是多项式级的）。

3. 为什么离散最优传输在高维下表现不佳

在本节中，我们针对黎曼流形上的最优传输建立了一个一般性的维数灾难障碍：当学习得到的传输映射具有有限支撑时，即它只能将样本发送到有限多个目标位置。我们的结果广泛适用于任何其离散化要么以显式方式引入

（例如通过网格），要么通过模型参数化以隐式方式引入的方法，并且不论具体的算法或优化过程如何均成立。作为一个具体实例，我们证明，据我们所知，RCPM是目前唯一直接构造黎曼最优传输映射的现有方法；它们在任何有限模型规模下都会产生离散输出映射，因此继承了对维度的相同指数依赖。

离散输出映射与近似误差。 我们固定 $\mu, \nu \in \mathcal{P}(\mathcal{M})$，并令 $T_\star$ 表示将 μ 推前至 ν 的二次代价函数 $c(x,y) = \frac{1}{2}d(x,y)^2$对应的最优传输映射（在 μ-几乎处处意义下定义）。为了在源测度 μ下比较可测映射 $T, S\colon \mathcal{M} \to \mathcal{M}$，我们使用 μ-均方根误差

$$\mathrm{RMSE}_\mu(T,S) := \left(\int_{\mathcal{M}} d\big(T(x), S(x)\big)^2 \, \mathrm{d}\mu(x) \right)^{1/2}.$$

在离散 最优传输参数化中，一个常见的结构限制是，学习到的映射只能输出有限多个点；例如，它可能将 $\mathcal{M}$ 划分为若干单元，并将这些单元映射到有限个点位。我们通过离散输出映射这一类来刻画这种离散化效应。对于 $m \in \mathbb{N}$，定义 $\mathsf{D}_m(\mu)$ 为可测映射$T\colon \mathcal{M} \to \mathcal{M}$ 的集合，其中推前 $T_{\#}\mu$ 的支撑集至多包含 m 个点，即 $\#\mathrm{supp}(T_{\#}\mu) \le m$。等价地，$T \in \mathsf{D}_m(\mu)$ 当且仅当存在 $y_1, \dots, y_m \in \mathcal{M}$，使得对于 μ-几乎处处的 x，$T(x) \in \{y_1, \dots, y_m\}$。

离散输出映射会引发维数灾难壁垒。下一个定理揭示了离散输出参数化的核心瓶颈：对 $T_\star$ 进行 $\mathsf{D}_m(\mu)$内的任何近似，m都会必然导致对 ν的原子数为$\nu \ll \mathrm{vol}_{\mathcal{M}}$ 的近似（如定理3.1所示）。这一步量化会带来一个无法避免的下界，其依赖于流形内在维数，且呈指数关系 p。

定理 3.1 (离散输出映射的维数灾难壁垒)。 假设 $\mu, \nu \ll \mathrm{vol}_{\mathcal{M}}$，并令 $T_\star$ 为从 μ 到 ν、针对 $c(x,y) = \frac{1}{2}d(x,y)^2$的最优传输映射。那么存在$C > 0$ （依赖于 $\mathcal{M}$ 和 ν），使得对所有 $m \in \mathbb{N}$,

$$\inf_{T \in \mathsf{D}_m(\mu)} \mathrm{RMSE}_\mu(T, T_\star) \;\ge\; C\, m^{-1/p}. \tag{6}$$

特别地，要达到 均方根误差$\mu(T, T_\star) \le \delta$，需要 $m \ge (C/\delta)^p$。

定理 3.1 的证明推迟至附录 H.1。定理 3.1 表明，这种维数灾难壁垒是基于离散化的黎曼最优传输方法所固有的，与架构和优化无关。

现在我们将讨论重点放在 RCPM 上。对于 $m \in \mathbb{N}$，RCPM 考虑如下形式的 $\mathcal{C}_m(\mathcal{M})$ 类 离散 c-凹势函数

$$\phi_m(x) \;=\; \min_{i \in [m]} \big(c(x, y_i) + \alpha_i \big), \quad y_i \in \mathcal{M},\ \alpha_i \in \mathbb{R}.$$

Cohen 等人 (2021年) 的一项核心定理表明，在任意紧致光滑无边界流形上，族$\{\mathcal{C}_m(\mathcal{M})\}_{m \in \mathbb{N}}$ 在所有 c-凹势函数的集合上是稠密的（在适当的意义下）。特别地，相关的传输映射可以任意精确地逼近 OT 映射，当 $m \to \infty$。

尽管具有这种渐近表达能力，RCPM 在任何有限 m下仍然会受到维数灾难的影响。事实上，对于$\phi \in \mathcal{C}_m(\mathcal{M})$，相关映射

$$T_\phi(x) := \exp_x\big(-\nabla\phi(x)\big),$$

取值于有限站点集 $\{y_1, \dots, y_m\}$。因此，$(T_\phi)_{\#}\mu$ 是一个 m-原子测度，而定理3.1给出了一个依赖于维度的逼近误差下界。

推论 3.2(RCPM 的维数灾难屏障)。在定理*3.1*的假设下，对于每个 $m \in \mathbb{N}$ 以及每个$\phi \in \mathcal{C}_m(\mathcal{M})$，相关映射 T_ϕ 满足 $T_\phi \in \mathsf{D}_m(\mu)$。因此，存在 $C > 0$ 使得对于所有 $m \in \mathbb{N}$,

$$\inf_{\phi \in \mathcal{C}_m(\mathcal{M})} \mathrm{RMSE}_\mu(T_\phi, T_\star) \;\ge\; C\, m^{-1/p}, \tag{7}$$

并达到 均方根误差$\mu(T_\phi, T_\star) \le \delta$ 需要 $m \ge (C/\delta)^p$。

推论 3.2 的证明见附录 H.2。

4. 黎曼神经最优传输

我们已经证明，离散化是黎曼流形上实现可扩展最优传输的一项根本障碍。因此，我们采取一种互补的方法，基于最优传输映射的连续参数化，完全避免离散表示。具体而言，我们引入黎曼神经最优传输（*RNOT*）映射，利用神经网络对传输势函数进行参数化，并通过黎曼指数映射恢复相应的传输映射。由此得到一个流形值生成器，其推前不受限于有限多个位置，并且在构造上保留了最优传输的几何结构。

隐式连续表示。 为定义 RNOT，我们从一个通用框架入手，用于在黎曼流形上构造最优传输映射的连续参数化。核心思想是在一个具有表达能力的函数类中近似传输势，同时通过 c-变换，在构造上强制满足 c-凹性这一结构要求。

回顾第 2.2 节可知，$\mathcal{F}$ 在 $C(\mathbb{R}^n, \mathbb{R})$ 中关于 ucc拓扑是稠密的。给定一个特征映射 $\varphi\colon \mathcal{M} \to \mathbb{R}^n$，它

诱导出拉回类

$$\varphi^*\mathcal{F} = \{f \circ \varphi : f \in \mathcal{F}\} \subset C(\mathcal{M}, \mathbb{R}).$$

由于最优传输势函数是 c-凹的，我们通过采用 c-变换来施加这一性质。对于任意 $\mathcal{G} \subset C(\mathcal{M}, \mathbb{R})$，定义相应的隐式 c-凹类

$$\mathfrak{C}(\mathcal{G}) := \{\psi^c : \psi \in \mathcal{G}\}.$$

特别地，我们在 $\mathfrak{C}(\varphi^*\mathcal{F})$中近似最优传输势函数，从而将 c-凹性内置其中，而非通过外部约束来强制施加。以下定理表明，在这一隐式 c-凹类中近似势函数，会导出收敛到真实黎曼最优传输映射的传输映射。

定理 *4.1*（隐式 c-凹势函数的普适性。）假设 μ, $\nu \in \mathcal{P}(\mathcal{M})$，其中 $\mu \ll vol_{\mathcal{M}}$，并令 $T_\star$为针对 $c(x,y) = \frac{1}{2}d(x,y)^2$、从 μ到 ν的最优传输映射。假设 $\mathcal{F}$ 是 $C(\mathbb{R}^n, \mathbb{R})$在 *ucc*拓扑下的稠密子集，且特征映射 φ: $\mathcal{M} \to \mathbb{R}^n$ 满足假设 *2.2*。那么存在一个势函数序列 $\{\phi_k\}_{k\geq 1} \subset \mathfrak{C}(\varphi^*\mathcal{F})$，其对应的映射

$$T_k(x) := \exp_x\big(-\nabla\phi_k(x)\big)$$

使得

$$T_k(x) \longrightarrow T_\star(x) \qquad \textit{for } \mu\textit{-almost every } x \in \mathcal{M}.$$

特别地，$T_k \to T_\star$ 在 μ下依概率收敛。

定理 4.1的证明推迟至附录H.3。

RNOT 势函数与映射。 现在，我们用神经网络实例化该框架，从而得到一类可训练的黎曼最优传输势函数与映射。固定一个激活函数σ： $\mathbb{R} \to \mathbb{R}$并令 $\mathcal{NN}_n^\sigma \subset C(\mathbb{R}^n, \mathbb{R})$ 表示一个前馈神经网络实现函数类，该类在$C(\mathbb{R}^n, \mathbb{R})$ 中关于 ucc 拓扑稠密。对于σ 的标准选择（例如，非多项式激活函数），由通用逼近定理可知其在 ucc 拓扑下稠密；参见 Leshno 等人(1993)； Lu 等人 (2017)；Zhou (2020）。由此得到以下定义。

定义 4.2（黎曼神经最优传输势函数与映射）。令 φ： $\mathcal{M} \to \mathbb{R}^n$ 为满足假设 2.2的特征映射，并令 $\mathcal{NN}_n^\sigma$ 如上所述。将*RNOT* 势函数定义为以下类：

$$\mathfrak{C}\big(\varphi^*\mathcal{NN}_n^\sigma\big) = \big\{(f \circ \varphi)^c : f \in \mathcal{NN}_n^\sigma\big\}.$$

相关的 *RNOT* 映射由指数映射公式得到

$$T(x) = \exp_x\big(-\nabla\phi(x)\big), \qquad \phi \in \mathfrak{C}\big(\varphi^*\mathcal{NN}_n^\sigma\big).$$

这一特例构建了一类实用的神经网络假设类，其中包含 RNOT 势函数和映射，可通过隐式层实现，并利用样本和 Kantorovich 半对偶目标函数进行端到端训练，从而避免归一化流模型基于似然/KL 的训练中所需的昂贵雅可比行列式计算（例如 Cohen 等人 (2021年); Rezende 等人 (2020年); Mathieu 和 Nickel (2020年)）。训练与实现细节见附录 E。

5. 黎曼神经最优传输突破维数灾难

在第4节引入RNOT之后，我们现在证明*RNOT*突破维数灾难：对于充分正则的问题，Kantorovich势函数与所得传输映射均可被逼近至精度 ε，且所用RNOT模型的参数数量与深度随ε^{-1} 多项式增长（且不会随流形维度指数增长）。全文中，我们采用非仿射分段线性激活函数以及架构类$\mathcal{NN}^\sigma_{n,\,W,\,L}$，输入维度为 n，可训练参数数量为 W，层数为L层（定义H.7–H.8）。我们将论证分为三个步骤。

步骤 1：在 $\mathcal{M}$上具有多项式ε^{-1}复杂度的函数逼近。我们的第一个结果给出了 $\mathcal{M}$ 上充分光滑函数的一致逼近，并对参数数量 W_ε 与深度 L_ε建立了显式界。

定理 5.1（多项式 ε复杂度，适用于流形上的函数）。设 σ： $\mathbb{R} \to \mathbb{R}$ 为任意非仿射分段线性激活函数。固定 $k \in \mathbb{N}$，并令 $\psi \in C^{kp,\,1}(\mathcal{M}, \mathbb{R})$。则对每一个 $\varepsilon \in (0,\,1)$，存在特征映射 $\varphi_\star$： $\mathcal{M} \to \mathbb{R}^{2p}$ 满足假设2.2、整数 W_ε、$L_\varepsilon \in \mathbb{N}$，以及神经网络$\hat{}g_\varepsilon \in \mathcal{NN}^\sigma_{2p,\,W_\varepsilon,\,L_\varepsilon}$，使得拉回网络

$$\hat{\psi}_\varepsilon := \hat{g}_\varepsilon \circ \varphi_\star \in \varphi_\star^*\mathcal{NN}^\sigma_{2p,W_\varepsilon,L_\varepsilon}$$

一致近似 ψ：

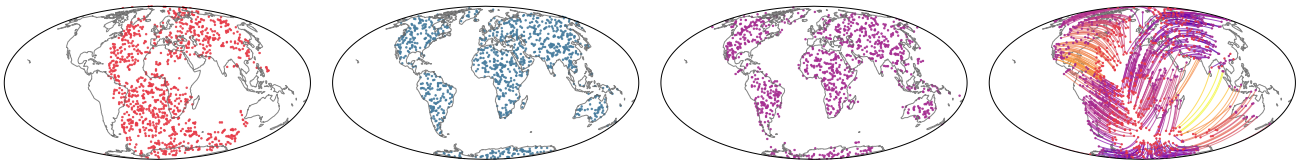
$$\|\psi - \hat{\psi}_\varepsilon\|_\infty < \varepsilon.$$

此外，$\hat{}g_\varepsilon$ 满足以下复杂度估计：

- $W_\varepsilon = \mathcal{O}\left(\varepsilon^{-\frac{4p}{3(kp+1)}}\right)$,
- $L_\varepsilon = \mathcal{O}\left(\varepsilon^{-\frac{2p}{3(kp+1)}}\right)$.

定理 5.1 的证明推迟至附录H.4。

定理5.1 强化了Kratsios & Papon的逼近保证(2022年)：他们将误差控制在预先指定的有限数据集上，使其保持一致 $\mathcal{D} \subset \mathcal{M}$，而我们则在整个流形上获得了一致逼近界 $\mathcal{M}$（在所述光滑性假设下）。



(a) 源 μ （1.5亿年前） (b)目标分布 ν （现在） (c) 传输后的分布 $T_{\#}\mu$ (d) 传输映射 T

图 1。 $\mathbb{S}^2$上的大陆漂移最优传输。从左到右依次为：源质量分布 μ （$\sim$ 150 百万年前）、目标分布 ν （现在）、传输后的分布 $T_{\#}\mu$，以及由学习得到的传输映射 T所诱导的测地线轨迹。

步骤 2：通过 c-变换一致逼近 Kantorovich 势函数。现在，我们将定理 5.1 的函数逼近保证推广到最优传输势函数。RNOT 将对偶势的搜索限制在假设类 $\mathfrak{C}(\varphi_\star^*\mathcal{NN}^\sigma_{2p,W,L})$ 中，即拉回网络的 c-变换。因此，要逼近一个最优传输势函数 $\phi_\star$，只需将其预势 $\psi_\star := \phi_\star^c$ 作为定义在 $\mathcal{M}$ 上的普通标量函数进行逼近。我们证明，若 $\hat{}\psi_\varepsilon$ 一致逼近 $\psi_\star$，则其 c-变换 $\hat{}\phi_\varepsilon = \hat{}\psi_\varepsilon^c$ 属于 $\mathfrak{C}(\varphi_\star^*\mathcal{NN}^\sigma_{2p,W_\varepsilon,L_\varepsilon})$，并以相同的 ε- 复杂度一致逼近 $\phi_\star$。因此，RNOT 类中的最优传输势函数在参数数量和深度方面具有多项式级缩放，从而避免了第 3 节指出的离散化引起的维数灾难。

推论 5.2 (RNOT 势函数的多项式 ε^{-1}复杂度)。 假设 μ, $\nu \in \mathcal{P}(\mathcal{M})$ 满足 $\mu \ll$ 体积$_{\mathcal{M}}$，并令 $\phi_\star$ 为 $c(x,y) = \frac{1}{2}d(x,y)^2$的一个最优传输势函数。固定 $k \in \mathbb{N}$，并假设其预势 $\psi_\star := \phi_\star^c$ 属于 $C^{kp,1}(\mathcal{M},\mathbb{R})$。令 σ 为任意非仿射分段线性激活函数。

对于任意 $\varepsilon \in (0,1)$，令 $\hat{}\psi_\varepsilon$ 为定理 5.1 应用于 $\psi_\star$时给出的拉回网络 $\|\psi_\star - \hat{}\psi_\varepsilon\|_\infty < \varepsilon$，使得 c-凹势函数 $\hat{}\phi_\varepsilon := \hat{}\psi_\varepsilon^c$。定义诱导的 $\|\psi_\star -$

$$\|\phi_\star - \hat{\phi}_\varepsilon\|_\infty \le \varepsilon,$$

实现 $\hat{}\psi_\varepsilon$ 的网络具有参数数量 W_ε和深度 L_ε，满足与定理 5.1中相同的界。

推论的证明5.2 推迟到附录H.5。

步骤 3：利用极小值点的稳定性逐点近似最优传输映射。传输映射通过内层问题 $T(x) \in \arg\min_{y\in\mathcal{M}}\{\frac{1}{2}d(x,y)^2 - \psi(y)\}$隐式定义，因此，势函数的微小扰动原则上可能导致所选极小化元发生较大变化。我们证明，在温和的正则性假设下，这种情况不会发生：远离奇异的几何构型时，对势函数进行一致的 ε近似，可在 $\mathcal{M}$的一个满μ测度子集上，对诱导映射进行逐点控制，误差上界为 $O(\sqrt{\varepsilon})$。

定理 5.3 （RNOT 映射的逐点稳定性）。 设μ, $\nu \in \mathcal{P}(\mathcal{M})$ 满足 μ, $\nu \ll \mathrm{vol}_{\mathcal{M}}$，并考虑二次代价 $c(x,y) = \frac{1}{2}d(x,y)^2$.设 $\phi_\star$ 是从 μ 到 ν的最优传输势函数，并令其预势为$\psi_\star := \phi_\star^c$.固定 $k \in \mathbb{N}$并假设 $\psi_\star \in C^{kp,1}(\mathcal{M})$。

于是存在一个可测集 $\Omega \subset \mathcal{M}$，满足 $\mu(\Omega) = 1$，使得对每个 $x \in \Omega$，都存在常数 $C(x) > 0$ 和$\varepsilon_x > 0$，具有以下性质。

对于任意 $\varepsilon \in (0, \min\{1, \varepsilon_x\})$ 以及任意满足以下条件的连续函数 $\psi_\varepsilon \in C(\mathcal{M})$

$$\|\psi_\varepsilon - \psi_\star\|_\infty \le \varepsilon,$$

通过选取任意一个极小化解，定义映射（可能是集合值映射）T_ε

$$T_\varepsilon(x) \in \arg\min_{y\in\mathcal{M}}\Big\{\tfrac{1}{2}d(x,y)^2 - \psi_\varepsilon(y)\Big\}.$$

令 $T_\star$ 为由 $\phi_\star$诱导的最优传输映射。那么，每个这样的选取都满足

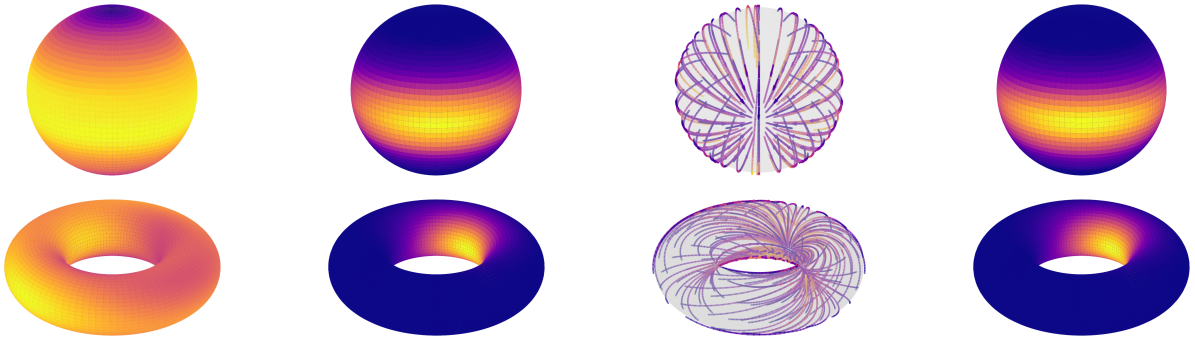
$$d(T_\varepsilon(x), T_\star(x)) \le 2\sqrt{\frac{\varepsilon}{C(x)}}.$$

定理 5.3 的证明推迟至附录 H.6。

结合定理 5.3 与推论 5.2，可在全测度集上以多项式 ε^{-1}复杂度逐点近似最优传输映射 $T_\star$，逐点误差为μ, 测度$O(\sqrt{\varepsilon})$。

6. 实验

在本节中，我们以流形归一化流基线为对照，对 RNOT 映射的实证性能进行基准评测。我们首先在真实世界地质数据集上评估建模保真度，该数据集源于发生在 $\mathbb{S}^2$ 上的大陆漂移。随后，我们利用高维球面 $\mathbb{S}^n$ 与环面 $\mathbb{T}^n$ 上的合成实验，研究相对于流形维数的可扩展性。对于 RNOT 实现，我们采用命题 2.3 中的格罗莫夫嵌入，并使用最远点采样（FPS）或均匀随机采样（RND）选择其地标集，从而评估地标几何对性能的影响；详见附录 E 。实验细节见附录 F。代码可应要求提供。



(a) 源 μ　(b) 目标分布 ν　(c) 传输 T　(d) 推前 $T_{\#}\mu$

图2. 在 $\mathbb{S}^2$ (上)和 $\mathbb{T}^2$ (下)上的最优传输，分别对应表 1 和 2 中的一次代表性运行。从左到右：均匀源 μ、包裹正态目标分布 ν、由学习到的传输诱导的测地线轨迹 T以及推前 $T_{\#}\mu$。

6.1. 真实案例研究：大陆漂移

根据 Cohen 等人 （2021年），我们研究流形生成建模的一项地球物理应用：球面 $\mathbb{S}^2$ 上的大陆漂移，并使用来自 M¨uller 等人（2018年）的古地理重建。我们考虑两种陆地质量分布：一种对应于地球 ~ 150百万年前（图 1（a）），另一种对应于现今地球（图1（b））。我们的目标是学习一个摊销映射 T，将源分布传输到目标分布。

我们在 $\mathbb{S}^2$ 上训练了一个 RNOT 映射，使用 FPS 地标嵌入和基于样本的 Kantorovich 半对偶目标函数（附录 E 和 F）。如图 1(c) 所示，学得的推前 $T_{\#}\mu$ 与当今的质量分布高度吻合。除了密度匹配之外，学得的传输 T还在“古代”地球与当今地球的各个位置之间建立了明确的对应关系。

为了直观呈现诱导出的传输动态，我们考虑映射的一参数族

$$T_t(x) \;=\; \exp_x\big(-t\,\nabla\phi(x)\big), \qquad t \in [0,1],$$

该映射族在恒等映射（$t = 0$）与学得的传输映射（$t = 1$）之间进行插值。图 1(d) 展示了从 $\mathbb{S}^2$ 中选取的起点所诱导出的轨迹，这些起点采样自 μ。这些曲线直观地呈现了与板块构造运动相符的质量位移：欧亚大陆与北美洲交界处附近的点向相反方向分开，反映了两个板块在地质时期中的分离（参见 Wilson，1963年）。

6.2. 球面和环面上的合成实验

我们通过在球面和环面上开展受控合成实验，补充真实场景案例研究。我们首先在 $\mathbb{S}^2$ 和 $\mathbb{T}^2$上比较低维情况下的方法，随后在 $\mathbb{S}^n$和 $\mathbb{T}^n$ 上改变维度，重复相同的实验，以评估可扩展性。

在这两类流形中，源分布均为流形上的均匀分布，目标分布则是尺度为 σ =0.3 的包裹正态分布，其中心位于 $\mathbb{S}^n$ 上的南极（$-1, 0, \ldots, 0$），以及 $\mathbb{T}^n$上的对应点（$\pi, \ldots, \pi$）。我们将其与 RCPM （ Cohen 等人，2021 ）、RCNF（Mathieu 和 Nickel，2020）和莫泽流（ Rozen 等人，2021）进行比较（理由和实现细节见附录 F）。与 Cohen 等人（2021）相同，我们引入正则化参数 γ并对 $\gamma \in \{1.0, 0.1, 0.05, 0.01, 0.005, 0.001\}$进行扫描，其中极限值 $\gamma \to 0$ 对应于硬（离散）最小值。我们通过前向 KL 散度 $D_{\mathrm{KL}}(T_{\#}\mu \,\|\, \nu)$ 和有效样本量（ESS）衡量性能，并报告 5 次独立运行（每次 1024 个样本）的均值 $\pm$ 置信区间。我们还报告每次运行的平均墙钟时间（秒）。

对于此处报告的所有 RCPM 结果（在 $\mathbb{S}^2$ 和 $\mathbb{T}^2$ 上），我们使用 $\gamma = 1$，这能取得最佳性能。在 $\mathbb{S}^2$ 上，RCPM 的 KL 散度最低（0.0037）、有效样本量最高（0.996），而采用 FPS 地标点的 RNOT 映射表现与之相当。

表 1. 各方法在 $\mathbb{S}^2$ 上的 KL 散度和有效样本量。粗体表示最佳结果（在统计显著性范围内），下划线表示次佳结果。报告值为 5 次独立运行的均值及置信区间。

模型	KL ↓	有效样本量 ↑	时间（秒）↓
本文方法（最远点采样）	0.03 ± 0.00	0.97 ± 0.00	986 ± 2
本文方法（RND）	0.04 ± 0.00	0.95 ± 0.00	1100 ± 1
RCPM	**0.0037 ± 0.0008**	**0.996 ± 0.000**	**37.3 ± 0.6**
RCNF	2.38 ± 0.10	0.48 ± 0.02	352 ± 6
莫泽流	1.16 ± 0.03	0.82 ± 0.00	758 ± 16

在 $\mathbb{T}^2$ 上，采用 FPS 地标点的 RNOT 映射显著优于 RCPM，KL 散度为 0.13（而 RCPM 为 0.93），有效样本量为 0.93（而 RCPM 为 0.55）。RCNF 和莫泽流的表现都明显逊于 RCPM 和 RNOT，在

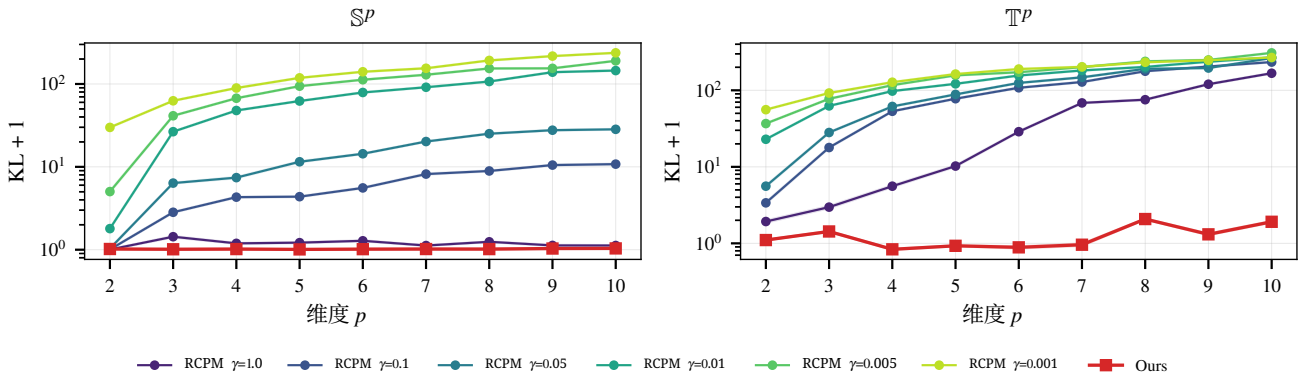


图3. KL 散度（对数尺度）随维度 $p \in \{2, \ldots, 10\}$ 的变化，对应从均匀源到包裹正态目标分布的传输，分别位于$\mathbb{S}^p$（左图）和 $\mathbb{T}^p$（右图）。曲线展示了在一组正则化值 γ扫描下的 RNOT（本文方法）与 RCPM。随着 p 增大，RCPM 的性能下降——在较小的 γ时尤为明显，此时映射较为尖锐——而 RNOT 在各维度上保持稳定。

两者 $\mathbb{S}^2$ 和 $\mathbb{T}^2$。图 2 展示了学习到的 RNOT 传输映射 $\mathbb{S}^2$（上排）以及 $\mathbb{T}^2$（下排）。

表2. 各方法在 $\mathbb{T}^2$ 上的 KL 散度和有效样本量。粗体表示最佳结果（达到统计显著性时），下划线表示次佳结果。报告值为 5 次独立运行的均值及置信区间。

模型	KL ↓	有效样本量 ↑	时间（秒）↓
本文方法（最远点采样）	**0.13 ± 0.01**	**0.93 ± 0.01**	1022 ± 10
本文方法 (RND)	0.22 ± 0.04	0.85 ± 0.02	1057 ± 1
RCPM	0.93 ± 0.02	0.55 ± 0.03	**62.3 ± 0.5**
RCNF	6.31 ± 0.56	0.36 ± 0.04	265 ± 6
莫泽流	6.84 ± 0.21	0.56 ± 0.01	769 ± 5

这种竞争力也伴随着更长的训练时间，因为 RNOT 通过迭代内层最小化来计算隐式的 c-变换。

接下来，我们扫描维度。在图 3中，随着 n 增大，RCPM 的表现会崩溃——尤其是当 γ较小时，此时对数和指数 最接近硬最小值，正如推论 3.2所预测的那样。较大的 γ 会使目标函数更平滑，从而有所帮助，但在实际应用中仍不足以避免维数灾难。相比之下，RNOT 在所有测试维度上的表现基本保持平稳。

由于计算不可行性，图3 的最大维度限制为 10，在此维度下，KL 散度在较小的 γ处便已饱和。维度扩展至 40 的结果见附录图 5。最后，消融研究（附录 G.1）表明，使用 FPS 进行地标选择（而非 RND）以及基于 对数和指数的初始化对性能影响最大，凸显了地标几何和初始化的重要性，也为未来改进指明了明确方向。

7. 讨论与局限性

本文提出了首个在理论上有依据的黎曼流形神经最优传输框架。我们的出发点是一项负面结论：任何输出传输映射离散近似的流形最优传输方法都不可避免地受到维数灾难的影响，需要指数级增长的参数才能达到固定精度。黎曼神经最优传输（RNOT）通过彻底避免离散化来绕过这一障碍：我们在 $\mathcal{M}$ 上参数化一个连续的预势函数，并通过 c-变换以构造方式确保其满足 c-凹性，从而得到内禀流形值映射。

对于非仿射分段线性激活函数，我们是首个证明在 ε^{-1} 下，逼近最优传输势函数所需的神经网络参数数量和深度具有显式多项式界，并通过稳定性论证证明，由此诱导的最优传输映射在一个满μ测度子集上逐点成立。这些保证也体现在维度扫描的实证结果中：图 3显示，随着 n 增大，RCPM 的表现急剧下降，而 RNOT 在所有测试维度上的表现基本保持平稳。实证结果表明，RNOT 在低维情况下具有竞争力，并且在维度扫描中保持稳定：图 3显示，随着 n 增大，RCPM 的表现急剧下降，而 RNOT 在所有测试维度上的表现基本保持平稳；大陆漂移案例研究进一步凸显了所学习的传输轨迹的可解释性。

局限性包括我们对紧流形的关注、速率背后的正则性假设，以及训练成本；训练成本主要来自评估 c-变换所需的迭代式内层最小化。未来的工作包括将理论拓展到更广泛的代价函数和设定，并加速这一内层求解过程（例如通过改进优化或采用摊销）。总体而言，RNOT 为在流形上实现可扩展的摊销最优传输提供了一条有理论依据的途径；据我们所知，它是在这一设定下首个提供维度友好型保证的框架。

影响声明

我们开发了一个用于黎曼流形上最优传输的深度学习框架。这使得我们能够对生物学、医学和物理科学等领域中的几何结构数据进行更有理论依据的建模。

致谢

S.B. 感谢诺和诺德基金会通过诺和诺德青年研究者奖（NNF20OC0059309）提供支持。S.B. 感谢埃里克与温迪·施密特战略创新基金通过施密特博学家奖（G-22-63345）提供支持，该奖项也支持 A.Micheli。S.B. 感谢诺和诺德基金会通过全球病原体防范平台（GPAP）（NNF26SA0109818）提供支持。Y.C. 获数字未来博士后奖学金支持。A.Monod 获英国工程与物理科学研究理事会人工智能数学基础中心支持：人工智能“埃尔朗根纲领”，编号 EP/Y028872/1。

贡献声明

作者贡献按照 CRediT（贡献者角色分类法）进行报告。

- **AlessandroMicheli**：概念构思；方法学；软件；正式分析；监督；调查；项目管理；可视化；验证；撰写初稿；审阅与编辑。
- **Yueqi Cao**：审阅与编辑。
- **AntheaMonod**：撰写初稿；审阅与编辑。
- **Samir Bhatt**：软件；经费获取；资源；可视化；验证；撰写初稿；审阅与编辑。

参考文献

Abadi, M., Agarwal, A., Barham, P., Brevdo,E., Chen, Z.,Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin,M., Ghemawat, S., Goodfellow,I., Harp,A.,Irving,G., Isard,M., Jia, Y., Jozefowicz,R., Kaiser, L.,Kud-lur, M.,Levenberg, J., Man´e, D., Monga, R., Moore,S., Murray, D., Olah, C.,Schuster,M.,Shlens,J.,Steiner, B., Sutskever, I.,Talwar, K., Tucker, P.,Van-houcke, V., Vasudevan,V.,Vi´egas, F., Vinyals, O., War- den, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow：异构系统上的大规模机器学习，2015年。软件可从以下网址获取 https://www.tensorflow.org/。

Ambrosio, L., Bru´e, E., and Semola, D. 最优传输讲义。施普林格·自然瑞士，2024年。Bangert, V. 黎曼流形上凸函数的解析性质。*Journalf¨ur diereine und angewandte Mathematik*，310:309–324，1979年。Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-Lopez, F., Pedregosa, F., and Vert, J.-P. 高效且模块化的隐式微分。载于神经信息处理系统进展，第5230–5242页，2022年。

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A. 和 Hamelryck, T. 局部蛋白质结构的生成式概率模型。美国国家科学院院刊，105(26)：8932–8937，2008年。

Brenier, Y. 向量值函数的极分解与单调重排。纯粹与应用数学通讯，44(4):375–417，1991年。

Burago, D., Burago, Y., Ivanov, S., 等。度量几何教程，第 33 卷。美国数学会，普罗维登斯，2001。

Chen, M., Jiang, H., Liao, W., 和 Zhao, T. 面向低维流形上函数的深层 ReLU 网络高效近似。载于神经信息处理系统进展，2019。

Cohen, S., Amos, B., 和 Lipman, Y. 黎曼凸势映射。载于第 *38* 届国际机器学习会议论文集，第 2028–2038 页，2021 年。

Cordero-Erausquin, D., McCann, R. J., 和 Schmucken-schl¨ager, M. 一种类似于 Borell、Brascamp 和 Lieb 的黎曼插值不等式 `。数学新进展，146(2):219–257，2001。

Cuturi, M. 和 Peyr´e, G. 半对偶正则化最优传输。*SIAM*评论，60(4):941–965，2018年。

De Bortoli, V.、Mathieu, E.、Hutchinson, M.、Thornton, J.、Teh, Y. W. 和 Doucet, A. 黎曼评分生成建模。载于神经信息处理系统进展，第 2406–2422 页，2022年。

Eppstein, D.、Har-Peled, S. 和 Sidiropoulos, A. 面向图度量的近似贪心聚类与距离选择。计算几何学杂志，2020年。

Falorsi, L., de Haan, P., Davidson, T. R. 和 Forr´e, P. 李群上的分布重参数化。载于第二十二届国际人工智能与统计学会议论文集，第 3244–3253 页，2019。

Fan, J., Liu, S., Ma, S., Zhou, H.-M. 和 Chen, Y. 神经 Monge 映射估计及其应用。《机器学习研究汇刊》，2023年。国际标准连续出版物编号 2835-8856。

Feiten, W., Lang, M. 和 Hirche, S. 使用投影高斯混合进行刚体运动估计。载于第十六届国际信息融合会议论文集，第 1465–1472 页，2013年。

Figalli, A. 最优部分传输问题。《理性力学与分析档案》，195(2):533–560，2010。

Fukushima, K. 通过模拟阈值元件构成的多层网络进行视觉特征提取。*IEEE*《系统科学与控制论汇刊》，5(4):322–333，1969。

Genevay, A., Cuturi, M., Peyr´e, G., and Bach, F. 大规模最优传输的随机优化。载于神经信息处理系统进展，2016年。Geuter, J., Kornhardt, G., Tomasson, I., and Laschos, V. 通用神经最优传输。载于第四十二届国际机器学习会议，2025年。Gilbarg, D. and Trudinger, N. S. 二阶椭圆偏微分方程。Springer Berlin Heidelberg，2001。Gromov, M. 黎曼流形的填充。《微分几何学杂志》，18(1)，1983年。Hamelryck, T., Kent, J. T., and Krogh, A. 利用局部结构偏置采样逼真的蛋白质构象。《*PLoS* 计算生物学》，2(9)，2006年。He, K., Zhang, X., Ren, S., and Sun, J. 深入探究整流器：在 ImageNet 分类中超越人类水平的表现。载于*IEEE* 国际计算机视觉会议（*ICCV*）论文集，第1026–1034页，2015年。Huang, C.-W., Aghajohari, M., Bose, J., Panangaden, P., and Courville, A. C. 黎曼扩散模型。载于神经信息处理系统进展，第2750–2761页，2022年。Iacobelli, M. 黎曼流形上概率测度的渐近量化。*ESAIM*：控制、优化与变分法，22(3):770–785，2016年。Kantorovitch, L. 论质量转移。管理科学，5(1):1–4，1958年。Korotin, A., Selikhanovych, D., and Burnaev, E. 神经最优传输。载于第十一届国际学习表征会议，2023年。

Kratsios, A. and Bilokopytov, I. 非欧几里得通用逼近。载于神经信息处理系统进展，第10635–10646页，2020。

Kratsios, A. 和 Papon, L. 可微几何深度学习的通用逼近定理。机器学习研究期刊，23(1):1–73，2022年。

Lee, J. M. 光滑流形导论。数学研究生教材。施普林格，第二版，2012。

Lee, J. M. 黎曼流形导论。施普林格国际出版公司，2018年。

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. 具有非多项式激活函数的多层前馈网络可以逼近任意函数。《神经网络》，6(6):861–867，1993。

Lu, Z.、Pu, H.、Wang, F.、Hu, Z.与Wang, L. 神经网络的表达能力：从宽度的视角。载于 神经信息处理系统进展，2017年。Maas, A. L.、Hannun, A. Y.与Ng, A. Y. 整流非线性改进神经网络声学模型。载于 第*30*届国际机器学习会议论文集，第28卷，2013年。可获取于
h
t
t
p
s://
a
p
i.
semanticscholar.org/CorpusID:16489696。

Makkuva, A.、Taghvaei, A.、Oh, S.与Lee, J. 基于输入凸神经网络的最优传输映射。载于第*37*届国际机器学习会议论文集，第6672–6681页，2020年。Mardia, K. V.与Jupp, P. E. 方向统计学。威利概率与统计丛书。约翰·威利父子公司，1999年。Mardia, K. V.、Taylor, C. C.与Subramaniam, G. K. 蛋白质生物信息学与用于角度数据的二元冯·米塞斯分布混合。生物统计学，63(2):505–512，2006年。Mathieu, E.与Nickel, M. 黎曼连续归一化流。载于神经信息处理系统进展，第2503–2515页。Curran Associates, Inc.，2020年。McCann, R. 黎曼流形上映射的极分解。几何与泛函分析，11(3): 589–608，2001 年 8 月。Micchelli, C. A.、Xu, Y.与Zhang, H. 通用核。机器学习研究期刊，7(95):2651–2667，2006年。

M¨uller, R. D.、Cannon, J.、Qin, X.、Watson, R. J.、Gurnis, M.、Williams, S.、Pfaffelmoser, T.、Seton, M.、Russell, S. H. J.与Zahirovic, S. GPlates：穿越深时构建虚拟地球。《地球化学、地球物理学与地球系统》，19(7): 2243–2261，2018年7月。

Peel, D., Whiten, W. J., 和 McLachlan, G. J. 拟合 Kent 分布混合模型以辅助联合集合识别。《美国统计学会会刊》, 96(453): 56–63, 2001。

Petersen, P. 黎曼几何。数学研究生教材。施普林格，2016年。

Rezende, D. J. 和 Racani`ere, S. 隐式黎曼凹势函数映射，2021年。预印本。在线获取：https://arxiv.org/abs/2110.01288。

Rezende, D. J., Papamakarios, G., Racaniere, S., Albergo, M., Kanwar, G., Shanahan, P., 和 Cran mer, K. 环面和球面上的归一化流。发表于第*37*届国际机器学习会议论文集，第8083–8092页，2020年。

Rout, L.、Korotin, A. 与 Burnaev, E. 使用最优传输映射的生成建模。载于国际学习表征会议，2022年。

Rozen, N.、Grover, A.、Nickel, M. 与 Lipman, Y. 莫泽流：流形上的基于散度的生成式建模。载于神经信息处理系统进展，第17669–17680页，2021年。

Sakai, T. 黎曼几何。数学专著译丛。美国数学会，1996年。Schmidt-Hieber, J. 流形上函数的深度ReLU网络逼近，2019。预印本。可在线获取：https://arxiv.org/abs/1908.00695。Senanayake, R. 和 Ramos, F. 方向网格地图：动态环境中的多模态角度不确定性建模。载于 *2018*年 *IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第3241–3248页，2018年。Stern, A. l^p 流形上的变量替换不等式。《数学不等式与应用》，16(1):55–67，2013年。Teschl, G. 实分析专题。讲义，2025年。可在线获取：https://www.mat.univie.ac.at/~gerald/ftp/book-ra/ra.pdf。Uscidda, T. 和 Cuturi, M. Monge 间隙：一种用于学习所有传输映射的正则化项。载于 第*40*届国际机器学习会议论文集，第34709–34733页，2023年。

Vacher, A. 和 Vialard, F.-X. 半对偶非平衡二次最优传输：快速统计速率与收敛算法。载于 第*40*届国际机器学习会议论文集，第34734–34758页，2023年。

Villani, C. 最优传输。施普林格，2016年。

Williamson, D. P. 与 Shmoys, D. B. 近似算法的设计。剑桥大学出版社，2012。

Wilson, J. T. 大陆漂移。《科学美国人》，208(4)：86–103，1963年。

Yarotsky, D. 深层 ReLU 网络逼近的误差界。《神经网络》，94:103–114，2017年。Yarotsky, D. 和 Zhevnerchuk, A. 深度神经网络逼近率的相图。载于神经信息处理系统进展，第 13005–13015 页，2020年。Zahirovic, S., Eleish, A., Doss, S., Pall, J., Cannon, J., Pis- tone, M., Tetley, M. G., Young, A. 和 Fox, P. 俯冲与碳酸盐台地相互作用。《地球科学数据杂志》，9(2):371–383，2022年。Zhou, D.-X. 深度卷积神经网络的普适性。《应用与计算调和分析》，48(2):787–794，2020年。

A. 黎曼几何综述

本附录汇集了 d　　本文中使用的微分几何记号和事实。标准参考文献包括 deLee (2018年);Peters　(2016年)。

A.1. 光滑流形

固定 $p \in \mathbb{N}$。p维拓扑流形 是一个拓扑空间 $\mathcal{M}$，它是豪斯多夫且第二可数的，并且每个点 $x \in \mathcal{M}$都有一个与 $\mathbb{R}^p$的某个开子集同胚的邻域。等价地，$\mathcal{M}$ 具有一个图册：由若干对 $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in A}$ 组成的族，使得

$$U_\alpha \subset \mathcal{M} \text{ are open}, \qquad \bigcup_{\alpha \in A} U_\alpha = \mathcal{M}, \qquad \phi_\alpha : U_\alpha \to \phi_\alpha(U_\alpha) \subset \mathbb{R}^p \text{ are homeomorphisms.}$$

A 光滑（C^∞）流形 是配备了一个图册的拓扑流形，其过渡映射　　s

$$\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$$

均为 C^∞ 微分同胚；如果两个这样的图册的并仍然光滑，则称它们等价。

带边界流形 的定义类似，只需将 $\mathbb{R}^p$ 替换为半空间 $\mathbb{H}^p = \{(x_1, \dots, x_p) \in \mathbb{R}^p : \text{即图表范围内的 } x_p \geq 0\}$。除非另有说明，“流形”均指无边界流形。

切丛。对于 $x \in \mathcal{M}$，切空间 $T_x\mathcal{M}$ 是一个 p维实向量空间，由 x处的切向量组成。具体而言，若 $(U, x_1, \dots, x_p)$ 是 x 附近的坐标图，则切空间可识别为由 $\partial/\partial x_1, \dots, \partial/\partial x_p$ 在 x 处的偏导数张成的实向量空间。切丛定义为不交并 $T\mathcal{M} := \bigsqcup_{x \in \mathcal{M}} T_x\mathcal{M}$。在 $T\mathcal{M}$ 上存在一个典范的光滑流形结构，使自然投影 $\mathrm{pr} : T\mathcal{M} \to \mathcal{M}$ 成为光滑映射。具体来说，在 $\mathcal{M}$的坐标图 $(U, x_1, \dots, x_p)$ 下，每个切向量都可唯一写成

$$v = \sum_{i=1}^{p} v_i \frac{\partial}{\partial x_i},$$

从而得到 $T\mathcal{M}$ 的诱导坐标 $(\mathrm{pr}^{-1}(U)\ x_1,\ \dots,\ x_p,\ v_1,\ \dots,\ v_p)$。

A.2. 黎曼度量

$\mathcal{M}$ 上的黎曼度量是一个光滑变化的内积

$$g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}, \qquad x \in \mathcal{M}.$$

It 在每个 $T_x\mathcal{M}$ 上诱导一个范数 $\|v\|_x := \sqrt{g_x(v,\ v)}$。数对 $(\mathcal{M},\ g)$ 是一个黎曼流形　　.

在所有附录中，只要含义明确，我们都省略下标，并将 $v \in T_x\mathcal{M}$ 的范数简单记为 $\|v\|$。

给定一条分段 C^1 曲线 $\gamma : [0, 1] \to \mathcal{M}$，其长度定义为

$$L(\gamma) := \int_0^1 \|\dot\gamma(t)\|\, dt.$$

黎曼距离是诱导的长度度量

$$d(x, y) := \inf\{L(\gamma) : \gamma(0) = x,\ \gamma(1) = y\}.$$

这将　　$\mathcal{M}$ 变为度量空间（$\mathcal{M},\ d$）。若 $\mathcal{M}$ 是紧致的，则 $\mathrm{diam}(\mathcal{M}) < \infty$，且（$\mathcal{M},\ d$）是完备的 .

列维-奇维塔联络。黎曼度量决定了 $T\mathcal{M}$ 上唯一的仿射联络 ∇，该联络无挠且与度量相容，称为列维-奇维塔联络。对于光滑函数 $f \in C^\infty(\mathcal{M})$，（黎曼）梯度$\nabla f$是由 $g(\nabla f, X) = df(X)$对所有向量场 X所定义的向量场。

子流形与法丛。设$(\mathcal{M}, g)$为黎曼流形，并设 $\mathcal{V} \subseteq \mathcal{M}$ 为光滑子流形。在每个 $x \in \mathcal{V}$处，切空间 $T_x\mathcal{V}$ 是 $T_x\mathcal{M}$的线性子空间。由于 g_x 是 $T_x\mathcal{M}$上的内积，存在正交分解

$$T_x\mathcal{M} = T_x\mathcal{V} \bigoplus (T_x\mathcal{V})^{\perp},$$

其中$(T_x\mathcal{V})^{\perp}$ 是 $T_x\mathcal{V}$ 在 $T_x\mathcal{M}$中的正交补。空间$(T_x\mathcal{V})^{\perp}$ 称为 $\mathcal{V}$ 在 x处的法空间，记为 $N_x\mathcal{V}$。法丛定义为不交并 $N\mathcal{V} := \bigsqcup_{x\in\mathcal{V}} N_x\mathcal{V}$。 $N\mathcal{V}$ 上存在典范的光滑流形结构，使得自然投影$\mathrm{pr}: N\mathcal{V} \to \mathcal{V}$ 成为光滑映射。

A.3. 测地线与指数映射

一条光滑曲线 $\gamma: [0,1] \to \mathcal{M}$ 是（仿射参数化的）测地线，如果它满足测地线方程 n

$$\nabla_{\dot\gamma}\dot\gamma = 0.$$

给定 $x \in \mathcal{M}$ 和 $v \in T_x\mathcal{M}$，存在唯一一条测地线 $\gamma_{x,v}: (-\varepsilon, \varepsilon) \to \mathcal{M}$ 满足 $\gamma_{x,v}(0) = x$ 且 $\dot\gamma_{x,v}(0) = v$。

E指数映射。 当$\gamma_{x,v}$ 在时刻 $t = 1$有定义时，位于 x 的黎曼指数映射 是 s

$$\exp_x(v) := \gamma_{x,v}(1).$$

由于正规邻域的存在，指数映射总是在局部良定义。

命题 A.1（正规邻域）。 对于任意 $x \in \mathcal{M}$，存在 $V \subset T_x\mathcal{M}$ 的邻域和 x的 $U \subset \mathcal{M}$ 邻域，使得 $\exp_x: V \to U$ 是一个 C^∞-微分同胚。

U 上的逆映射是对数映射$\log_x: U \to T_x\mathcal{M}$，定义为 $\log_x(y) = (\exp_x)^{-1}(y)$。

内射半径。位于内射半径于 x是

$$\mathrm{inj}(x) := \sup\big\{r > 0 : \exp_x \text{ is a diffeomorphism on } B_{T_x\mathcal{M}}(0, r)\big\},$$

其中 $B_{T_x\mathcal{M}}(0, r) = \{v \in T_x\mathcal{M} : \|v\|_x < r\}$。若 $\mathcal{M}$ 紧致，则 $\mathrm{inj}(\mathcal{M}) := \inf_{x\in\mathcal{M}} \mathrm{inj}(x) > 0$。

完备性。 我们始终假设 $\mathcal{M}$是连通且完备的。以下定理至关重要。**定理 A.2**（Hopf–Rinow）。 设 $(\mathcal{M}, g)$ 是一个连通黎曼流形。以下条件等价 *nt:*

1. （$\mathcal{M}, d$）是一个完备度量空间；*2.* （$\mathcal{M}, g$）是测地完备的，即对于每个 $x \in \mathcal{M}$，$\exp_x$ 在整个 $T_x\mathcal{M}$ 上都有定义；*3.* （$\mathcal{M}, d$）的闭有界子集都是紧致的。

特别地，任何紧黎曼流形都是完备的，且任意两点之间都可以由一条最短测地线连接。

A.4. 割迹

对于 $x \in \mathcal{M}$，将 $T_x\mathcal{M}$ 中的单位球面定义为

$$U_x\mathcal{M} := \{u \in T_x\mathcal{M} : \|u\|_x = 1\}, \qquad U\mathcal{M} := \bigsqcup_{x\in\mathcal{M}} U_x\mathcal{M}.$$

集合 $U\mathcal{M}$是 $T\mathcal{M}$的余维为 1(参见 Sakai (1996年，第 III 章)。

定义 A.3（割时与割迹（Sakai，1996年，定义 4.3））。假设 $(\mathcal{M}, g)$ 是完备的，并固定 $x \in \mathcal{M}$。对于$u \in U_x\mathcal{M}$，令 $\gamma_u(t) := \exp_x(tu)$ 表示满足 $\gamma_u(0) = x$ 的单位速度测地线。将割时 $t(u) \in (0, \infty]$ 定义为

$$t(u) := \sup\Big\{t > 0 : \ d\big(x, \gamma_u(s)\big) = s \text{ for all } s \in [0, t]\Big\}.$$

如果 $t(u) < \infty$，则向量 $t(u)u \in T_x\mathcal{M}$ 是沿 γ_u 的切空间割点，而 $\exp_x(t(u)u)$ 是相应的割点。x 的切空间割迹和割迹为

$$\widetilde{\mathrm{Cut}}(x) := \{t(u)u : u \in U_x\mathcal{M},\ t(u) < \infty\}, \qquad \mathrm{Cut}(x) := \exp_x\big(\widetilde{\mathrm{Cut}}(x)\big).$$

另定义 $T_x\mathcal{M}$ 中的内部区域 及其像：

$$\widetilde{\mathcal{I}}_x := \{tu : u \in U_x\mathcal{M},\ 0 < t < t(u)\}, \qquad \mathcal{I}_x := \exp_x(\widetilde{\mathcal{I}}_x).$$

关于切空间割点，我们有以下刻画。

L引理 A.4. *(Petersen，2016年，引理 5.7.9）若 $u \in T_x\mathcal{M}$ 是 x的切空间割点，则以下情况之一成立 r*

1. 存在 $w \neq u \in T_x\mathcal{M}$，使得 $\exp_x(u) = \exp_x(w)$，或者
2. 微分 $d(\exp_x)$在 u处是奇异的。

在第二种情形下，引理 A.4 中的点 $\exp_x(u) \in$ 割迹(x)也称为第一个共轭点。

下面总结 割迹(x)(参见Sakai (1996年年，引理 4.4))。

引理 A.5 (基本 性质：割迹(x) (Sakai，1996年，引理 4.4))。 设 $(\mathcal{M}, g)$ 是完备的，且 $x \in \mathcal{M}$。则：

1. $\mathcal{I}_x \cap$割迹$(x) = \varnothing$，$\mathcal{M} = \mathcal{I}_x \cup$割迹$(x)$，以及 $\overline{\mathcal{I}}_x = \mathcal{M}$;
2. $\mathcal{I}_x$ 是包含 $0 \in T_x\mathcal{M}$ 的一个极大定义域，在该定义域上 $\exp_x$ 是微分同胚;
3. 割迹(x) 是 $\mathrm{vol}_{\mathcal{M}}$-零测的，且 dim 割迹$(x) \leq p - 1$，因此

$$\mathrm{vol}_{\mathcal{M}}\big(\mathrm{Cut}(x)\big) = 0, \qquad \forall x \in \mathcal{M}.$$

平方距离函数。 固定 $y \in \mathcal{M}$。函数 $x \mapsto \frac{1}{2}d(x,y)^2$在 $\mathcal{M} \setminus$Cut(y) 上是 C^∞。此外，对于 $x \notin$Cut(y)，对数映射 $\log_x(y)$ 有定义，且

$$\nabla_x\Big(\tfrac{1}{2}d(x,y)^2\Big) = -\log_x(y). \tag{8}$$

类似地，对于 $y \notin$割迹(x)，有 $\nabla_y(\frac{1}{2}d(x, y)^2) = -\log_y(x)$。这些恒等式在对涉及二次代价的最优传输目标函数求导时经常使用。

B. 黎曼最优传输综述

B.1. 黎曼流形上的表述

我们简要回顾紧黎曼流形上的最优传输（最优传输）形式体系（参见 Ambrosio 等人 (2024，第 7.3 章）。设 $(\mathcal{M},\ g)$ 是一个光滑、连通的 n 维紧黎曼流形，且无边界，并设 d 表示其黎曼距离。我们将 $\mathcal{B}(\mathcal{M})$ 记作 σ- 代数上的博雷尔 $\mathcal{M}$，并将 $\mathrm{vol}_{\mathcal{M}} \in \mathcal{M}_+(\mathcal{M})$ 记作黎曼体积测度，其中 $\mathcal{M}_+(\mathcal{M})$ 表示 $\mathcal{M}$上的有限非负博雷尔测度集合。令

$$\mathcal{P}(\mathcal{M}) := \{\mu \in \mathcal{M}_+(\mathcal{M}) : \mu(\mathcal{M}) = 1\}$$

为 $\mathcal{M}$ 上博雷尔概率测度的集合。给定博雷尔映射 $T\colon\ \mathcal{M} \to \mathcal{M}$ 和 $\mu \in \mathcal{P}(\mathcal{M})$，推前（像）测度 $T_{\#}\mu \in \mathcal{P}(\mathcal{M})$定义为

$$T_{\#}\mu(A) := \mu(T^{-1}(A)), \qquad \forall A \in \mathcal{B}(\mathcal{M}).$$

2-瓦瑟斯坦距离。对于 $\mu,\ \nu \in \mathcal{P}(\mathcal{M})$，我们用 $\Pi(\mu,\nu)$ 表示耦合（或传输计划）的集合，即边缘分布为 μ 和 ν 的概率测度 $\pi \in \mathcal{P}(\mathcal{M}\times\mathcal{M})$:

$$(\mathrm{pr}_1)_{\#}\pi = \mu, \qquad (\mathrm{pr}_2)_{\#}\pi = \nu,$$

其中 pr_i 为 $\mathcal{M}\times\mathcal{M}$上的坐标投影。$W_2$ 上的 2-瓦瑟斯坦距离 $\mathcal{P}(\mathcal{M})$ 定义为 y

$$W_2(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\mathcal{M}\times\mathcal{M}} d^2(x,y)\,\mathrm{d}\pi(x,y) \right)^{\frac{1}{2}}.$$

由于 $\mathcal{M}$ 是紧致的，下确界有限，且至少由一个最优计划 $\pi^\star \in \Pi(\mu,\ \nu)$ 取得。

Monge 与 Kantorovich 表述。在全文中，我们考虑二次代价

$$c(x,y) := \tfrac{1}{2}d^2(x,y).$$

相应的*Kantorovich*问题为

$$\mathrm{KP}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} c(x,y)\,\mathrm{d}\pi(x,y) = \tfrac{1}{2}\,W_2^2(\mu,\nu). \tag{9}$$

蒙日问题是将（9）限制在由映射诱导的耦合上:

$$\mathrm{MP}(\mu,\nu) := \inf_{T_{\#}\mu=\nu} \int_{\mathcal{M}} c\big(x,T(x)\big)\,\mathrm{d}\mu(x). \tag{10}$$

一般而言，MP 可能不存在极小化元（甚至可能定义不适定），而 KP 在紧致度量空间上总是至少存在一个最优计划。

B.2. 对偶性、 c-变换与 c-凹性

Kantorovich问题（9）的对偶表述为

$$\mathrm{KP}(\mu,\nu) = \sup_{\substack{\varphi,\psi\in C(\mathcal{M},\mathbb{R})\\ \varphi(x)+\psi(y)\le c(x,y)}} \left\{ \int_{\mathcal{M}} \varphi(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{M}} \psi(y)\,\mathrm{d}\nu(y) \right\}. \tag{11}$$

给定 $\psi\colon\ \mathcal{M} \to \mathbb{R}\cup\{-\infty\}$，其 c-变换定义为

$$\psi^c(x) := \inf_{y\in\mathcal{M}} \big(c(x,y) - \psi(y)\big). \tag{12}$$

按照这一约定，公式（11）可以写成以下半对偶形式

$$\mathrm{KP}(\mu,\nu) = \sup_{\psi\in C(\mathcal{M})} \left\{ \int_{\mathcal{M}} \psi^c(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{M}} \psi(y)\,\mathrm{d}\nu(y) \right\}. \tag{13}$$

备注B.1（符号约定）。许多流形上的最优传输构造（以及若干学习目标）倾向于采用等价的重新参数化 $\tilde{\psi} := -\psi$，在这种情况下

$$\tilde{\psi}^c(x) = \inf_{y \in \mathcal{M}} \big(c(x,y) + \tilde{\psi}(y)\big), \qquad \mathrm{KP}(\mu,\nu) = \sup_{\tilde{\psi} \in C(\mathcal{M},\mathbb{R})} \left\{ \int \tilde{\psi}^c \, \mathrm{d}\mu - \int \tilde{\psi} \, \mathrm{d}\nu \right\}. \tag{14}$$

两种约定只需作替换 $\tilde{\psi} = -\psi$即可等价。

称函数 $\varphi\colon \mathcal{M} \to \mathbb{R} \cup \{-\infty\}$ 为 c-凹函数，若 $\varphi = \psi^c$ 对某个 $\psi\colon \mathcal{M} \to \mathbb{R} \cup \{-\infty\}$成立。已知，$\varphi$ 是c-凹的，当且仅当 $\varphi = \varphi^{cc}$，其中 $\varphi^{cc} := (\varphi^c)^c$。对于定义在 $\mathcal{M}$上的二次代价 $c(x,y) = \frac{1}{2}d^2(x,y)$，半对偶形式 (13) 中的上确界总能在某个 c-凹函数处取得。

B.3. 最优性条件与麦肯定理

定义 B.2 （c-次微分）。设 $\varphi\colon \mathcal{M} \to \mathbb{R} \cup \{-\infty\}$ 且 $x \in \{\varphi > -\infty\}$。$\varphi$ 在 x 处的 c-次微分为

$$\partial^c \varphi(x) := \Big\{ y \in \mathcal{M} : \ \varphi(x) + \varphi^c(y) = c(x,y) \Big\}.$$

等价地，若 $\varphi = \psi^c$ 如式（12）所示，则

$$y \in \partial^c \varphi(x) \quad \Longleftrightarrow \quad y \in \arg\min_{z \in \mathcal{M}} \big(c(x,z) - \psi(z) \big).$$

在另一种符号约定（14）下，同一集合可表示为 $\arg\min_z(c(x,z) + \tilde{\psi}(z))$。

若 (φ, ψ) 是（11）中的一个最优对（其中 $\varphi = \psi^c$），且 $\pi^\star$ 是（9）中的一个最优计划，则

$$\varphi(x) + \psi(y) = c(x,y) \qquad \pi^\star\text{-a.e. on } \mathcal{M} \times \mathcal{M}.$$

特别地，$\mathrm{spt}(\pi^\star) \subset \{(x,\ y) : \ y \in \partial^c \varphi(x)\}$。

黎曼流形上二次代价的一个核心事实是，在温和条件下，最优耦合由传输映射诱导，这一结果称为 McCann 定理。

定理 B.3 ($c = \frac{1}{2}d^2$ ((McCann，2001)))。设 μ，$\nu \in \mathcal{P}(\mathcal{M})$，并假设$\mu$ 相对于 $\mathrm{vol}_{\mathcal{M}}$绝对连续。则存在一个 c-凹势函数 φ，使得（(9)）的最优计划由一个（本质上唯一的）映射 $T\colon \mathcal{M} \to \mathcal{M}$所诱导，且该映射满足

$$T(x) \in \partial^c \varphi(x) \qquad \textit{for } \mu\textit{-a.e. } x \in \mathcal{M}.$$

此外，φ 在 μ-几乎处处可微，并且在可微点处，

$$T(x) = \exp_x\big(-\nabla \varphi(x)\big), \tag{15}$$

其中，∇ 表示黎曼梯度，而$\exp_x$： $T_x\mathcal{M} \to \mathcal{M}$ 是指数映射。

注记 B.4（割迹与光滑性）。函数 $(x,\ y) \mapsto \frac{1}{2}d^2$（$x,\ y$）在割迹之外是光滑的，但通常在割迹上不光滑。因此，φ 以及映射公式（15）应按几乎处处的意义理解，而许多定量结论自然仅适用于避开割迹邻域的区域。

C. H¨older 空间综述

多重指标记法。 设 $n \in \mathbb{N}$。对于多重指标 $J = (j_1, \ldots\ldots, j_n) \in \mathbb{N}_0^n$，定义

$$|J| := j_1 + \cdots + j_n, \qquad D^J := \frac{\partial^{|J|}}{\partial x_1^{j_1} \cdots \partial x_n^{j_n}}.$$

我们用 $\|x\|$表示 $x \in \mathbb{R}^n$的欧几里得范数。

C.1. 欧几里得区域上的 H¨older 空间

设 $k \in \mathbb{N}_0, \alpha \in (0, 1]$, 和 $K \subset \mathbb{R}^n$ 为一个紧集。假设 $f\colon K \to \mathbb{R}$ 是一个函数，使得所有满足 $|J| \le k$ 的导数 $D^J f$均存在，并连续延拓到 K的边界。定义半范数

$$[f]_{C^{k,\alpha}(K)} := \sum_{|J|=k} \sup_{\substack{x,y \in K^\circ \\ x \neq y}} \frac{|D^J f(x) - D^J f(y)|}{\|x - y\|^\alpha},$$

以及 H¨older 范数

$$\|f\|_{C^{k,\alpha}(K)} := \sum_{j=1}^{k} \sup_{|J|=j} \|D^J f\|_\infty \; + \; [f]_{C^{k,\alpha}(K)}.$$

H¨older 空间 $C^{k,\alpha}(K)$ 定义为所有 H¨older 范数有限的此类函数所构成的集合 .

在证明中，我们通常取 $K = [0, 1]^n$。对于 $r = k + \alpha$，记

$$\mathcal{H}_{r,n} := \Big\{ f \in C^{k,\alpha}([0,1]^n) : \|f\|_{C^{k,\alpha}([0,1]^n)} \le 1 \Big\}. \tag{16}$$

H¨older 空间 $C^{k,\alpha}([0, 1]^n)$ 中的单位球。

C.2. 光滑流形上的 H¨older 空间

设 $\mathcal{M}$ 是一个紧致的 p 维光滑流形，并固定 $k \in \mathbb{N}_0, \alpha \in (0, 1]$。选取一个有限光滑图册 $\{(U_i, \varphi_i)\}_{i=1}^N$，使得每个 $\varphi_i\colon U_i \to V_i \subset \mathbb{R}^p$ 都是到某个开集 V_i 上的 C^∞- 微分同胚。设 $\{\chi_i\}_{i=1}^N$ 是从属于 $\{U_i\}_{i=1}^N$ 的光滑单位分解:

$$\chi_i \in C^\infty(\mathcal{M}), \quad 0 \le \chi_i \le 1, \quad \operatorname{supp}(\chi_i) \subset U_i, \quad \sum_{i=1}^{N} \chi_i \equiv 1 \text{ on } \mathcal{M}.$$

对于 $f\colon \mathcal{M} \to \mathbb{R}$，定义

$$f_i := (\chi_i f) \circ \varphi_i^{-1} \quad \text{on } V_i, \qquad K_i := \varphi_i(\operatorname{supp} \chi_i) \Subset V_i.$$

我们称 $f \in C^{k,\alpha}(\mathcal{M})$, 如果对所有 i，都有 $f_i \in C^{k,\alpha}(K_i)$, 并且

$$\|f\|_{C^{k,\alpha}(\mathcal{M})} := \max_{1 \le i \le N} \|f_i\|_{C^{k,\alpha}(K_i)} < \infty.$$

有限图册和从属单位分解的不同选择会给出等价的范数。特别地，H¨older 空间$C^{k,\alpha}(\mathcal{M})$ 在范数等价的意义下定义良好。一个常用的替代刻画是: $f \in C^{k,\alpha}(\mathcal{M})$ 当且仅当对于每个坐标图 (U, φ) 和每个紧致 $K \Subset \varphi(U)$，坐标表示 $f \circ \varphi^{-1}$ 属于 $C^{k,\alpha}(K)$。

C.3. $C^{k,\alpha}$ 区域与延拓

我们回顾椭圆型偏微分方程理论中使用的边界正则性概念，以及一个基本的延拓引理（Gilbarg& Trudinger，2001，第 6.2 节）。

定义 C.1 ($C^{k,\,\alpha}$ 区域)。 设 $\Omega \subset \mathbb{R}^n$ 为有界开集，且 $k \ge 1, \alpha \in [0, 1]$。若对每个 $x_0 \in \partial\Omega$，均存在一个半径 $r > 0$ 和一个 $C^{k,\,\alpha}$ 微分同胚，则称 Ω 为 $C^{k,\,\alpha}$区域

$$\psi : B(x_0, r) \to D \subset \mathbb{R}^n, \qquad \psi,\, \psi^{-1} \in C^{k,\alpha},$$

使得

$$\psi\big(B(x_0,r)\cap\Omega\big)\subset\mathbb{R}^n_+,\qquad \psi\big(B(x_0,r)\cap\partial\Omega\big)\subset\partial\mathbb{R}^n_+,$$

其中 $\mathbb{R}^n_+ := \{\ x\in\mathbb{R}^n:\ \ x_n>0\}$且 $\partial\mathbb{R}^n_+ :=\{x_n=0\}$。等价地，$\psi$ 拉直边界附近 x_0.

若 $T\subset\partial\Omega$，我们称 T 为一个 $C^{k,\alpha}$ 边界部分，如果上述性质在每个 $x_0\in T$处成立，并附加要求 $B(x_0,\ r)\cap\partial\Omega\subset T$。

引理 C.2 （从 $C^{k,\alpha}$ 区域的延拓；引理 6.7（Gilbarg & Trudinger，2001））。设 $\Omega\subset\mathbb{R}^n$ 为有界 $C^{k,\alpha}$ 区域，具有 $k\geq 1$，并设 Ω' 为一个开集，满足 $\Omega\subset\Omega'$。若 $u\in C^{k,\alpha}(\Omega)$，则存在 $w\in C^{k,\alpha}(\Omega')$ 使得$w=u$在Ω上且

$$\|w\|_{C^{k,\alpha}(\Omega')}\leq C\,\|u\|_{C^{k,\alpha}(\overline{\Omega})},$$

其中 C 仅依赖于 k、α、Ω和Ω'。

备注 C.3（关于 Ω 上的范数）。当我们写作 $C^{k,\alpha}(\overline{\Omega})$ 时，意指 u 具有最高至 k 阶的导数，这些导数可连续延拓至 Ω，并且在 Ω 上满足阶数为 α 的 H¨older 条件；其范数定义如上文的紧集定义所示（取 $K=\Omega$）。

D. 相关工作

欧几里得空间中的神经最优传输。大量研究利用传输映射或康托罗维奇势的神经参数化，探索欧几里得空间中概率分布之间最优传输（OT）的学习。对于二次代价函数 $c(x,y) = \frac{1}{2}\|x-y\|^2$，一种具有代表性的方法是学习凸康托罗维奇势，并通过其梯度恢复蒙日映射；Makkuva 等人¯（2020年）采用输入凸神经网络（ICNN）和极小极大形式来实现这一策略。除了 Brenier/ICNN 参数化之外，一种实现可扩展性的标准方法是采用（熵正则化）OT 的半对偶形式，其中一个对偶变量经过优化，另一个则以闭式形式恢复；由此可得到基于样本的随机优化方案，参见例如 Genevay 等人（2016年）；Cuturi & Peyr´e（2018年）；Vacher & Vialard（2023年）。较新的"连续"神经 OT 求解器直接学习对偶变量和/或传输计划，并支持对耦合和映射进行样本外评估；参见例如 Korotin 等人（2023年）。神经方法还利用与半对偶训练密切相关的单势目标函数，直接从样本中学习蒙日映射；参见例如 Rout 等人（2022年）；Fan 等人（2023年）。

流形上的最优传输与生成建模。从理论角度来看，带二次代价的黎曼流形上的最优传输已有充分研究：在温和的假设条件下（例如 $\mu \ll$ 体积$_{\mathcal{M}}$），最优计划会集中于一个（几乎处处唯一的）映射，该映射可通过指数映射和一个 c-凹势函数表示（例如参见McCann(2001)； Villani (2016年)； Ambrosio 等人 (2024)）。在机器学习中，流形值生成建模采用了黎曼归一化流和流形上的连续时间动力学等方法（例如Mathieu 和 Nickel(2020年)），也采用了针对特定紧致流形（如球面和环面）设计的构造（例如Rezende 等人 (2020年)。

这与度量空间中经典的到地标点的距离嵌入理念密切相关，即用点到一组参考点/集合的距离来表示该点；本文采用的到地标点的距离嵌入见命题2.3（根据Gromov(1983)）。在我们的设定中，这种基于距离的特征映射发挥着关键的解析作用：它们为 $\mathcal{M}$ 提供单射且正则的欧几里得表示，从而能够引入定量逼近速率。

诸如 Schr¨odinger 桥之类的随机方法，为最优传输提供了熵正则化松弛，将最优传输与扩散过程联系起来（De Bortoli 等人，2022年）。这些方法得到的是传输计划（或随机流），而非确定性的蒙日映射，因此无法直接保证基于映射的传输，除非在小噪声极限下。另一项相关研究提出了 Monge 间隙的概念（Uscidda 和 Cuturi，2023年），并将其用作正则化项，以促使学习得到的传输映射满足蒙日型最优性。

与本文直接相关的是基于 c-凹势函数的受最优传输启发的流形流。 Cohen 等人（2021年）提出了 RCPM，这是一类构建于紧黎曼流形上的表达能力丰富的微分同胚，由离散 c-凹势函数构成；所得映射通过将指数映射与学习得到的势函数的（次）梯度复合而得。在此观点基础上，Rezende 和 Racani`ere （2021年）提出了隐式黎曼凹势映射（IRCPM），通过允许更一般的 $c$-凹势函数，并通过求解内层最小化问题（以及隐式微分）隐式地求得映射，从而扩展了 RCPM，实现高效的基于似然的训练和对称性处理。具体而言，据我们所知，Rezende 和 Racani`ere （2021年）是最早将这一受最优传输启发的内层最小化明确实现为流形流模型中的隐式层，并通过隐式微分对极小值点进行微分的研究者之一。

流形上的通用逼近与维数灾难考量。 我们的逼近论证结合了两个主题：(i) 通过单射／足够正则的特征映射，将流形上的逼近约化为欧几里得空间中的逼近；(ii) 引入欧几里得逼近理论中的定量近似率。在普适性方面，Kratsios & Bilokopytov (2020年) 给出了一般条件，在这些条件下，将欧几里得通用逼近器与适当的特征映射／读出映射复合，即可实现非欧几里得学习问题（包括流形值情形）的普适性。 Kratsios& Papon(2022年) 从定量与可微几何视角研究了几何约束下的流形逼近，并指出了数据依赖条件能够减轻或避免经典维数灾难的情形。在欧几里得逼近方面， Yarotsky& Zhevnerchuk (2020年) 证明了某些架构／激活函数具有快速（"无维数灾难"）逼近速率，包括适当的周期性／结构化激活函数所实现的近乎指数速率。相比之下，针对流形上一般 ReLU 网络的逼近结果，通常给出由内在维度决定的（多项式级）ε^{-1}速率，至多相差对数因子；例如参见 Schmidt-Hieber (2019)； Chen 等人 (2019)。我们的函数逼近定理与之互为补充：它针对由平方距离最优传输产生的特定几何函数类（特别是 c-凹势函数），并结合全局正则特征映射约化与欧几里得无维数灾难的逼近情形；当存在此类欧几里得速率时，由此得到不同的复杂度标度。

我们贡献的定位。 上述研究启发了我们的研究重点：我们研究具有二次代价的紧黎曼流形上的最优传输势函数和蒙日映射，并针对几何最优传输结构提出逼近保证。与欧氏神经最优传输不同，我们的分析必须明确考虑几何奇异性（例如割迹），因为在这些位置，平方距离及其相应的 c-变换会失去光滑性。与强调似然训练和经验表达能力的流形流构造不同，我们利用 RNOT 参数化，为 c-凹最优传输势函数（及其诱导的传输映射）提供定量逼近理论；一旦具备适当的特征映射／嵌入机制，其速率便与无维数灾难的欧几里得逼近相当。

E. 实现细节

E.1. 格罗莫夫嵌入实现

地标选择。 我们的嵌入层通过每个点 $x \in \mathcal{M}$到一组地标点 $L = \{\ell_j\}_{j=1}^{M} \subset \mathcal{M}$的距离来表示它，

$$\varphi(x) := (d(x, \ell_j))_{j=1}^{M} \in \mathbb{R}^M.$$

我们考虑两种实用策略来选择 L:

(i) 随机地标点 *(RND)*：我们从 $\mathcal{M}$ 上的参考测度中独立同分布地采样 ℓ_j （在我们的实验中，使用 $\mathcal{M}$上的均匀测度）。

(ii) 最远点采样（*FPS*）：为了获得空间分布更均匀的地标集，我们在候选池 $C = \{c_n\}_{n=1}^{N} \subset \mathcal{M}$ （从同一参考测度中采样或取自训练数据）上执行贪心 k-中心过程。

FPS 以一个随机种子 $\ell_1 \in C$ 初始化，然后迭代选择

$$\ell_{t+1} := \arg\max_{c \in C} \min_{1 \le j \le t} d(c, \ell_j), \qquad t = 1, \dots, M-1,$$

即，每个新地标都是与当前地标集的测地距离最远的候选点。为了量化空间填充特性，我们将地标集 L 相对于 C 的（经验）覆盖半径定义为

$$R(L; C) := \max_{c \in C} \min_{\ell \in L} d(c, \ell),$$

使得 r 满足$C \subset \bigcup_{\ell \in L} B$（$\ell$， r）的最小半径。离散的 k-中心目标是最小化 R（L； C），遍历所有$|L| = M$ 的子集 C。FPS 是该目标的标准贪心近似算法（通常称为最远优先遍历），倾向于生成分布均匀且覆盖半径较小的地标点（例如参见Williamson & Shmoys（2012）；Eppstein 等人（2020年））。此外，它还能生成嵌套的地标集：前 M 个地标是针对任意更大的$M' > M$所获得结果的前缀。

嵌入诊断与 M 的选择。 距离坐标嵌入的理论单射条件取决于几何常数，而这些常数通常无法以闭式形式获得。因此，我们使用留出集 $X_{\text{val}} \subset \mathcal{M}$上的经验诊断指标来选择 M。对于每个候选项 M，我们计算 $x_n \in X_{\text{val}}$ 的嵌入 $z_n = \varphi(x_n)$，并监测嵌入空间中的非塌缩 统计量。具体而言，我们估计最小两两间距

$$s_M := \min_{a \ne b} \|z_a - z_b\|_2,$$

以及近碰撞的比例

$$\rho_M(\varepsilon) := \frac{1}{|P|} \sum_{(a,b) \in P} \mathbb{1}_{\{\|z_a - z_b\|_2 < \varepsilon\}},$$

其中， P是从 X_{val} 中采样得到的一组不同的点对（为提高效率，我们对点对进行子采样）。等价地， $\rho_M(\varepsilon)$是在用于抽取点对（x， y）的采样方案下，对 $\Pr(\|\varphi_M(x) - \varphi_M(y)\|_2 < \varepsilon)$ 的经验估计。

嵌入诊断与 M 的选择。 距离坐标嵌入的理论单射条件取决于几何常数，而这些常数通常无法以闭式形式获得。因此，我们使用留出验证集 $X_{\text{val}} \subset \mathcal{M}$上的经验诊断指标来选择 M。对于每个候选项 M，我们计算 $\in \mathbb{R}^M$ 中 $x_n \in X_{\text{val}}$的嵌入 $z_n = \varphi(x_n)$，其中 $\varphi(x) = (d(x, \ell_1), \dots\dots, d(x, \ell_M))$ 是地标距离嵌入。我们监测嵌入空间中的非塌缩统计量：最小两两间距 s_M 和近碰撞比例 $\rho_M(\varepsilon)$。此外，我们还报告经验覆盖半径

$$R_M^{\text{val}} := \max_{x \in X_{\text{val}}} \min_{1 \le j \le M} d(x, \ell_j),$$

用于衡量地标集对感兴趣区域的覆盖程度。

实践中，我们会在多个 X_{val} 随机子集上重复这些估计以降低方差，并选择最小的 M，使得 s_M 稳定地高于数值容差，且 $\rho_M(\varepsilon) \approx 0$；这为数据分辨率下的单射性提供了一种实用的替代指标。图4 对比了 RAND 地标点采样与最远点采样，结果表明，使用更少的地标点，最远点采样即可实现更低的覆盖半径和相当的分离度（见图4）。

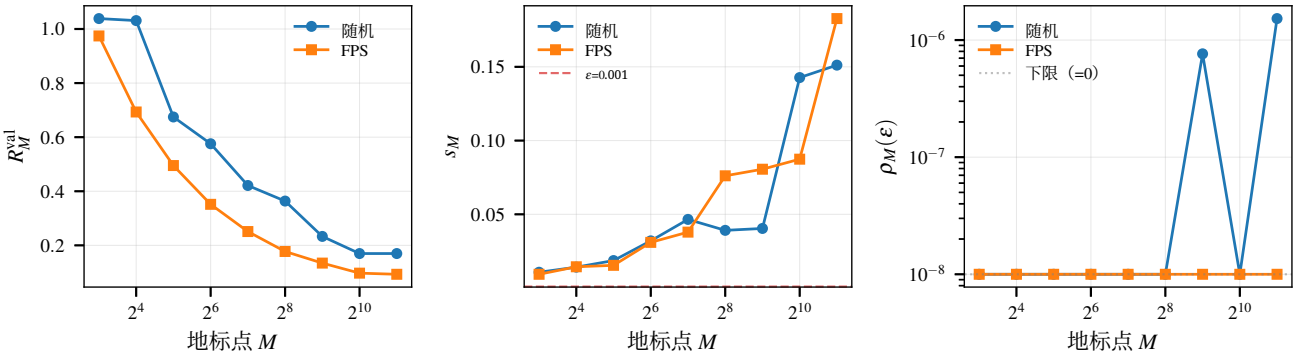


图 4. 在 $\mathbb{S}^2$ 上的嵌入诊断：覆盖半径 R_M^{val}（左）、最小两两间距 s_M（中）以及近碰撞比例 $\rho_M(\varepsilon)$（右），比较随机采样与最远点采样（FPS）。FPS 在保持分离度相近的同时实现了更低的覆盖半径，且所有 M 测试均满足 $\rho_M \approx 0$。

E.2. 通过 Kantorovich 半对偶训练 RNOT 映射

现在我们将介绍 RNOT 映射在实践中如何训练。本节详细说明了用于对流形上的预势函数进行参数化的特征表示、优化目标，以及 c-变换的数值处理方法。我们提出的流程仅使用来自源分布和目标分布的样本，即可对 RNOT 映射进行端到端训练，无需离散化流形，也无需计算雅可比行列式。此处详述的实现与先前的理论相一致，并旨在高维场景下保持稳定且可扩展。

我们首先介绍用于表示 $\mathcal{M}$ 上函数的特征映射。

到地标距离特征表示。我们采用一种受命题2.3启发的距离坐标特征映射。原则上，该结果保证存在一个几何阈值 $\delta_0 > 0$，使得对于任意 $\delta \in (0, \delta_0)$ 以及任意极大 δ-分离集 $S = \{x_i\}_{i=1}^I \subset \mathcal{M}$，映射

$$\varphi_S(x) \;=\; \big(d(x, x_i)\big)_{i=1}^I$$

是单射，因此在紧致流形 $\mathcal{M}$上是一个拓扑嵌入。由于 δ_0 取决于 $\mathcal{M}$ 的几何（通常未知），且没有闭式表达，我们不会尝试显式构造这样的 S。

相反，我们通过随机采样或最远点采样（附录E.1）构造一个（规模更大的）地标集 $L = \{\ell_j\}_{j=1}^M$

$$\varphi(x) \;:=\; \big(d(x, \ell_j)\big)_{j=1}^M \in \mathbb{R}^M. \tag{17}$$

该特征映射是连续的，增大 M 不会破坏单射性：如果 L 包含一个单射子集（如命题2.3所示），那么 φ 也同样是单射的。实践中，我们将 M 视为可调的分辨率参数，并逐步增大它，直到该表示在留出样本上经验性地不发生坍缩。重要的是，我们强调，这些地标点仅用于构造定义在 $\mathcal{M}$上的函数的连续特征表示；它们不会离散化传输问题本身，所得的 RNOT 映射仍然完全连续，且输出支撑集不受限制。具体而言，我们在验证集上监测附录 E.1 所述的嵌入诊断指标，并选择 M，使这些量始终高于固定的数值容差。

基于这一表示，我们现在给出用于训练 RNOT 映射的目标函数。

训练目标：Kantorovich 半对偶。令 μ 表示定义在 $\mathcal{M}$ 上的源分布，ν 表示目标分布。RNOT 映射会诱导出传输映射 $T_\theta\colon \mathcal{M} \to \mathcal{M}$，进而得到模型分布 $\nu_\theta := T_{\theta\#}\mu$。一种标准替代方案是通过最大似然进行训练（等价于在常数项之外最小化 $\mathrm{KL}(\nu_\theta \| \nu)$），但这需要通过变量替换公式计算 $\log \nu_\theta$，并计算 T_θ的雅可比对数行列式；在高维场景中，这种做法代价高昂，而且往往不稳定。与基于离散化的最优传输方法不同，这里使用地标点不会限制学习得到的传输映射的支撑集。

相反， 我们使用 Kantorovich 半对偶目标函数进行训练，该目标函数仅取决于来自 μ 的样本以及 ν

$$\mathcal{J}(\psi) := \mathbb{E}_{x\sim\mu}\big[\psi^c(x)\big] \;+\; \mathbb{E}_{y\sim\nu}\big[\psi(y)\big],$$
$$W_2(\mu,\nu) := \sup_{\psi} \mathcal{J}(\psi).$$

我们将 ψ 参数化为 $\psi_\theta = f_\theta \circ \varphi$，其中 f_θ 是一个作用于地标特征（17）的多层感知机，并最大化 $\mathcal{J}(\psi_\theta)$。等价地，我们最小化负的半对偶损失

$$\mathcal{L}(\theta) \;:=\; -\mathbb{E}_{x\sim\mu}[\psi_\theta^c(x)] - \mathbb{E}_{y\sim\nu}[\psi_\theta(y)]$$

给定小批量 $\{x_i\}_{i=1}^B \sim \mu$ 和 $\{y_j\}_{j=1}^B \sim \nu$，我们使用蒙特卡洛估计量

$$\widehat{\mathcal{L}}(\theta) \;=\; -\frac{1}{B}\sum_{i=1}^{B}\psi_\theta^c(x_i) - \frac{1}{B}\sum_{j=1}^{B}\psi_\theta(y_j).$$

评估 半对偶目标函数需要求解一个内层优化问题，我们将在下文解决这一问题 t.

初始化内层最小化。计算 c-变换需要在 $\mathcal{M}$ 上求解一个内层优化问题。为了提高稳定性并加快收敛，我们首先在有限个目标分布样本 $\{y_k\}_{k=1}^K \sim \nu$上计算软最小值（对数和指数）近似，并以此启动内循环。具体而言，我们定义

$$y_0(x) = \Pi_{\mathcal{M}}\Big(\sum_{k=1}^{K} \mathrm{softmax}_k\Big(\tfrac{\psi_\theta(y_k)-c(x,y_k)}{\gamma}\Big)\, y_k\Big).$$

其中，$\gamma > 0$ 是温度参数，$\Pi_{\mathcal{M}}$ 表示向 $\mathcal{M}$ 的投影。此初始化过程中停止梯度传播。从 $y_0(x)$出发，可以通过黎曼梯度下降（有或无动量）求解内层问题。

对 c-变换求导。 主要的计算难点在于计算并求导 $\psi_\theta^c(x)$，其中涉及内层最小化。定义

$$F_\theta(x,y) := c(x,y) - \psi_\theta(y),$$
$$\psi_\theta^c(x) := \min_{y\in\mathcal{M}} F_\theta(x,y),$$
$$y_\theta^\star(x) \in \arg\min_{y\in\mathcal{M}} F_\theta(x,y).$$

假设对于当前的 (x,θ)，极小化元 $y_\theta^\star(x)$（局部）唯一，且 $F_\theta(x,\cdot)$ 在 $y_\theta^\star(x)$ 处可微，并满足一阶最优性条件

$$\nabla_y F_\theta\big(x, y_\theta^\star(x)\big) = 0, \tag{18}$$

其中，∇_y 表示关于第二个自变量的黎曼梯度。随后，通过链式法则求导，并利用一阶最优性条件（18）消去经过极小化元的导数，可得

$$\nabla_\theta \psi_\theta^c(x) = \nabla_\theta F_\theta\big(x, y_\theta^\star(x)\big) = \nabla_\theta \psi_\theta\big(y_\theta^\star(x)\big), \tag{19}$$

因为 c 不依赖于 θ。在实践中，这相当于对 $\psi_\theta(y)$ 求导，同时对极小值点 $y_\theta^\star(x)$ 停止梯度传播。

我们的内层求解器返回一个近似极小化点 $\tilde{y}_\theta(x)$，因此 (19) 仅近似成立。由此产生的偏差由驻点残差控制

$$r_\theta(x,\tilde{y}) \;:=\; \big\|\nabla_y F_\theta(x,\tilde{y})\big\|_2.$$

对于平方距离代价 $c(x,y) = \frac{1}{2}d(x,y)^2$，，我们有

$$\nabla_y\Big(\tfrac{1}{2}d(x,y)^2\Big) = -\log_y(x),$$

因此，驻点条件变为 $-\log_y(x) - \nabla\psi_\theta(y) = 0$。相应地，我们监测诊断指标

$$g_\theta(x,\tilde{y}) \;:=\; -\log_{\tilde{y}}(x) - \nabla\psi_\theta(\tilde{y}), \qquad \|g_\theta(x,\tilde{y})\|_2,$$

在训练过程中，以确保内层最小化足够精确，从而实现稳定的外循环优化。

E.3. 算法

算法 1 下文的算法可以通过更改第 10 行，采用重球法（动量）或黎曼 Adam 实现。为简化说明，我们使用标准梯度下降法介绍该算法。此外，我们不介绍任何初始化方法（例如，基于对数和指数的方法）。

算法 1 采用地标距离嵌入的半对偶 RNOT 映射

要求： 流形 $(\mathcal{M}, g)$；源 μ、目标分布 ν位于 $\mathcal{M}$上；地标点 $\{\ell_m\}_{m=1}^L \subset \mathcal{M}$；代价函数 $c(x,y) = \frac{1}{2}\, d(x,y)^2$；批量大小 B；步数 T；内层步数 K；步长 η, α。

对 $m = 1, \ldots, L$，定义嵌入 $\varphi(x) \in \mathbb{R}^L$ by $\varphi_m(x) = d(x, \ell_m)$。

将势函数 $\psi_\theta(x)$ 参数化为多层感知机$_\theta(\varphi(x))$。

初始化 θ。

对于 $t = 1, \ldots, T$ **do**

  采样 $\{x_i\}_{i=1}^B \sim \mu$ 和 $\{y_j\}_{j=1}^B \sim \nu$。

  对于 $i = 1, \ldots, B$ **do**

    $y \leftarrow x_i$。

    对于 $k = 1, \ldots, K$ **do**

      $g \leftarrow \nabla_y\big(c(x_i,\ y) - \psi_\theta(y)\big) = -$对数$_y(x_i) - \nabla\psi_\theta(y)$。

      $y \leftarrow$指数映射$_y(-\alpha g)$。

    结束循环

    $\psi_\theta^c(x_i) \leftarrow c(x_i, y) - \psi_\theta(y)$　　（y 将 ∇_θ [包络定理]视为常数）。

  结束循环

  $\mathcal{L}(\theta) \leftarrow -\frac{1}{B}\sum_{j=1}^B \psi_\theta(y_j) - \frac{1}{B}\sum_{i=1}^B \psi_\theta^c(x_i)$。

  $\theta \leftarrow$ 更新$(\theta, \nabla_\theta \mathcal{L}(\theta))$。

</style}结束循环

返回： ψ_θ。

F. 实验设置

F.1. 真实世界实验

数据集与几何构造。 遵循 Cohen 等人(2021年) 的方法，我们将框架应用于大陆漂移研究（M¨uller 等人，2018年）。我们在 $\mathbb{S}^2$ 上构建源分布和目标经验分布，分别表示距今 1.5 亿年（侏罗纪）和现今（距今 0 年）的陆块。古地理海岸线重建数据来自 GPlates 网络服务（M¨uller 等人，2018年），并采用 Zahirovic 等人(2022年) 的板块运动模型。我们通过拒绝采样，在 $\mathbb{S}^2$ 上均匀采样 50,000 个点，并限制这些点位于重建的大陆边界内。训练期间，我们从该经验点云中均匀采样；进行密度评估时，使用带宽为 $h = 0.1$ 的高斯核密度估计。

训练设置。 我们使用 1024 的批量大小和 1024 个地标点训练模型，共进行 500 次迭代，每次迭代最多进行 1000 次内部求解。由此得到的最优传输映射 $T\colon \mathbb{S}^2 \to \mathbb{S}^2$ 捕捉了 1.5 亿年来大陆漂移的整体运动，且无需了解板块构造。

定性结果。 图 1展示了传输后的映射，以及基础密度和目标密度。传输后的构型还原了大陆漂移中典型的大尺度模式，包括连贯的板块运动和远距离质量传输，表明模型学到了古地理构型之间有意义的几何对应关系。

定性验证。 我们通过测量板块纯度来验证学习到的映射是否保留了构造结构。板块纯度定义为：来自每个源板块的点中，被传输至同一个目标板块的比例。通过在现今点云中进行最近邻查找，为传输后的点分配板块标签。学习到的映射的平均板块纯度为 90%（即来自每个源板块并被传输至同一目标板块的点所占比例），且 Monge 间隙

$$\mathbb{E}[c(x, T(x))] + \mathcal{L}(\theta) < 0.1\%,$$

表明近乎最优的传输（Uscidda 和 Cuturi，2023年）。

F.2. 合成实验

数据与几何设置。我们在两类黎曼流形上评估传输质量： n 维单位球面 $\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$，以及 n 维平坦环面 $\mathbb{T}^n = (\mathbb{S}^1)^n$（乘积流形），维度为$n \in 2,$ ……，10。在两种设置中，源分布均为流形上的均匀分布，目标分布则是尺度为 $\sigma = 0.3$ 的包裹正态分布；在球面上以南极（-1，0，……，0）为中心，在环面上则以相应的点（π，……，π）为中心。

基线方法与指标。我们将所提方法与 RCPM（Cohen 等人，2021年）、RCNF（Mathieu 和 Nickel，2020年）以及莫泽流（Rozen 等人，2021年）进行比较。对于 RCPM，我们对正则化参数进行了遍历

$$\gamma \in \{1.0, 0.1, 0.05, 0.01, 0.005, 0.001\},$$

其中，较小的 γ 会使代价函数变得更尖锐，并在 $\gamma \to 0$ 的极限情况下趋近于精确传输映射。

性能通过正向 Kullback–Leibler（KL散度） $D_{\mathrm{KL}}(T_{\#}\mu|\nu)$ 和有效样本量（有效样本量）进行报告；各项指标均取 5 个批次、每批 1024 个样本的平均值。

训练方案。 我们使用每层 128 个节点的 2 层神经网络、128 个地标点、批量大小为 256，以及 1,000 个训练步骤来训练 RNOT 映射（我们的方法）；内部求解迭代次数最多为 2,500 次。RCPM 使用（Cohen 等人，2021年）的超参数进行训练，使用 68 个地标点、批量大小为 256 和 5,000 次迭代。RCNF 和莫泽流使用默认架构和超参数进行训练，批量大小为 512；RCNF 训练 5,000 次迭代，莫泽流训练 10,000 次迭代。所有实验均在配备 192 GB 内存的 AMD MI300X GPU 上进行。我们观察到，随着维度增加，RCPM 会变得数值不稳定，尤其是在 较小时，这限制了我们的比较范围，使其仅包括 $n \leq 10$具有 5,000 次训练迭代的情况。

消融与敏感性分析。 消融研究（附录 G.1）显示，FPS地标点选择和基于LogSumExp的初始化是影响最大的设计选择，表明地标几何和初始化策略是影响性能的重要因素，也是未来值得探索的方向。

F.3. 度量

为定量评估学习得到的传输映射与目标分布的匹配程度，我们报告基准测度经推前后所得分布与目标分布之间的Kullback–Leibler（KL散度）估计值，同时报告一个基于重要性权重的有效样本量（ESS）诊断指标。

令 μ 表示流形 $\mathcal{M}$上的源分布， ν 表示目标分布，并令 $T_\theta\colon \mathcal{M}\to\mathcal{M}$ 为学习得到的传输映射。模型分布是推前

$$\nu_\theta = (T_\theta)_{\#}\mu.$$

KL 散度。 我们跟踪 KL 散度

$$\mathrm{KL}(\nu_\theta\|\nu) := \int_{\mathcal{M}} \log\left(\frac{\mathrm{d}\nu_\theta}{\mathrm{d}\nu}(y)\right)\mathrm{d}\nu_\theta(y) = \mathbb{E}_{y\sim\nu_\theta}[\log\nu_\theta(y) - \log\nu(y)]$$

在实践中，我们通过蒙特卡洛方法，使用样本 $x_i\sim\mu$ 及其传输后的点 $y_i = T_\theta(x_i)$来估计该量。

假设 μ 和 ν 都有相对于 $\mathrm{vol}_{\mathcal{M}}$的密度（仍记为 $\mu(\cdot)$、 $\nu(\cdot)$）。令 $T_\theta\colon \mathcal{M}\to\mathcal{M}$ 为一个 C^1 微分同胚，并定义推前 $\nu_\theta := (T_\theta)_{\#}\mu$。那么 ν_θ 也有相对于 $\mathrm{vol}_{\mathcal{M}}$的密度，流形上的变量替换公式给出，对于 $y = T_\theta(x)$,

$$\nu_\theta(T_\theta(x))\left|\det(\mathrm{d}T_\theta(x))\right| = \mu(x).$$

其中 $\left|\det(\mathrm{d}T_\theta(x))\right|$ 表示线性映射 $\mathrm{d}T_\theta(x)$ 的绝对雅可比行列式：$T_x\mathcal{M}\to T_{T_\theta(x)}\mathcal{M}$ 其在标准正交基下表示（Falorsi 等人，2019年，引理 D.1）；另见 Stern (2013年，第 2 节）。等价地，

$$\log\nu_\theta(y) = \log\mu(x) - \log\left|\det(\mathrm{d}T_\theta(x))\right|, \qquad y = T_\theta(x).$$

对于光滑映射 $T_\theta\colon\ \mathcal{M}\to\mathcal{M}$，其在 $x\in\mathcal{M}$ 处的微分是如下线性映射

$$\mathrm{d}T_\theta(x) : T_x\mathcal{M}\to T_{T_\theta(x)}\mathcal{M}$$

由 T_θ 如何将 x 处的切向量向前映射来定义。具体而言，若 $\gamma(t)$ 是一条满足 $\gamma(0) = x$ 且$\dot\gamma(0) = v\in T_x\mathcal{M}$的光滑曲线，则

$$\mathrm{d}T_\theta(x)[v] = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} T_\theta(\gamma(t))\in T_{T_\theta(x)}\mathcal{M}.$$

由于 $\mathrm{d}T_\theta(x)$ 在不同的向量空间 $T_x\mathcal{M}$ 和 $T_y\mathcal{M}$ 之间进行映射（其中 $y = T_\theta(x)$），其雅可比行列式需使用黎曼度量定义。假设 $\mathcal{M}$ 嵌入在环境空间 $\mathbb{R}^D$ 中，且 T_θ 由光滑环境映射 $\tilde T_\theta\colon\ \mathbb{R}^D\to\mathbb{R}^D$ 表示，其中 $\tilde T_\theta|_{\mathcal{M}} = T_\theta$。令 $E_x\in\mathbb{R}^{D\times p}$ 和 $E_y\in\mathbb{R}^{D\times p}$ 的列分别构成 $T_x\mathcal{M}$ 和 $T_y\mathcal{M}$ 的标准正交基。$\mathrm{d}T_\theta(x)$ 在这些基下的矩阵为

$$J(x) = E_y^\top(\mathrm{d}\tilde T_\theta(x))E_x\in\mathbb{R}^{p\times p}.$$

内在雅可比行列式为

$$\left|\det(\mathrm{d}T_\theta(x))\right| \equiv \left|\det(J(x))\right|, \qquad \log\left|\det(\mathrm{d}T_\theta(x))\right| = \log\left|\det(J(x))\right|.$$

该量不依赖于标准正交基的具体选择：改变基会在左侧/右侧将 J 乘以正交矩阵，这只会使 $\det(J)$ 的符号发生变化，因此 $|\det(J)|$ 保持不变。

将其代入KL散度的定义，即可得到该估计量

$$\widehat{\mathrm{KL}}(\nu_\theta\|\nu) = \frac{1}{N}\sum_{i=1}^{N}\left(\log\mu(x_i) - \log\left|\det(\mathrm{d}T_\theta(x_i))\right| - \log\nu(y_i)\right).$$

有效样本量。 除KL散度之外，我们还基于权重计算一项重要性采样诊断，这些权重在传输后的样本上将目标密度与模型密度进行比较。定义（未归一化的）重要性权重

$$w_i = \frac{\nu(y_i)}{\nu_\theta(y_i)} = \exp(\log \nu(y_i) - \log \nu_\theta(y_i)).$$

经验归一化常数估计为

$$\widehat{Z} = \frac{1}{N}\sum_{i=1}^{N} w_i,$$

当 ν_θ 和 ν 的支撑集相近，且密度比表现良好时，该值应接近 1。利用这些权重，我们报告有效样本量 (ESS):

$$\mathrm{ESS} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2} = \frac{1}{\sum_{i=1}^{N} \tilde{w}_i^2},$$

其中， $\tilde{w}_i = w_i / \sum_j w_j$ 是归一化权重。当 T_θ 是最优传输映射且 $\nu_\theta = \nu$ 时，所有权重相等，且 ESS $= N$。我们报告归一化比值 $\mathrm{ESS}/N \in [0, 1]$。

ESS 用于衡量权重退化：ESS $\approx N$ 表示权重近乎均匀（重叠良好），而 ESS 较小则表示大部分质量集中在少数样本上（重叠较差）。

通过隐函数定理计算雅可比矩阵。 虽然我们提供了通过迭代内层求解器进行反向传播的代码，但在大规模运算时，展开许多次迭代可能会导致数值不稳定（且耗费大量内存）。因此，我们使用隐式微分 $\mathrm{d}T_\theta(x)$ 来计算传输雅可比矩阵，方法是 应用隐函数定理（IFT）（例如Lee(2012，第 4 章）；另请参阅 Blondel 等人(2022年）中的隐式层讨论。

传输 映射 $T_\theta(x) = y^\star(x)$ 由一阶最优性条件的解隐式定义

$$F(x, y) := -\log_y(x) - \nabla\psi_\theta(y) = 0,$$

其中 $\log_y(x) \in T_y\mathcal{M}$ 是黎曼对数映射（在不位于割迹上的点对 (y, x) 处），而 $\nabla\psi_\theta(y) \in T_y\mathcal{M}$ 是对偶势函数的黎曼梯度。特别地， $F(x, y) \in T_y\mathcal{M}$ 是输出点 y处的切向量。
假设关于第二个自变量的偏微分

$$D_y F(x, y^\star(x)) : T_{y^\star(x)}\mathcal{M} \to T_{y^\star(x)}\mathcal{M},$$

是可逆的，则隐函数定理意味着 $y^\star(x)$ 在局部可微，且其微分满足

$$\mathrm{d}T_\theta(x) = -\big(D_y F(x, y^\star(x))\big)^{-1} \circ D_x F(x, y^\star(x)),$$

其中 $D_x F(x, y^\star(x)) : T_x\mathcal{M} \to T_{y^\star(x)}\mathcal{M}$。
为获得内在的 $p \times p$ 雅可比矩阵，我们在正交归一切基下进行计算。设 $E_x \in \mathbb{R}^{D\times p}$ 和 $E_y \in \mathbb{R}^{D\times p}$ 的列分别构成 $T_x\mathcal{M}$ 和$T_{y^\star(x)}\mathcal{M}$的标准正交基（在环境表示 $\mathcal{M} \subset \mathbb{R}^D$中）。对 F的环境实现进行自动微分，得到环境雅可比算子

$$\big(D_y F\big)_{\mathrm{amb}} \in \mathbb{R}^{D\times D}, \qquad \big(D_x F\big)_{\mathrm{amb}} \in \mathbb{R}^{D\times D},$$

并将其投影到切坐标:

$$[D_y F] = E_y^\top \big(D_y F\big)_{\mathrm{amb}} E_y \in \mathbb{R}^{p\times p}, \qquad [D_x F] = E_y^\top \big(D_x F\big)_{\mathrm{amb}} E_x \in \mathbb{R}^{p\times p}. \tag{20}$$

随后，通过求解 $p \times p$ 线性系统，得到内在雅可比矩阵 $J(x)$

$$J(x) = -[D_y F]^{-1}[D_x F], \tag{21}$$

并通过 log |det($J(x)$)| 计算变量变换项。

单样本成本主要由 $O(p^3)$ 线性代数运算（求解和求行列式）以及计算（20) 中投影雅可比矩阵的开销构成（可通过与 p 基向量计算 JVP/VJP 乘积来高效实现，而无需构造完整的 $D \times D$ 矩阵）。

在流形上估计 KL 散度在数值上可能较为棘手，因为它需要结合对数密度和雅可比行列式，因此微小误差也可能累积。此外，由于抽样变异，有限 N 情况下的 KL 蒙特卡洛估计量可能为负，数值误差还可能加剧这一现象；我们在 *Cohen* 等人的工作 (2021年) 以及我们自己的实现中，都偶尔观察到了负的估计值。

基于隐函数定理的雅可比矩阵也对内层求解器精度十分敏感：如果传输后的点 $y^\star(x)$ 未能以足够精度满足驻点条件 $F(x, y^\star(x)) \approx 0$，或者 $[D_y F]$ 的条件数过大，那么（21）就会变得不可靠。根据经验，我们发现残差范数低于 10^{-2} 时，KL 散度估计较为稳定，不过这一阈值取决于具体问题。

最后，我们强调，半对偶目标函数直接优化传输质量，而非 KL 散度。因此，训练良好的模型可能会产生出色的传输映射（例如，在代价函数或视觉诊断指标方面表现出色），但 KL 估计值仍可能存在较大噪声。因此，我们主要将 KL 视为诊断性总结，而非训练目标。

提出替代方法的理由。 我们将 RNOT 与流形上的代表性归一化流基线进行比较：RCNF（Mathieu 和 Nickel, 2020 年）、莫泽流（ Rozen 等人， 2021 年）以及黎曼凸势映射（ RCPM ）（ Cohen 等人， 2021 年）。这些方法代表了在黎曼流形上学习传输映射的主要方向：基于常微分方程的流（RCNF）、基于散度的密度匹配（莫泽流），以及通过凸势函数实现的离散最优传输（RCPM）。

如前所述，我们的方法优化了半对偶损失；但为了进行比较，我们采用 KL 散度和有效样本量（ESS），作为衡量传输质量的另一种指标。 值得注意的是，许多近期的黎曼扩散模型主要面向生成模型设计，因此采用不同的评估方案；具体而言，基于分数的扩散模型（Huang 等人，2022年；De Bortoli 等人，2022年）通常报告留出数据上的负对数似然，而非显式学习得到的传输映射的 KL 散度。我们选用的基线方法 RCNF、莫泽流和 RCPM 均可通过变量替换公式进行可处理的密度评估，因此能够基于相同指标进行严谨比较。对于作者未实现 KL 散度、且计算并非显而易见的方法（例如 RCNF 或莫泽流），若要评估则需要临时补充，这可能会不公平地呈现其方法，或引入意料之外的错误。

G. 更多结果

G.1. 消融与敏感性分析

我们进行了消融研究，以了解方法中关键设计选择的作用。我们考虑了四种流形设置：$\mathbb{S}^2$、$\mathbb{T}^2$、$\mathbb{S}^{10}$ 和 $\mathbb{T}^{10}$，并使用 KL 散度评估性能（越低越好）。

研究结果总结。表3 总结了各项消融实验相对于基线的表现。我们重点指出三项关键发现：

1. **FPS 地标点选择始终能提升性能**。与随机采样相比，使用最远点采样（FPS）进行地标选择，在所有流形上都能获得最佳或接近最佳的 KL 散度；在球面上的提升为 27–52%，在 $\mathbb{T}^2$上提升为 34%。

2. **LogSumExp 初始化对环面流形至关重要**。禁用 LogSumExp 初始化可能导致环面上的结果失效（在$2.2\times$ 上更差，在 $\mathbb{T}^2$上更差，在 $3.5\times$ 上更差，在 $\mathbb{T}^{10}$上更差），而对球面的影响则微乎其微。这表明，该初始化对于处理环面的周期结构至关重要。

3. **内层学习率对收敛有显著影响**。 较低的内层学习率（10^{-3} 而非 5×10^{-2}）会严重降低在环面上的性能，在 $\mathbb{T}^2$ 上导致KL散度差 $6.2\times$，在 $\mathbb{T}^{10}$上差 $12.3\times$。

详细分析。**求解器设计**。 极小值点求解器的配置对性能影响最大。基于LogSumExp的初始化可提供良好的初始值，这对环面流形尤为重要，因为周期边界条件使优化更具挑战性。与标准随机梯度下降相比，Adam 优化器能带来稳定的收益，尤其是在损失景观更复杂的高维环面上。

网络架构。 FPS地标点选择带来了最稳定的性能提升，这表明分布均匀的地标点能更好地捕捉流形的几何特征。网络宽度和深度的影响较小：较小的网络（[32, 32]）在 $\mathbb{S}^{10}$上也能达到甚至超过基线性能，这表明对于较简单的流形，基线模型可能略微过度参数化。

训练超参数。 较低的外层学习率（10^{-4} 而非 10^{-3}）会持续损害性能，这可能是由于固定步数预算内训练不足，或缺乏摆脱较差解的能力。较大的批量大小（512）能改善球面上的结果，但对环面的影响较小。这些结果表明，随着维度增加，较大的批量大小并非必不可少。

建议。 基于这些消融实验，我们建议：

- 使用 FPS地标点选择，以改善几何覆盖范围
- 保持启用 LogSumExp 初始化，尤其适用于环面
- 在内循环中使用 Adam 优化器
- 对于球面，使用 256–512 的批量大小

G.2. 高维流形的进一步结果

表3. 展示不同流形上 KL 散度（↓）的消融研究结果。粗体表示最佳结果；下划线表示相较于基线性能下降 $> 50\%$。所有结果均取自 5 个评估批次，每批包含 1024 个样本，并取其平均值。

配置	$\mathbb{S}^2$	$\mathbb{T}^2$	$\mathbb{S}^{10}$	$\mathbb{T}^{10}$
基线	0.040	0.125	0.048	1.07
求解器消融				
不使用对数和指数初始化	0.043	<u>0.270</u>	0.048	<u>3.79</u>
内层步数：500	0.041	0.129	0.048	1.11
内层步数：4000	0.040	0.122	0.048	1.06
内层学习率：10^{-3}	0.055	<u>0.775</u>	0.048	<u>13.13</u>
γ_{LSE}: 1.0	0.040	0.124	0.048	1.24
γ_{LSE}: 0.01	0.040	0.127	0.048	1.07
γ_{LSE}: 0.001	0.040	0.126	0.048	1.06
不使用 Adam（随机梯度下降）	0.046	<u>0.219</u>	0.048	1.44
架构消融实验				
地标点：最远点采样	**0.019**	**0.083**	**0.035**	**0.03**
地标点：32	0.036	0.099	0.054	<u>3.71</u>
地标点：64	0.041	0.203	0.048	1.47
地标点：256	0.036	0.036	0.053	—
隐藏层：[32, 32]	0.046	0.202	0.040	1.16
隐藏层：[64, 64]	0.034	0.170	0.044	1.23
隐藏层：[256, 256]	0.037	0.104	0.045	1.04
隐藏层：[128, 128, 128]	0.040	0.128	0.047	1.19
训练消融实验				
外层学习率：10^{-4}	<u>0.067</u>	0.184	<u>0.060</u>	1.59
批量大小：512	0.036	0.118	**0.034**	1.15
批量大小：128	<u>0.072</u>	0.183	0.040	1.16

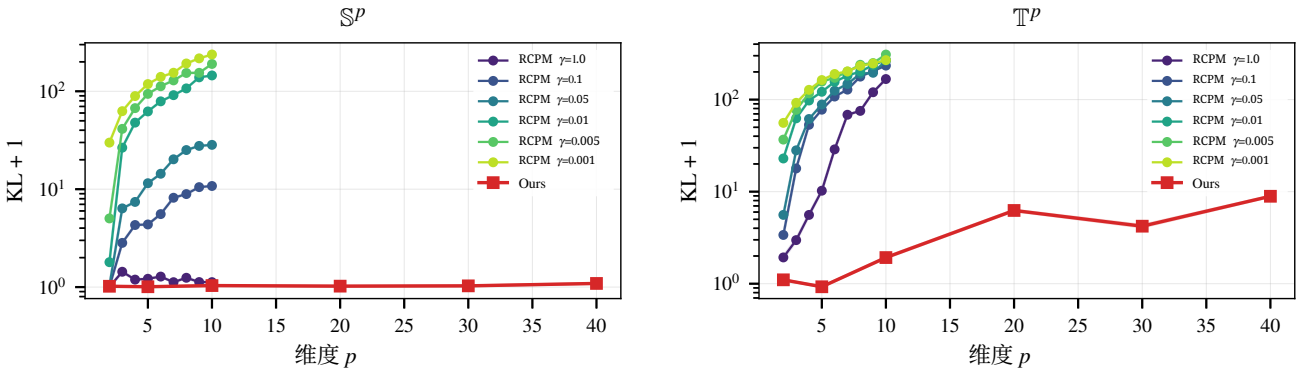


图5. 高维流形上 KL 散度随维度的变化。我们将所提方法与采用不同正则化参数 $\gamma \in \{0.001, 0.005, 0.01, 0.05, 0.1, 1.0\}$ 的 RCPM 进行比较。由于更高维度下计算不可行性，RCPM 的结果仅展示 $p \in \{2, \ldots, 10\}$；而我们的方法可扩展至 $p = 40$。在所有测试维度上，我们的方法均实现了更低的 KL 散度，并能良好地扩展至高维场景，从实证角度支持了我们的理论结果。

H. 证明

H.1. 定理3.1 的证明

回顾以下 Iacobelli (2016年) 中对量化误差的定义会很有帮助。

定义 H.1 (量化误差；Iacobelli (2016年) 中的定义 1.1)。 设 $(\mathcal{M}, d)$ 是一个完备黎曼流形，$\rho \in \mathcal{P}(\mathcal{M})$， $r \geq 1$，且 $N \in \mathbb{N}$。第 N 阶、规模为 r 的量化误差为

$$V_{N,r}(\rho) := \inf_{\alpha \subset \mathcal{M}:\ |\alpha| \leq N} \int_{\mathcal{M}} \min_{a \in \alpha} d(a, y)^r \, \mathrm{d}\rho(y).$$

我们在证明定理 3.1 时将用到的一个关键结果是 Iacobelli (2016年) 的定理 1.4，现为方便起见在此陈述。为了详细陈述我们的主要结果，我们需要引入 Iacobelli (2016年) 中的一些记号：给定一点 $x_0 \in \mathcal{M}$，我们可以在 (ρ, θ) 上考虑由常值度量 $T_{x_0}\mathcal{M} \simeq \mathbb{R}^p$ 所诱导的极坐标 g_{x_0}，其中 θ 表示单位球面 $\mathbb{S}^{p-1}$上的一个向量。接着，我们可以定义以下量，用来衡量指数映射的微分限制在半径为 $\mathbb{S}^{p-1}_\rho \subset T_{x_0}\mathcal{M}$ 的球面 ρ上的规模：

$$A_{x_0}(\rho) := \sup_{v \in \mathbb{S}^{p-1}_\rho, w \in T_v \mathbb{S}^{p-1}_\rho, |w|_{x_0} = \rho} \| d_v \exp_{x_0}[w] \|_{\exp_{x_0}(v)}.$$

现在我们可以给出Iacobelli (2016年)中的定理 1.4。

定理 H.2 (Iacobelli，2016年中的定理 1.4)。设 $(\mathcal{M},\ g)$ 是一个 p维无边界完备黎曼流形，并设 $\mu = h\mathrm{dvol} + \mu^s$ 是定义在 $\mathcal{M}$上的概率测度。假设存在一点 $x_0 \in \mathcal{M}$ 和 $\delta > 0$ 使得

$$\int_{\mathcal{M}} d(x, x_0)^{r+\delta} \mathrm{d}\mu(x) + \int_{\mathcal{M}} A_{x_0}(d(x, x_0))^r \mathrm{d}\mu(x) < \infty. \tag{22}$$

Then

$$\lim_{N \to \infty} N^{r/p} V_{N,r}(\mu) = Q \left(\int_{\mathcal{M}} h^{p/(p+r)} \mathrm{d}x \right)^{(p+r)/p}.$$

其中 Q 是一个正常数。

我们的下一个结果表明，定理 H.2 可直接应用于我们的设定（即当 $(\mathcal{M},\ g)$ 是紧黎曼流形时），因为充分条件总是成立。

引理 H.3 （Iacobelli 的可积性条件在紧流形上自动成立）。设 $(\mathcal{M},\ g)$ 是一个紧黎曼流形（因此完备且无边界），设 $\mu \in \mathcal{P}(\mathcal{M})$，并固定 $x_0 \in \mathcal{M}$， $r \geq 1$，以及 $\delta > 0$。则可积性条件 (22)成立，即

$$\int_{\mathcal{M}} d(x, x_0)^{r+\delta} \, \mathrm{d}\mu(x) \ + \ \int_{\mathcal{M}} A_{x_0}(d(x, x_0))^r \, \mathrm{d}\mu(x) < \infty.$$

引理的证明H.3 推迟至附录H.7。

引理 H.4 (来自$V_{m,\ 2}$ 至 W_2)。 设 $\rho \in \mathcal{P}_2(\mathcal{M})$ 且设 η 的支撑集至多包含 m个点。则

$$V_{m,2}(\rho) \ \leq \ W_2(\rho, \eta)^2$$

引理的证明H.4 推迟至附录H.8。

引理 H.5 （共同源测度下的 Wasserstein 稳定性）。 设 $\mu \in \mathcal{P}(\mathcal{M})$ 并设 t， $s\colon \mathcal{M} \to \mathcal{M}$ 是可测的。令 $\nu_t := t_{\#}\mu$ 和 $\nu_s := s_{\#}\mu$。则 ν_t， $\nu_s \in \mathcal{P}_2(\mathcal{M})$ 且

$$W_2(\nu_t, \nu_s) \ \leq \ \mathrm{RMSE}_\mu(t, s). \tag{23}$$

引理的证明H.5 推迟至附录 H.9。

定理 *3.1* 的证明。 令 $p := $ 维数$(\mathcal{M})$ 并定义 $\nu := (T_\star)_{\#}\mu$。
步骤 1：下界$V_{m,\,2}(\nu)$。 由于 $\nu \ll$ 体积$_{\mathcal{M}}$，记 $\nu = h$体积$_{\mathcal{M}}$，其中 $h \geq 0$ 且 $\int_{\mathcal{M}} h$ 体积$_{\mathcal{M}} = 1$。应用定理H.2，取 $r = 2$ 和 $\rho = \nu$。由于 $\mathcal{M}$ 是紧致的，引理H.3 表明可积性条件（22）成立，因此该定理给出渐近公式

$$\lim_{m\to\infty} m^{2/p} V_{m,2}(\nu) = Q\left(\int_{\mathcal{M}} h^{p/(p+2)}\,\mathrm{dvol}_{\mathcal{M}}\right)^{(p+2)/p} =: C_{\mathrm{quant}}.$$

此外 $C_{\mathrm{quant}} > 0$：事实上，由于 ν 是一个概率测度， $h > 0$ 在一个黎曼体积为正的集合上，因此$\int_{\mathcal{M}} h^{p/(p+2)}$ 体积$_{\mathcal{M}} > 0$。
根据极限的定义，存在 $m_0 \in \mathbb{N}$，使得对所有 $m \geq m_0$，

$$m^{2/p} V_{m,2}(\nu) \;\geq\; \frac{C_{\mathrm{quant}}}{2}, \qquad \text{equivalently} \qquad V_{m,2}(\nu) \;\geq\; \frac{C_{\mathrm{quant}}}{2}\, m^{-2/p}.$$

对于有限多个指标 $1 \leq m < m_0$，定义

$$c_\star := \min_{1\leq m<m_0} m^{2/p} V_{m,2}(\nu).$$

这个最小值中的每一项都严格为正：由于 $\nu \ll$ 体积$_{\mathcal{M}}$，它不可能支撑在有限多个点上，因此对于每个固定的 m，$V_{m,\,2}(\nu) > 0$。所以 $c_\star > 0$。最后令

$$C_0 := \min\Big\{\frac{C_{\mathrm{quant}}}{2},\, c_\star\Big\} > 0.$$

那么，对每个 $m \in \mathbb{N}$，

$$V_{m,2}(\nu) \;\geq\; C_0\, m^{-2/p}, \tag{24}$$

这就是所需的一致下界。

步骤 2：离散输出映射蕴含 Wasserstein下界。 固定 $m \in \mathbb{N}$并令 $T \in \mathsf{D}_m(\mu)$。设 $\nu_T := T_{\#}\mu$，根据定义，其支撑集至多包含 m 个点 $\mathsf{D}_m(\mu)$。应用引理H.4 可得

$$V_{m,2}(\nu) \;\leq\; W_2(\nu, \nu_T)^2.$$

结合式 (24，得到

$$W_2(\nu, \nu_T) \;\geq\; \sqrt{C_0}\, m^{-1/p}. \tag{25}$$

步骤 3：从 Wasserstein距离到映射均方根误差。 应用引理 H.5 其中 $t := T$ 且 $s := T_\star$，可得

$$W_2(\nu_T, \nu) \leq \mathrm{RMSE}_\mu(T, T_\star).$$

结合（25）可得，对每个 $T \in \mathsf{D}_m(\mu)$，

$$\mathrm{RMSE}_\mu(T, T_\star) \;\geq\; \sqrt{C_0}\, m^{-1/p}.$$

对 $T \in \mathsf{D}_m(\mu)$ 取下确界即可证明（6），其中 $C := \sqrt{C_0}$。条件 均方根误差$_\mu(T,\ T_\star) \leq \delta$ 通过重新排列可推出 $m \geq (C/\delta)^p$。 □

H.2. 推论 3.2 的证明

推论的证明 *3.2*。 固定 $m \in \mathbb{N}$ 并令 $\phi \in \mathcal{C}_m(\mathcal{M})$。根据 $\mathcal{C}_m(\mathcal{M})$的定义，存在点位 $\{y_i\}_{i\in[m]} \subset \mathcal{M}$ 和标量 $\{\alpha_i\}_{i\in[m]} \subset \mathbb{R}$ 使得，对所有 $x \in \mathcal{M}$,

$$\phi(x) = \min_{i\in[m]} \Big(\tfrac{1}{2} d(x, y_i)^2 + \alpha_i\Big).$$

Set

$$f_i(x) := \tfrac{1}{2} d(x, y_i)^2 + \alpha_i, \qquad i \in [m].$$

步骤 1：定义单元和并列集。对于每个 $i \in [m]$，定义（严格）单元

$$A_i := \{x \in \mathcal{M} : \ f_i(x) < f_j(x) \text{ for all } j \neq i\},$$

并定义并列集

$$B := \Big\{x \in \mathcal{M} : \ \exists\, i \neq j \text{ with } f_i(x) = f_j(x) = \phi(x)\Big\}.$$

（McCann，2001）中的引理 1 表明，$x \mapsto \frac{1}{2}d(x,\ y_i)^2$ 在 $\mathcal{M}$ 上对每个固定的 y_i都是利普希茨的，因此每个 f_i 都是利普希茨的（因而是连续的）。对于 $i \neq j$，定义 $g_{ij} := f_i - f_j$；于是 g_{ij} 是连续的，并且

$$A_i = \bigcap_{j \neq i} \{x \in \mathcal{M} : \ g_{ij}(x) < 0\} = \bigcap_{j \neq i} g_{ij}^{-1}((-\infty, 0))$$

它是连续映射下某个开集的有限个原像的交，因此是开集。每个点 $x \in \mathcal{M}$要么有唯一极小值点，要么存在并列，因此

$$\mathcal{M} = \Big(\bigcup_{i=1}^{m} A_i\Big) \cup B, \qquad A_i \cap A_j = \varnothing \ (i \neq j).$$

步骤 2： ϕ **几乎处处可微**。由于 ϕ 是有限个利普希茨函数的最小值，因此它在 $\mathcal{M}$. 上是利普希茨的。根据黎曼流形上的拉德马赫定理（Burago 等人，2001，定理 5.5.7），存在一个集合 $\mathcal{U} \subset \mathcal{M}$，满足 $_{\mathcal{M}}(\mathcal{U}) \ = 0$，使得 ϕ 在 $\mathcal{M} \setminus \mathcal{U}$上可微。

步骤 3：识别 $\nabla\phi$ **在每个** A_i**上**。固定 $i \in [m]$ 和 $x \in A_i \setminus \mathcal{U}$。根据 A_i 的定义，对所有$j \neq i$，我们有 $f_i(x) < f_j(x)$。由连续性，存在 $r > 0$，使得严格不等式在 $B(x,\ r)$上仍然成立：

$$f_i(z) < f_j(z) \quad \forall j \neq i,\ \forall z \in B(x, r).$$

因此，对所有 $z \in B(x,\ r)$，有 $\phi(z) \ = f_i(z)$，故 ϕ 在局部与光滑函数 f_i 重合，并且

$$\nabla\phi(x) = \nabla f_i(x) = \nabla\Big(\tfrac{1}{2}d(x, y_i)^2\Big).$$

设割迹(y_i) 表示 y_i的割迹。由(8)，若 $x \in \mathcal{M} \setminus$ (割迹$(y_i) \cup \{y_i\})$，则 $y_i \notin$ 割迹(x) 并且

$$\nabla\Big(\tfrac{1}{2}d(x, y_i)^2\Big) = -\log_x(y_i),$$

因此，对每一个满足 $x \notin$ 割迹$(y_i) \cup \{y_i\}$ 的 $x \in A_i \setminus \mathcal{U}$，我们得到

$$\nabla\phi(x) = -\log_x(y_i).$$

因此，对于这样的 x,

$$T_\phi(x) := \exp_x\big(-\nabla\phi(x)\big) = \exp_x\big(\log_x(y_i)\big) = y_i.$$

Let

$$\mathrm{Cut} := \Big(\bigcup_{i=1}^{m} \mathrm{Cut}(y_i)\Big) \cup \{y_1, \ldots, y_m\}.$$

一点的割迹具有零黎曼体积，因此$\mathrm{vol}_{\mathcal{M}}$(割迹) $= 0$ 作为有限并 n.

步骤 4: Th e 并列集 B **的黎曼体积为零**。对于每个 $i \neq j$，定义超曲面候选

$$H_{ij} := \{x \in \mathcal{M} : \ f_i(x) = f_j(x)\}.$$

于是 $B \subseteq \bigcup_{1 \leq i < j \leq m} H_{ij}$，因此只需证明 $\mathrm{vol}_{\mathcal{M}}(H_{ij}) \ = 0$。固定 $i \neq j$ 并定义

$$h_{ij}(x) := f_i(x) - f_j(x) = \tfrac{1}{2}\big(d(x, y_i)^2 - d(x, y_j)^2\big) + (\alpha_i - \alpha_j).$$

Let

$$S_{ij} := \mathrm{Cut}(y_i) \cup \mathrm{Cut}(y_j) \cup \{y_i, y_j\}.$$

在 $\mathcal{M} \setminus S_{ij}$ 上，两个平方距离函数都是光滑的，因此 h_{ij} 在此处光滑，并且

$$\nabla h_{ij}(x) = -\log_x(y_i) + \log_x(y_j) \qquad (x \in \mathcal{M} \setminus S_{ij}).$$

我们声称，对所有 $x \in H_{ij} \setminus S_{ij}$，$\nabla h_{ij}(x) \neq 0$。事实上，若在某个 $x \notin S_{ij}$ 处有 $\nabla h_{ij}(x) = 0$，则 $\log_x(y_i) = \log_x(y_j)$，应用 $\exp_x$ 后得到 $y_i = y_j$，这与 $i \neq j$ 矛盾。因此，0 是 h_{ij}：$\mathcal{M} \setminus S_{ij} \to \mathbb{R}$ 的正则值，所以 $H_{ij} \setminus S_{ij} = h_{ij}^{-1}(0)$ 是 $\mathcal{M} \setminus S_{ij}$ 中一个嵌入的 $(p-1)$ 维子流形，特别地，其 p 维黎曼体积为零。由于 $\mathrm{vol}_{\mathcal{M}}(S_{ij}) = 0$，我们得出 $\mathrm{vol}_{\mathcal{M}}(H_{ij}) = 0$。由并集的有限性，

$$\mathrm{vol}_{\mathcal{M}}(B) = 0.$$

步骤 5：离散输出性质 μ-几乎处处 定义例外集

$$N := \mathcal{U} \cup \mathrm{Cut} \cup B.$$

我们有 体积$_{\mathcal{M}}(N) = 0$。根据定理 3.1的假设，我们有 $\mu \ll$体积$_{\mathcal{M}}$，因此 $\mu(N) = 0$。在 $\mathcal{M} \setminus N$上，每个x 都恰好属于一个单元 A_i 并满足 $T_\phi(x) = y_i$；因此

$$T_\phi(x) \in \{y_1, \ldots, y_m\} \qquad \text{for } \mu\text{-a.e. } x \in \mathcal{M}.$$

等价地， T_ϕ 在一个具有全 μ测度的集合上至多取 m 个不同的值（并且 $T_{\phi\#}\mu$ 的支撑集至多包含 m 个点）。因此 $T_\phi \in \mathsf{D}_m(\mu)$。

S第 6 步：维数灾难下界。由于 $\{T_\phi : \ \phi \in \mathcal{C}_m(\mathcal{M})\} \subseteq \mathsf{D}_m(\mu)$，我们有 e

$$\inf_{\phi \in \mathcal{C}_m(\mathcal{M})} \mathrm{RMSE}_\mu(T_\phi, T_\star) \ \geq \ \inf_{T \in \mathsf{D}_m(\mu)} \mathrm{RMSE}_\mu(T, T_\star).$$

应用定理 3.1 可得一个常数 $C > 0$，使得对所有 $m \in \mathbb{N}$,

$$\inf_{\phi \in \mathcal{C}_m(\mathcal{M})} \mathrm{RMSE}_\mu(T_\phi, T_\star) \ \geq \ C\, m^{-1/p},$$

这就是（7）。样本复杂度结论随即得出：若 均方根误差$_\mu(T_\phi, \ T_\star) \ \leq \ \delta$，则必有 $m \geq (C/\delta)^p$。

□

H.3. 定理4.1的证明

在证明定理4.1之前，我们先给出以下收敛结果。

引理 *H.6*（沿 c-变换逼近序列的梯度收敛）。设 $\mu \in \mathcal{P}(\mathcal{M})$ 满足 $\mu \ll \mathrm{vol}_{\mathcal{M}}$。设$\varphi : \mathcal{M} \to \mathbb{R}^n$ 是满足假设 *2.2* 的特征映射，并假设 $\mathcal{F}$ 在 $C(\mathbb{R}^n, \mathbb{R})$ 中关于 *ucc*拓扑稠密。那么对每个 $\phi \in \Psi_c(\mathcal{M})$，都存在一个序列 $\psi_k \in \varphi^* \mathcal{F}$ 和一个可测集 $N_\star \subset \mathcal{M}$，使得 $\mu(N_\star) = 1$，并且令 $\phi_k := \psi_k^c$,

$$\|\phi_k - \phi\|_\infty \to 0,$$

且对每个 $x \in N_\star$，函数 ϕ 和 ϕ_k 在 x 处可微，并且

$$\nabla \phi_k(x) \to \nabla \phi(x).$$

定理的证 引理H.6 的证明推迟到附录H.10。现在我们准备证明定理4.1 .

定理的证明 *4.1*。 根据定理 B.3，由于 $\mu \ll \mathrm{vol}_{\mathcal{M}}$和 $c(x, \ y) \ = \ \frac{1}{2} d(x, \ y)^2$, 存在一个 c-凹势函数$\phi \in \Psi_c$（$\mathcal{M}$），使得最优传输映射 $T_\star$ 从 μ 到 ν 几乎处处（相对于 μ）由下式给出：

$$T_\star(x) = \exp_x\big(-\nabla \phi(x)\big).$$

步骤 1：构造几乎处处梯度收敛的逼近势函数。 由于 $\mathcal{F}$ 在 $C(\mathbb{R}^n, \mathbb{R})$ 中关于 ucc拓扑稠密，且 φ 满足假设 2.2，因此拉回类 $\varphi^*\mathcal{F}$ 在 $C(\mathcal{M}, \mathbb{R})$ 中关于 ucc拓扑稠密。因此，引理 H.6 的假设均得到满足。将引理 H.6 应用于上述 ϕ 可得到一个序列 $\psi_k \in \varphi^*\mathcal{F}$ 及相应的 c-变换

$$\phi_k := \psi_k^c \in \mathfrak{C}(\varphi^*\mathcal{F})$$

使得

$$\|\phi_k - \phi\|_\infty \to 0 \qquad \text{and} \qquad \nabla\phi_k(x) \to \nabla\phi(x) \ \text{ for } \mu\text{-a.e. } x \in \mathcal{M}.$$

设 $N_\star \subset \mathcal{M}$ 为一个博雷尔集，满足 $\mu(N_\star) = 1$，在其上 ϕ 以及所有 ϕ_k 均可微，且对每个$x \in N_\star$有 $\nabla\phi_k(x) \to \nabla\phi(x)$。

步骤 2：定义处处有定义的映射 T_k 与 $T_\star$。 固定任意一点 $y_0 \in \mathcal{M}$，并定义映射 T_k，$T_\star$：$\mathcal{M} \to \mathcal{M}$如下

$$T_k(x) := \begin{cases} \exp_x\big(-\nabla\phi_k(x)\big), & x \in N_\star, \\ y_0, & x \notin N_\star, \end{cases} \qquad T_\star(x) := \begin{cases} \exp_x\big(-\nabla\phi(x)\big), & x \in N_\star, \\ y_0, & x \notin N_\star. \end{cases}$$

对 $N_\star^c$的这一修改无关紧要，因为 $\mu(N_\star^c) = 0$，但它确保 T_k与 $T_\star$处处有定义（因而可视为$(\mathcal{M}, \mu)$上的随机变量）。

步骤 3：在 $N_\star$上的逐点收敛。 固定 $x \in N_\star$。由于 $\mathcal{M}$ 是紧致的，故它是完备的，并且由霍普夫–里诺定理（见定理A.2），指数映射$\exp_x$：$T_x\mathcal{M} \to \mathcal{M}$在整个 $T_x\mathcal{M}$ 上有定义，并且作为切向量的函数是光滑的，因而连续。因此，由 $\nabla\phi_k(x) \to \nabla\phi(x)$ 在 $T_x\mathcal{M}$中，我们得到

$$T_k(x) = \exp_x\big(-\nabla\phi_k(x)\big) \longrightarrow \exp_x\big(-\nabla\phi(x)\big) = T_\star(x).$$

因此，对每个 $x \in N_\star$，$T_k(x) \to T_\star(x)$；又由于 $\mu(N_\star) = 1$，我们有

$$T_k(x) \to T_\star(x) \qquad \text{for } \mu\text{-a.e. } x \in \mathcal{M}.$$

步骤 4：几乎必然收敛蕴含依概率收敛。 回顾概率论中的标准结论：若$X_k \to X$ 在概率测度 μ下几乎必然成立，则 $X_k \to X$ 依概率成立。将此结论应用于在分布 $x \sim \mu$下的随机变量$X_k(x) := T_k(x)$ 和 $X(x) := T_\star(x)$。由于 $T_k \to T_\star$ μ-几乎处处成立，因此

$$T_k \to T_\star \qquad \text{in probability under } \mu.$$

这就证明了序列 (T_k) 在 μ下依概率一致逼近最优传输映射 $T_\star$，如所声称。

□

H.4. 定理5.1的证明

我们研究紧集上函数的神经网络逼近。我们所依赖的逼近结果来自 Yarotsky& Zhevnerchuk (2020年)；这些结果针对的是 Yarotsky (2017年) 引入的网络模型，其中网络被描述为一个前馈有向无环图（DAG）。这种基于图的观点比传统的全连接分层结构稍具一般性：它自然支持稀疏连接和跳跃连接，并直接以图中所含标量可训练参数的数量衡量复杂度。为尽量减少符号使用上的不便，并逐字引用Yarotsky & Zhevnerchuk (2020年) 的表述，我们在接下来的定义中采用这一约定。

定义 H.7（前馈神经网络架构）。固定输入维度 $n \in \mathbb{N}$和激活函数 σ：$\mathbb{R} \to \mathbb{R}$。前馈网络架构 是一个有向无环图，其顶点（计算单元）被划分为 层 $V_0, \ldots, V_{L-1}$，其结构如下：

- $V_0 = \{1, \ldots, n\}$ 是输入层；对于 $x \in \mathbb{R}^n$，我们对每个 $i \in V_0$设定 $z_i = x_i$。
- $V_{L-1} = \{\text{out}\}$ 是由单个计算单元组成的输出层 。

对于每个 $\ell \in \{1, \dots, L-1\}$ 和每个计算单元 $v \in V_\ell$，其入邻居集合满足

$$\mathrm{In}(v) \subseteq \bigcup_{j=0}^{\ell-1} V_j,$$

因此，$\mathrm{In}(v)$ 可以包含来自任意先前层的单元（允许使用跳跃连接），并且对于所有 $u \in \mathrm{In}(v)$，架构都包含有向边 $(u \to v)$。

ˆ 给定权重和偏置 $(w_{vu})_{(u\to v)\in E}$ 以及 $(b_v)_{v \in \cup_{\ell=1}^{L-1} V_\ell}$，相应的实现函数 $f\colon \mathbb{R}^n \to \mathbb{R}$ 由以下递归定义

$$z_v \;=\; \sigma\left(\sum_{u \in \mathrm{In}(v)} w_{vu}\, z_u + b_v \right), \qquad v \in \bigcup_{\ell=1}^{L-2} V_\ell,$$

以及输出规则（无激活函数）

$$\hat{f}(x) = z_{\mathrm{out}} \;=\; \sum_{u \in \mathrm{In}(\mathrm{out})} w_{\mathrm{out},u}\, z_u + b_{\mathrm{out}}.$$

前述定义规定了架构（分层有向无环图）以及分配参数后对应的计算规则。现在，我们确定Yarotsky （2017年）和Yarotsky & Zhevnerchuk （2020年）中使用的复杂度度量，即底层图的深度（层数）和规模（标量参数数量）。

定义 H.8（深度和权重数量）。设架构如定义H.7所述。其深度为层数 L （特别地，具有一个隐藏层的网络深度为3）。令 E 表示有向边（连接）的集合，并定义计算单元数量

$$U \;:=\; \sum_{\ell=1}^{L-1} |V_\ell|$$

（隐藏层计算单元加上输出单元）。权重数量（网络规模）为

$$W \;:=\; |E| + U,$$

即每条连接对应一个标量参数，每个计算单元对应一个标量偏置。我们用 $\mathcal{NN}^{\sigma}_{n,\ L,\ W}$ 表示由深度至多为 L 、且权重数量至多为 W （按上述含义）的架构所能表示的所有ˆ函数 $f\colon \mathbb{R}^n \to \mathbb{R}$ 。

定义H.8 的关键在于，复杂度是根据图的内在结构来衡量的：每条有向边贡献一个标量权重，每个计算单元贡献一个标量偏置，因此 $W = |E| + U$。这与传统的稠密矩阵参数化形成对比，后者会统计每个层矩阵中的所有元素，包括稀疏图中结构上不存在的边所对应的系数。最后请注意，按照此约定，深度也计入输入层和输出层；因此，“$L = 3$”对应一个隐藏层。下文中，每当我们引用 Yarotsky & Zhevnerchuk(2020年) 时，符号 L和 W 均指此约定。

我们的分析仅限于分段线性激活函数。也就是说，我们考虑 $\sigma \in C(\mathbb{R}, \mathbb{R})$，其中存在$B \in \mathbb{N}$ 以及两两不同的断点 $x_1, \dots, x_B \in \mathbb{R}$，使得每个点 $x \in \mathbb{R} - \{x_b\}_{b=1}^{B}$ 都落在某个开区间内，且 σ 在该区间上是仿射的；而在任何断点 x_b 处（对于 $b = 1, \dots, B$），都不存在这样的仿射邻域。特别地，如果 σ是分段线性且非仿射的，那么必有 $B \geq 1$。该类包括 Fukushima （ 1969 ）的 ReLU 激活函数、 Maas 等（ 2013 ）的 Leaky-ReLU 激活函数、 He 等（ 2015 ）的 pReLU 激活函数，以及 Heaviside 函数的常用分段线性近似（例如 Abadi 等（2015）中实现的近似）。

证明定理 5.1 的一个关键要素，是将 Yarotsky & Zhevnerchuk(2020年，定理 3.3）（“深度不连续阶段”）直接特化到 H¨older 空间 $\mathcal{H}_{r,\ n}$ （见（16））。我们将其表述为一种便于学习理论使用的 ε-形式。

定理 H.9 （Yarotsky & Zhevnerchuk（2020年）中定理 3.3 对分段线性激活函数的实例化）。 固定输入维度 $n \in \mathbb{N}$ 和光滑性级别 $r > 0$。令 β 满足

$$\frac{r}{n} < \beta \leq \frac{2r}{n}.$$

~令 σ： $\mathbb{R} \to \mathbb{R}$ 为任意非仿射分段线性激活函数。则存在常数 $C_{r,\ n,\ \beta,\ \sigma} > 0$ 和$C_{r,\ n,\ \sigma} > 0$，使得对于每个 $\varepsilon \in$ （0,1）以及每个 $f \in \mathcal{H}_{r,\ n}$，都存在整数 $W_\varepsilon \geq 1$ 和 $L_\varepsilon \geq 2$ 以及一个神经网络 $\hat{f}_\varepsilon \in \mathcal{NN}^{\sigma}_{n,\ L_\varepsilon,\ W_\varepsilon}$，使得

$$\|f - \hat{f}_\varepsilon\|_\infty \leq \varepsilon,$$

and

$$W_\varepsilon \leq \left(\frac{\widetilde{C}_{r,n,\beta,\sigma}}{\varepsilon}\right)^{1/\beta}, \qquad L_\varepsilon \leq C_{r,n,\sigma}\, W_\varepsilon^{\beta n/r - 1}.$$

定理的证明H.9 推迟到附录H.12。

现在我们利用定理 H.9 来证明定理5.1。为便于阅读，我们将证明分解为若干引理 s.

证明策略。 证明通过将 p维流形 $\mathcal{M}$ 上的近似问题化为2p维欧几里得立方体上的近似问题来展开。该化约包含三个概念性步骤：(i) 通过固定的光滑特征映射 $\varphi_\star$将 $\mathcal{M}$ 嵌入 $[0,1]^{2p}$；(ii) 将 f 传输到嵌入副本 $K' = \varphi_\star(\mathcal{M})$，随后将其延拓为立方体上全局定义的 H¨lder 函数；(iii) 在 $[0,1]^{2p}$ 上应用欧几里得逼近定理，并通过与 $\varphi_\star$复合，将得到的网络拉回到 $\mathcal{M}$ 上。我们现在用单独的引理形式化每个步骤。

引理 H.10 (嵌入与归一化到开立方体)。 设 $\mathcal{M}$ 是一个维度为 p的紧致光滑流形。那么存在

- 光滑嵌入 Φ： $\mathcal{M} \to \mathbb{R}^{2p}$,
- 可逆仿射映射 A： $\mathbb{R}^{2p} \to \mathbb{R}^{2p}$,

使得特征映射 $\varphi_\star$： $= A \circ \Phi$ 是一个光滑嵌入，其像为 K'： $= \varphi_\star(\mathcal{M}) \subset (0,1)^{2p}$，且 K' 是紧致的 .

证明。 根据强惠特尼嵌入定理（参见（Lee， 2012，定理 6.20））， 存在一个光滑嵌入Φ： $\mathcal{M} \to \mathbb{R}^{2p}$。由于 Φ 是光滑的，因此它是连续的；又由于 $\mathcal{M}$ 是紧致的，其像

$$K := \Phi(\mathcal{M}) \subset \mathbb{R}^{2p}$$

由于它是紧集的连续像，因此也是紧致的。特别地， K 是有界的（例如，由 $\mathbb{R}^{2p}$中的海涅–博雷尔定理可知）。取 $c \in \mathbb{R}^{2p}$ 并定义

$$R := \sup_{z \in K} \|z - c\|_2 < \infty.$$

等价地，由于 $z \mapsto \|z - c\|_2$ 连续且 K 紧致，根据极值定理，上确界实际上是最大值，因此 $R = \max_{z \in K} \|z - c\|_2$。

若 $R = 0$，则对每个 $z \in K$都有 $\|z - c\|_2 = 0$，从而 $K = \{c\}$。由于 Φ 是嵌入，特别地它是单射，故$\Phi(\mathcal{M}) = \{c\}$蕴含 $\mathcal{M}$ 为单点集，该结论是平凡的。假设 $R > 0$。

定义仿射映射 B： $\mathbb{R}^{2p} \to \mathbb{R}^{2p}$ 为

$$B(z) := \frac{z - c}{2R} + \frac{1}{2}\mathbf{1}.$$

此处线性部分为 $\frac{1}{2R} I$，故 B可逆，其逆为 $B^{-1}(u) = 2R\left(u - \frac{1}{2}\mathbf{1}\right) + c$。于是对每个 $z \in K$，我们有 $\|z - c\|_\infty \leq \|z - c\|_2 \leq R$。因此对每个坐标 j， $|(z - c)_j| \leq R$，所以

$$-\frac{1}{2} \leq \frac{(z-c)_j}{2R} \leq \frac{1}{2} \quad \Longrightarrow \quad 0 \leq \frac{(z-c)_j}{2R} + \frac{1}{2} \leq 1,$$

即 $B(z) \in [0,1]^{2p}$。因此

$$B(K) \subset [0,1]^{2p}.$$

为了得到严格包含关系 n 到开立方体中，任取 $\eta \in (0,\ 1/2)$，并定义仿射压缩映射 S_η： $\mathbb{R}^{2p} \to \mathbb{R}^{2p}$ by

$$S_\eta(u) := (1 - 2\eta)u + \eta\mathbf{1}.$$

由于 $1-2\eta>0$，其线性部分（$1-2\eta$）I 是可逆的，因此 S_η 可逆，其逆为

$$S_\eta^{-1}(v)=\frac{v-\eta\mathbf{1}}{1-2\eta}.$$

此外，对于每个 $u\in[0,1]^{2p}$ 以及每个坐标 $j=1,\ldots\ldots,\ 2p$,

$$0\le u_j\le 1 \quad\Longrightarrow\quad \eta\le(1-2\eta)u_j+\eta\le 1-\eta,$$

因此 $S_\eta(u)\in[\eta,1-\eta]^{2p}\subset(0,\ 1)^{2p}$。特别地,

$$S_\eta([0,1]^{2p})\subset[\eta,1-\eta]^{2p}\subset(0,1)^{2p}.$$

现在令

$$A:=S_\eta\circ B,\qquad \varphi_\star:=A\circ\Phi,\qquad K':=\varphi_\star(\mathcal{M})=A(K).$$

那么 A 是一个可逆仿射映射（由可逆仿射映射复合而成）。由于 A 是 $\mathbb{R}^{2p}$ 的微分同胚，且 Φ 是一个嵌入，因此 $\varphi_\star=A\circ\Phi$ 仍是一个光滑嵌入（嵌入与微分同胚复合后，仍保持单射性、浸入性以及到其像上的同胚性）。最后,

$$K'=A(K)=S_\eta(B(K))\subset S_\eta([0,1]^{2p})\subset[\eta,1-\eta]^{2p}\subset(0,1)^{2p}.$$

此外，K' 是紧致的，因为它是紧集 $\mathcal{M}$ 的连续像（等价地，$K'=A(K)$ 是紧致的，因为它是紧集 K 的连续像）。□

步骤 *1*：固定一个欧几里得模型 $\mathcal{M}$。引理 H.10 构造了一个光滑嵌入 $\varphi_\star:\ \mathcal{M}\to[0,1]^{2p}$，使得后续所有近似都可以在固定的立方体 $[0,1]^{2p}$上进行。记 $K':=\varphi_\star(\mathcal{M})$ 为该嵌入副本。下一个引理将目标函数 f 拉回到 K'，并比较 H¨older 范数。

引理 H.11（拉回 a nd H¨older 范数控制）。令 $\varphi_\star:\ \mathcal{M}\to K'$ 如引理 *H.10* 中所述。给定 $f\in C^{k,\ 1}(\mathcal{M})$，定义 e

$$h:K'\to\mathbb{R},\qquad h(u):=f(\varphi_\star^{-1}(u)).$$

则 $h\in C^{k,\ 1}(K')$。此外，存在一个常数 $C_{\mathrm{pb}}=C_{\mathrm{pb}}(\mathcal{M},\varphi_\star,k)$ 使得

$$\|h\|_{C^{k,1}(K')}\le C_{\mathrm{pb}}\|f\|_{C^{k,1}(\mathcal{M})}.$$

特别地，如果 $\|f\|_{C^{k,\ 1}(\mathcal{M})}\le 1$，那么 $\|h\|_{C^{k,\ 1}(K')}\le C_{\mathrm{pb}}$。

证明。由于 $\varphi_\star$ 是光滑嵌入，它在像 $K'=\varphi_\star(\mathcal{M})$ 上诱导出唯一的光滑结构，使得K' 成为 $\mathbb{R}^{2p}$的光滑嵌入子流形。逆映射 $\varphi_\star^{-1}\colon K'\to\mathcal{M}$ 是光滑的。对于任意 $f\colon\ \mathcal{M}\to\mathbb{R}$，函数 $h=f\circ\varphi_\star^{-1}$是在 K'上定义良好的映射。

为比较范数，固定 $\mathcal{M}$ 上的一个有限 C^∞ 图册 $\{(U_i,\ \varphi_i)\}_{i=1}^N$（有限是因为 $\mathcal{M}$ 是紧致的），并通过 $V_i:=\Psi(U_i)$ 和 $\tilde{\varphi}_i:=\varphi_i\circ\Psi^{-1}\colon\ V_i\to\Omega_i:=\varphi_i(U_i)$ 在 K' 上定义诱导图册。于是 $\tilde{\varphi}_i^{-1}=(\varphi_i\circ\Psi^{-1})^{-1}=\Psi\circ\varphi_i^{-1}$，并且

$$h\circ\tilde{\varphi}_i^{-1}=f\circ\Psi^{-1}\circ(\varphi_i\circ\Psi^{-1})^{-1}=f\circ\Psi^{-1}\circ(\Psi\circ\varphi_i^{-1})=f\circ\varphi_i^{-1}.$$

特别地，若令 $f_i:=f\circ\varphi_i^{-1}$ 且 $h_i:=h\circ\tilde{\varphi}_i^{-1}$，则在 Ω_i上逐点成立 $h_i=f_i$。由于 $f\in C^{k,\ 1}(\mathcal{M})$，根据定义$f_i\in C^{k,\ 1}(\Omega_i)$，即对于每个满足 $|\alpha|\le k$ 的多重指标 α，欧几里得导数 $D^\alpha f_i$ 在 Ω_i上存在且连续，并且对于 $|\alpha|=k$，其 Lipschitz 半范数

$$[D^\alpha f_i]_{\mathrm{Lip}(\Omega_i)}:=\sup_{x\neq y\in\Omega_i}\frac{|D^\alpha f_i(x)-D^\alpha f_i(y)|}{\|x-y\|}$$

是有限的。由于 $h_i=f_i$，t 导数在存在的地方相等，因此对于所有 $|\alpha|\le k$，我们都有 $D^\alpha h_i=D^\alpha f_i$，a nd尤其是对于 $|\alpha|=k$,

$$[D^\alpha h_i]_{\mathrm{Lip}(\Omega_i)}=[D^\alpha f_i]_{\mathrm{Lip}(\Omega_i)}<\infty.$$

因此 $h_i \in C^{k,1}(\Omega_i)$对每一个 i成立，而这正是 $h \in C^{k,1}(K')$。

最后，由于 K' 是紧致的，$\|\cdot\|_{C^{k,1}(K')}$ 的任意两种基于图册的定义都是等价的，因此存在常数 $C_{\mathrm{pb}} = C_{\mathrm{pb}}(\mathcal{M}, \Psi, k)$，使得

$$\|h\|_{C^{k,1}(K')} \le C_{\mathrm{pb}} \max_{1\le i\le N} \|h \circ \tilde{\varphi}_i^{-1}\|_{C^{k,1}(\Omega_i)} = C_{\mathrm{pb}} \|f\|_{C^{k,1}(\mathcal{M})}.$$

□

步骤 *2*：在嵌入副本 K' 上工作。给定 $f \in C^{k,1}(\mathcal{M})$，我们设 $h := f \circ \Psi^{-1}$ 在 K' 上。引理 H.11 表明$h \in C^{k,1}(K')$，并给出了$\|h\|_{C^{k,1}(K')}$ 关于$\|f\|_{C^{k,1}(\mathcal{M})}$的统一界。不过，要应用欧几里得逼近定理，我们还必须将 h扩展为定义在 K'的某个开邻域上的函数，并最终扩展到整个立方体上。这可以通过围绕 K'构造管状邻域来实现。

引理 H.12 (立方体内的有界光滑区域)。 设 $K' \Subset (0,1)^{2p}$ 是$\mathbb{R}^{2p}$中的一个紧致嵌入 C^∞ 子流形。那么存在一个半径 $\rho > 0$ 和一个开集$\Omega \subset \mathbb{R}^{2p}$，使得

$$K' \Subset \Omega \Subset (0,1)^{2p},$$

Ω 有界，且 $\partial\Omega$ 是一个 C^∞ 超曲面（特别地，Ω对每个 $k \ge 1$都是一个 $C^{k,1}$ 区域）。

证明。 对每个 $x \in K'$，定义 K' 在 x 处的法空间为由所有相对于欧几里得点积与 T_xK' 正交的向量构成的 p 维子空间 $N_xK' \subseteq T_x\mathbb{R}^{2p}$。此外，我们将 K' 的法丛定义为

$$NK' = \{(x, v) \in \mathbb{R}^{2p} \times \mathbb{R}^{2p} : x \in K', v \in N_xK'\}.$$

根据管状邻域定理（Lee，2012，定理 6.24），存在零截面$\mathcal{U}$的一个开邻域$K_0 = \{(x,\ 0)\colon\ x \in K'\} \subseteq NK'$，以及 U 的一个开邻域 K'，它位于 $\mathbb{R}^{2p}$ 中，并存在一个光滑微分同胚

$$F : \mathcal{U} \to U \subset \mathbb{R}^{2p}$$

使得 $F(x,\ 0) = x$ 对所有 $x \in K'$均成立。

步骤 1：找到一个统一半径 $\rho_0 > 0$，使得 ρ_0-圆盘丛包含于 $\mathcal{U}$。 由于 $\mathcal{U}$ 是开集且包含零截面，对每个 $x \in K'$，点 $(x,\ 0) \in NK'$都属于 $\mathcal{U}$。由于 $\mathcal{U}$是开集，存在 $(x,\ 0)$ 的一个开邻域 $W_x \subset NK'$，使得 $W_x \subset \mathcal{U}$。根据法丛的局部平凡性，我们可以选取形如以下形式的 W_x

$$W_x \supset \{(y, v) : y \in O_x, \|v\| < \rho_x\}$$

对于某个 $O_x \subset K'$ 的开邻域 x 以及某个 $\rho_x > 0$。因此

$$\{(y, v) : y \in O_x,\ \|v\| < \rho_x\} \subset \mathcal{U}.$$

族 $\{O_x\}_{x\in K'}$是 K'的一个开覆盖。由于 K'是紧致的，存在有限个点 x_1、.....、 $x_m \in K'$，使得

$$K' \subset \bigcup_{j=1}^{m} O_{x_j}.$$

定义 $\rho_0 :=$ 的最小值$1\le j\le m\ \rho_{x_j}$。由于有限个正数的最小值为正，因此有 $\rho_0 > 0$。现在任取满足 $\|v\| < \rho_0$的 $(x, v) \in NK'$。由于 $\{O_{x_j}\}$ 覆盖 K'，因此存在某个 j 满足 $x \in O_{x_j}$。于是 $\|v\| < \rho \le \rho_{x_j}$，所以

$$(x, v) \in \{(y, w) : y \in O_{x_j}, \|w\| < \rho_{x_j}\} \subset \mathcal{U}.$$

因此， ρ_0-圆盘丛

$$D_{\rho_0} := \{(x, v) \in NK' : \|v\| < \rho_0\}$$

满足

$$D_{\rho_0} \subset \mathcal{U}.$$

步骤 2：利用假设 $K' \Subset (0,1)^{2p}$ 使管状邻域保持在立方体内。 </style}记号 $(0,1)^{2p}$ 表示 K'是紧致的，并且包含在开集 $(0,1)^{2p}$中。等价地 $K' \cap \partial[0,1]^d = \varnothing$。考虑 K'到最近集合 $\partial[0,1]^{2p}$的距离：

$$\delta := \operatorname{dist}(K', \partial[0,1]^{2p}) := \inf\{\|x-y\| : x \in K', y \in \partial[0,1]^{2p}\}$$

我们断言 $\delta > 0$。事实上，函数 $x \mapsto \operatorname{dist}(x, \partial[0,1]^{2p})$ 在 $\mathbb{R}^{2p}$ 上连续，因此根据极值定理，它在紧集 K' 上的限制可取得最小值。由于 $K' \subset (0,1)^{2p}$ 且 $\partial[0,1]^{2p}$ 是立方体的边界，每个 $x \in K'$ 都满足 $\operatorname{dist}(x, \partial[0,1]^{2p}) > 0$。因此，最小值严格为正，从而 $\delta > 0$。

现在选取 $\rho := \min\{\rho_0/2, \delta/2\}$ 并定义

$$D_\rho := \{(x,v) \in NK' : \|v\| < \rho\}, \qquad \Omega := F(D_\rho) \subset \mathbb{R}^d.$$

步骤 3：Ω 是开集且包含 K'。 由于 D_ρ 在 NK' 中是开集（它由严格不等式 $\|v\| < \rho$ 定义），且 F 是一个微分同胚，因此

$$\Omega = F(D_\rho)$$

在 $\mathbb{R}^{2p}$ 中是开集。此外，$K' \subset \Omega$：如果 $x \in K'$，那么 $(x,0) \in D_\rho$，因此

$$x = F(x,0) \in F(D_\rho) = \Omega.$$

因此 $K' \subset \Omega$。然后 Ω 是开集（开集在微分同胚下的像），并且 $K' \subset \Omega$，因为 $(x,0) \in D_\rho$ 和$F(x,0) = x$。

步骤 4：$\Omega \Subset (0,1)^{2p}$，因此 Ω 有界。 我们证明 $\Omega \subset (0,1)^{2p}$。设 $z \in \Omega$。则 $z = F(x,v)$ 对某个 $(x,v) \in D_\rho$，因此 $\|v\| < \rho \le \delta/2$。对于 $\mathbb{R}^{2p}$ 中的管状邻域，其几何解释是，$F(x,v)$ 可通过从 $x \in K'$ 沿法向方向移动长度 $\|v\|$得到；特别地，从 z 到 x 的欧几里得距离为$\|v\|$，因此

$$||z-x||_2 \le \|v\| < \rho \le \delta/2.$$

现在反设 $z \notin (0,1)^{2p}$。由于 $(0,1)^{2p}$ 是 $[0,1]^{2p}$ 的内部，这意味着 $z \in \mathbb{R}^{2p} \setminus (0,1)^{2p}$。因此，从 $x \in (0,1)^{2p}$ 到 $z \notin (0,1)^{2p}$ 的线段必定穿过边界 $\partial[0,1]^{2p}$。特别地，

$$\operatorname{dist}(x, \partial[0,1]^{2p}) \le \|z-x\|_2.$$

但根据 δ 的定义，对于每个 $x \in K'$，$\operatorname{dist}(x, \partial[0,1]^{2p}) \ge \delta$，因此我们得到

$$\delta \le \|z-x\|_2 < \delta/2,$$

这就产生了矛盾。因此 $z \in (0,1)^{2p}$。这表明 $\overline{\Omega} \subset (0,1)^{2p}$。要证明 $\Omega \Subset (0,1)^{2p}$，还需证明 $\Omega \subset (0,1)^{2p}$。由于 $\|v\| \le \rho$ 蕴含 $(x,v) \in D_\rho$，其闭包满足

$$\overline{\Omega} = \overline{F(D_\rho)} = F(\overline{D_\rho}),$$

因为 F 是一个同胚，因此它将闭包映射到闭包。现在，同样的距离论证 $\|v\| \le \rho \le \delta/2$表明，$F(D_\rho)$中的每个点到某个 $x \in K'$的距离仍至多为 ρ，因此无法到达 $\partial[0,1]^{2p}$。于是 $\Omega \subset (0,1)^{2p}$，即 $\Omega \Subset (0,1)^{2p}$。由于 $(0,1)^{2p}$ 有界且 $\Omega \subset (0,1)^{2p}$，因此 $\overline{\Omega}$ 有界是显然的。

步骤 5：$\partial\Omega$ 是一个 $C^{k,1}$ 区域。 最后，考虑球面丛

$$S_\rho := \{(x,v) \in NK' : \|v\| = \rho\}.$$

我们声称 S_ρ是 NK'中的一个 C^∞超曲面。为此，首先回顾一下，NK'是一个光滑向量丛（因为 $K' \subset \mathbb{R}^d$是一个嵌入的 C^∞ 子流形），因此每个局部平凡化

$$\tau_O : NK'|_O \to O \times \mathbb{R}^p$$

是一个 C^∞ 微分同胚。这里 $O \times \mathbb{R}^p$ 赋予其标准乘积光滑结构，即由乘积图$(\varphi \times \mathrm{id})(U \times \mathbb{R}^p)$ 生成的结构，其中(U,φ) 是 O上的一个坐标图。因此，关于 $NK'|_O$子集的任何光滑性陈述，都可以在乘积坐标 $O \times \mathbb{R}^p$ (Lee，2012，第 10 章）中验证。

在开集 $O \subset K'$ 上固定这样一个平凡化。在这些坐标中定义

$$G : O \times \mathbb{R}^p \to \mathbb{R}, \qquad G(y, w) = \|w\|^2.$$

映射 G 是 C^∞ （实际上是纤维坐标的多项式），并且

$$S_\rho \cap (O \times \mathbb{R}^p) = \{(y, w) : \|w\| = \rho\} = \{(y, w) : \|w\|^2 = \rho^2\} = G^{-1}(\rho^2).$$

现在任取一个 $(y, w) \in G^{-1}(\rho^2)$。于是 $|w| = \rho > 0$，因此 $w \neq 0$。G 在 (y, w) 处的微分是线性映射

$$dG_{(y,w)} : T_{(y,w)}(O \times \mathbb{R}^p) \to \mathbb{R}.$$

利用典范等同 $T_{(y,w)}(O \times \mathbb{R}^p) \simeq T_y O \times T_w \mathbb{R}^p$ 和 $T_w \mathbb{R}^p \simeq \mathbb{R}^p$，我们可以将 (y, w) 处的切向量视为一对 $(\dot{y}, \dot{w}) \in T_y O \times \mathbb{R}^p$。由于 $G(y, w) = |w|^2$ 不依赖于 y，其微分不含 $\dot{y}$ 项，而沿纤维方向求导可得

$$dG_{(y,w)}(\dot{y}, \dot{w}) = \left.\frac{d}{dt}\right|_{t=0} G(y, w + t\dot{w}) = \left.\frac{d}{dt}\right|_{t=0} |w + t\dot{w}|^2 = 2\langle w, \dot{w}\rangle.$$

选取容许切方向 $(\dot{y}, \dot{w}) = (0, w) \in T_y O \times \mathbb{R}^p$，得到

$$dG_{(y,w)}(0, w) = 2\langle w, w\rangle = 2\|w\|^2 = 2\rho^2 > 0,$$

因此，$dG_{(y,w)}$ 不是零映射。由于目标是 $\mathbb{R}$，这等价于 $dG_{(y,w)}$的满射性。因此，ρ^2 是 G的正则值，根据正则水平集定理（原像定理），$G^{-1}(\rho^2)$ 是一个 C^∞ 余维为 1 的$O \times \mathbb{R}^p$子流形，也就是一个 C^∞ 超曲面（Lee，2012，推论5.14）。

由于上述描述在每个丛坐标图中都成立，下面我们证明这意味着 S_ρ 是 NK'中的一个全局 C^∞ 超曲面。设 $\{(O_i, \tau_i)\}_{i \in I}$是 NK'的一个光滑丛图册，其中

$$\tau_i : NK'|_{O_i} \longrightarrow O_i \times \mathbb{R}^p$$

是一个到其像上的 C^∞ 微分同胚，并且转移映射

$$\tau_{ij} := \tau_j \circ \tau_i^{-1} : (O_i \cap O_j) \times \mathbb{R}^p \longrightarrow (O_i \cap O_j) \times \mathbb{R}^p$$

是 C^∞ 且逐纤维线性的。在坐标图 τ_i中，球面丛表示为

$$\tau_i(S_\rho \cap NK'|_{O_i}) = \{(y, w) \in O_i \times \mathbb{R}^p : \ \|w\| = \rho\}.$$

根据 $G(y, w) = \|w\|^2$，由正则水平集定理可知，该集合是 $O_i \times \mathbb{R}^p$中的一个 C^∞ 超曲面。现在考虑重叠部分 $O_i \cap O_j \neq \varnothing$。由于$\tau_{ij}$ 是一个微分同胚，它将 C^∞ 超曲面映射为 C^∞ 超曲面。此外，由于$\tau_{ij} = \tau_j \circ \tau_i^{-1}$，我们在重叠部分上有如下集合恒等式：

$$\tau_{ij}\Big(\tau_i(S_\rho \cap NK'|_{O_i \cap O_j})\Big) = \tau_j(S_\rho \cap NK'|_{O_i \cap O_j}).$$

因此，在 τ_i 和τ_j 中得到的局部超曲面坐标图通过光滑转移映射τ_{ij}在重叠部分上彼此一致。因此，族 $\{S_\rho \cap NK'|_{O_i}\}_{i \in I}$定义了一个全局良定义的嵌入式 C^∞余维为1的 NK'子流形，即S_ρ 是 C^∞ 中的一个全局 NK' 超曲面。（Lee，2012，例如参见光滑向量丛、丛图册和光滑转移映射）

我们断言 $\partial D_\rho = S_\rho$。事实上，纤维范数 $(x, v) \mapsto \|v\|$ 是连续的，因此

$$\overline{D_\rho} = \{(x, v) : \|v\| \leq \rho\}.$$

若 $\|v\| < \rho$，则 $(x, v) \in D_\rho$（因此不在边界上）；而若 $\|v\| > \rho$，则 $(x, v) \notin D_\rho$（因此不在边界上）。因此，任一边界点都必须满足 $\|v\| = \rho$，故 $\partial D_\rho \subset S_\rho$。反之，若 $\|v\| = \rho$，则 (x, v) 在NK' 中的每个邻域都包含满足 $\|v\| < \rho$的点以及满足 $\|v\| > \rho$的点，故 $(x, v) \in \partial D_\rho$。因此 $\partial D_\rho = S_\rho$。

由于 $F\colon\ \mathcal{U}\to U$ 是微分同胚，且我们选取 $\rho<\rho_0$ 使得 $D_\rho\subset\overline{D_{\rho_0}}\subset\mathcal{U}$，因此 F 在 $\mathcal{U}$ 上的限制为同胚（特别是在 D_ρ的某个邻域上）。同胚将边界映射为边界：对任意$A\subset\mathcal{U}$,

$$F(\partial A)=\partial(F(A)),$$

因为它保持闭包和内部不变。将其应用于 $A=D_\rho$ 和 $\Omega:=F(D_\rho)$，可得

$$\partial\Omega=\partial(F(D_\rho))=F(\partial D_\rho)=F(S_\rho).$$

最后，由于 S_ρ是 C^∞中的一个超曲面，且 F是 $\mathcal{U}$上的微分同胚，因此限制映射 $F|_{S_\rho}\colon\ S_\rho\to F(S_\rho)=\partial\Omega$ 是一个到其像上的微分同胚。因此， $\partial\Omega$ 是 $\mathbb{R}^{2p}$中的一个超曲面。

为了将此与定义 C.1 联系起来，固定 $z_\star\in\partial\Omega$。由于 $\partial\Omega$ 是 $\mathbb{R}^{2p}$中的一个超曲面，因此它特别是一个 $(2p-1)$ 维嵌入子流形。由此，切空间 $T_{z_\star}(\partial\Omega)\ \subset\ \mathbb{R}^{2p}$ 是一个 $(2p-1)$ 维线性子空间。将其与 $\mathbb{R}^{2p}$ 的一个刚体运动复合（先平移 $-z_\star$，再进行正交旋转）；该刚体运动是一个具有 C^∞ 逆映射的微分同胚。因此，不失一般性，我们可以假设 $z_\star=0$，且 $0\in\partial\Omega$ 处的切超平面是标准的“水平”超平面，即最后一个坐标为零，也就是说

$$T_0(\partial\Omega)=\{x_{2p}=0\}.$$

由于 $\partial\Omega$ 是一个嵌入的 C^∞ 超曲面，因此它局部上是一个正则水平集：根据（Lee，2012，命题 5.16），存在一个包含 U 的邻域 0 以及一个 C^∞ 浸没 $\Phi\colon\ U\to\mathbb{R}$使得

$$\partial\Omega\cap U=\Phi^{-1}(0)=\{x\in U\colon\ \Phi(x)=0\}.$$

特别地，由于 Φ 是一个浸没， $d\Phi_0\neq0$，用欧几里得坐标表示时等价于 $\nabla\Phi(0)\neq0$。

Mor此外，由于 $\partial\Omega\cap U=\{\Phi=0\}$ 且 $\nabla\Phi(0)\neq0$，0 处水平集的切空间满足 es

$$T_0(\partial\Omega)=\ker(d\Phi_0)=\big\{v\in\mathbb{R}^{2p}\colon\ d\Phi_0(v)=0\big\}=\big\{v\in\mathbb{R}^{2p}\colon\ \nabla\Phi(0)\cdot v=0\big\},$$

即 $T_0(\partial\Omega)$ 是与 $\nabla\Phi(0)$ 正交的超平面。假设 $T_0(\partial\Omega)\ =\{v\in\mathbb{R}^{2p}\colon\ \ v_{2p}=0\}$ 因此意味着

$$\nabla\Phi(0)\in\big(T_0(\partial\Omega)\big)^\perp=\operatorname{span}\{e_{2p}\},$$

其中 $e_{2p}=(0,\ \ldots\ldots,\ 0,\ 1)$ 是 x_{2p} 轴方向。因此存在 $\lambda\neq0$ 使得

$$\nabla\Phi(0)=\lambda e_{2p},$$

特别是，梯度的最后一个分量非零:

$$\partial_{x_{2p}}\Phi(0)=(\nabla\Phi(0))_{2p}=\lambda\neq0.$$

由 $\partial_{x_{2p}}\Phi$ 的连续性可知，我们可以（必要时缩小 U）假设对所有 $x\in U$， $\partial_{x_{2p}}\Phi(x)\ \neq\ 0$，这正是应用隐函数定理并将 $\Phi(x',\ x_{2p})=0$ 作为 x' 的函数求解 x_{2p} 所需的条件。

对方程 $\Phi(x',\ x_{2p})\ =\ 0$ 应用隐函数定理，可知存在 $r\ >\ 0$ 以及一个 C^∞ 函数 $g:B^{2p-1}(0,\ r)\ \to\mathbb{R}$，使得

$$\partial\Omega\cap B(0,r)=\{(x',x_{2p})\in B(0,r)\colon\ x_{2p}=g(x')\}.$$

接下来，令 $H(x',\ x_{2p}):=x_{2p}-g(x')$。由于 g 连续， H 连续，因此集合

$$\begin{aligned}U_+&:=\{(x',x_{2p})\in B(0,r)\colon\ H(x',x_{2p})>0\}=\{x_{2p}>g(x')\}\cap B(0,r),\\U_-&:=\{(x',x_{2p})\in B(0,r)\colon\ H(x',x_{2p})<0\}=\{x_{2p}<g(x')\}\cap B(0,r)\end{aligned}$$

在 $B(0,\ r)$ 中是开集。此外， $U_+\cap U_-=\varnothing$，而 $B(0,\ r)\ \setminus\partial\Omega$ 中的每一点都满足 $x_{2p}\neq g(x')$，因此恰好属于 U_+ 或 U_- 中的一个。因此

$$B(0,r)\setminus\partial\Omega=U_+\dot\cup U_-.$$

由于 Ω 是一个区域（开且连通），交集 $\Omega \cap B(0,r)$ 是开集；如有必要，通过缩小 r，我们可以假设 $\Omega \cap B(0,r)$ 是连通的。[1] 由于 $\partial\Omega \cap B(0,r)$ 将该球分成恰好两个连通分支，连通集 $\Omega \cap B(0,r)$ 必须恰好包含于其中一个分支中。如有必要，将 g 替换为 $-g$（等价于交换两侧的标签），我们可以假设

$$\Omega \cap B(0,r) \subset U_+, \qquad \text{i.e.} \qquad \Omega \cap B(0,r) \subset \{(x', x_{2p}) \in B(0,r) : x_{2p} > g(x')\}.$$

定义边界拉直映射

$$\psi : B(0,r) \to \psi(B(0,r)), \qquad \psi(x', x_{2p}) := (x', x_{2p} - g(x')).$$

由于 $g \in C^\infty$，映射 ψ 是 C^∞。其逆映射可显式写为

$$\psi^{-1}(y', y_{2p}) = (y', y_{2p} + g(y')),$$

它也具有 C^∞。等价地，ψ 的雅可比矩阵为

$$D\psi(x) = \begin{pmatrix} I_{2p-1} & 0 \\ -\nabla g(x')^\top & 1 \end{pmatrix}, \qquad \text{so} \qquad \det D\psi(x) = 1,$$

因此，ψ 是从 $B(0,r)$ 到其像 $\psi(B(0,r))$ 的 C^∞ 微分同胚，其逆映射为 $\psi^{-1}(y', y_{2p}) = (y', y_{2p} + g(y'))$。特别地，对每个 $k \geq 1$，都有 $\psi, \psi^{-1} \in C^\infty \subset C^{k,1}$。

由上述单侧包含关系可知，若 $(x', x_{2p}) \in \Omega \cap B(0,r)$，则 $x_{2p} > g(x')$，因此 $\psi(x', x_{2p})$ 的最后一个坐标为 $x_{2p} - g(x') > 0$。因此

$$\psi\big(B(0,r) \cap \Omega\big) \subset \mathbb{R}^{2p}_+.$$

另一方面，如果 $(x', x_{2p}) \in \partial\Omega \cap B(0, r)$，那么 $x_{2p} = g(x')$，因此 $\psi(x', x_{2p})$ 的最后一个坐标为 0，从而

$$\psi\big(B(0,r) \cap \partial\Omega\big) \subset \partial\mathbb{R}^{2p}_+.$$

这验证了 定义 C.1 中的边界拉直性质。因此，对于每个 $k \geq 1$，Ω 都是一个 $C^{k,1}$ 区域。

□

步骤 *3*：在 K' 周围构造一个正则区域。引理 H.12 构造了一个有界的 $C^{k,1}$ 区域 Ω，使得 $K' \Subset \Omega \Subset (0,1)^d$。这提供了一个适用经典延拓结果的环境，可将 Hölder 函数从具有正则边界的区域上延拓。在下一个引理中，我们将管状收缩映射 π（以在 Ω 上定义一个与 h 在 K' 上一致的函数）与延拓引理 C.2 结合起来，从而在整个立方体上得到一个函数。

引理 H.13（从 K' 延拓到立方体，并控制 $C^{k,1}$ 范数）。设 $K' \Subset (0,1)^{2p}$ 是一个紧致嵌入的 C^∞ 子流形，并设 $h \in C^{k,1}(K')$。则存在 $H \in C^{k,1}([0,1]^{2p})$，使得 $H = h$ 在 K' 上，并且

$$\|H\|_{C^{k,1}([0,1]^{2p})} \leq C_{\text{ext}} \|h\|_{C^{k,1}(K')},$$

对于某个常数 $C_{\text{ext}} = C_{\text{ext}}(K', k)$。

证明。**步骤 1：构造管状邻域以及到 K' 上的光滑收缩映射。** 令 $F : \mathcal{U} \to U \subset \mathbb{R}^{2p}$ 是由管状邻域定理给出的管状邻域微分同胚，其定义在法丛 NK' 中零截面的一个开邻域 $\mathcal{U}$ 上，且满足 $F(x,0) = x$ 对所有 $x \in K'$ 均成立（参见（Lee，2012，定理 6.24））。与引理 H.12 中一样，选取 $\rho_0 > 0$ 使得

$$D_{\rho_0} := \{(x,v) \in NK' : \|v\| < \rho_0\} \subset \mathcal{U},$$

然后选取 $0 < \rho < \rho_0$ 使得 $D_\rho \subset \overline{D_{\rho_0}} \subset \mathcal{U}$（例如 $\rho := \rho_0/2$）。定义管状域

$$\Omega := F(D_\rho) \Subset (0,1)^{2p}.$$

[1] 例如，由于开连通集是局部道路连通的，因此可以将 $r > 0$ 选得足够小，使得 $\Omega \cap B(0, r)$ 中任意两点都能通过一条完全包含在 $\Omega \cap B(0, r)$ 中的路径连接起来。

根据构造， F和 F^{-1} 分别在 D_ρ 和 Ω 的开邻域上 C^∞ 。

令 $\mathrm{pr}: NK' \to K'$ 表示丛投影（$x,\ v$） $\mapsto x$，并定义

$$\pi := \mathrm{pr} \circ F^{-1} \quad \text{on an open neighborhood of } \overline{\Omega}.$$

则 π 是 C^∞，将 Ω 映射到 K'，并满足 $\pi|_{K'} = \mathrm{id}_{K'}$，因为对于 $x \in K'$，我们有 $F^{-1}(x) = (x,\ 0)$，因此 $\pi(x) = \mathrm{pr}(x,\ 0) = x$。

步骤 2：通过收缩映射将 h 从 $\overline{K'}$ 延拓到 Ω 。定义

$$u : \overline{\Omega} \to \mathbb{R}, \qquad u(z) := h(\pi(z)).$$

由于 $\pi|_{K'} = \mathrm{id}_{K'}$，因此 $u = h$ 在 K' 上成立。

我们声称 $u \in C^{k,\,1}(\overline{\Omega})$，并且

$$\|u\|_{C^{k,1}(\overline{\Omega})} \le C_0 \|h\|_{C^{k,1}(K')}$$

对某个常数 $C_0 = C_0(K',\ F,\ k)$。为此，通过 F 将 u 拉回到 D_ρ：对于 $(x,\ v) \in D_\rho$，

$$(u \circ F)(x,v) = h\big(\pi(F(x,v))\big) = h\big(\mathrm{pr}(x,v)\big) = h(x).$$

因此， $u \circ F$ 仅依赖于基变量 x ，并且沿每个纤维方向 v 保持为常数。

在 K' 上固定一个有限图册 $C^\infty(O_i,\ \varphi_i)\}_{i=1}^N$ 及其对应的局部平凡化 $NK'|_{O_i} \simeq O_i \times \mathbb{R}^p$。在诱导坐标 $(y,\ w) \in \varphi_i(O_i) \times B^p(0,\ \rho)$ 中，函数 $u \circ F$ 表示为

$$(u \circ F) \circ (\varphi_i^{-1} \times \mathrm{id})(y,w) = h \circ \varphi_i^{-1}(y),$$

它与 w无关。因此：

- w 变量的所有偏导数均为零；
- y 变量的偏导数在不超过 k 阶时，与 $h \circ \varphi_i^{-1}$ 的偏导数相同；
- k 阶导数的 Lipschitz 半范数（在 $(y,\ w)$ 中）由 y 中的 Lipschitz 半范数控制，因为不存在 w 依赖。

因此，对于每个坐标图 i，

$$\|(u \circ F) \circ (\varphi_i^{-1} \times \mathrm{id})\|_{C^{k,1}(\varphi_i(O_i) \times B^p(0,\rho))} \le \|h \circ \varphi_i^{-1}\|_{C^{k,1}(\varphi_i(O_i))}.$$

由于 $u = (u \circ F) \circ F^{-1}$ 且 F^{-1} 在 Ω 的某个邻域上是 C^∞，根据截至 k 阶的链式法则以及对 k 阶导数的 Lipschitz 控制，可得到所述估计，其中常数 C_0 取决于 F^{-1} 在 Ω 上截至 $k+1$ 阶导数的一致有界性。这证明了 $u \in C^{k,\,1}(\Omega)$ 以及所述界。

步骤 3：应用 $C^{k,\,1}$ 延拓定理并限制到立方体上。根据引理H.12，Ω 是有界的 C^∞ 区域，因此是有界的 $C^{k,\,1}$ 区域。选取 $\eta > 0$ 并令

$$\Omega' := (-\eta, 1+\eta)^{2p},$$

使得 $\overline{\Omega} \subset \Omega'$ 且 $[0,1]^{2p} \subset \Omega'$。应用引理 C.2 （其中 $\alpha = 1$ 和 $n = 2p$）到 $u \in C^{k,\,1}(\Omega)$，我们得到 $w \in C^{k,\,1}(\Omega')$ 使得 $w = u$ 在 Ω 上，并且

$$\|w\|_{C^{k,1}(\Omega')} \le C_1 \|u\|_{C^{k,1}(\overline{\Omega})}, \qquad C_1 = C_1(k,1,\Omega,\Omega').$$

最后定义 $H := w|_{[0,1]^{2p}}$。则有 $H \in C^{k,\,1}([0,1]^{2p})$，并且由于 $K' \subset \Omega$ 且 $w = u$ 在 Ω 上， $u = h$ 在 K' 上，因此我们有$H = h$ 在 K' 上。此外，限制操作不会增大 $C^{k,\,1}$ 范数，因此

$$\|H\|_{C^{k,1}([0,1]^{2p})} \le \|w\|_{C^{k,1}(\Omega')} \le C_1 \|u\|_{C^{k,1}(\overline{\Omega})} \le C_1 C_0 \|h\|_{C^{k,1}(K')}.$$

设定 $C_{\mathrm{ext}} := C_0 C_1$ 即可完成证明。 □

步骤 *4*：将问题约化为 $[0,1]^{2p}$ 上的欧几里得近似。引理 H.13 给出一个函数 $H \in C^{k,\,1}([0,1]^{2p})$，使得$H = h$ 在 K' 上成立，并且满足范数控制 $\|H\|_{C^{k,\,1}([0,1]^{2p})} \lesssim \|h\|_{C^{k,\,1}(K')}$。此时问题完全转化为欧几里得问题：在立方体上用 ReLU 网络一致近似 H。这正是定理 H.9的内容，只需先归一化到 H¨older 单位球。

引理 *H.14*（归一化后立方体上的 *Yarotsky* 近似）。设 $H \in C^{k,1}([0,1]^{2p})$，并令 $B := \sim \|H\|_{C^{k,1}([0,1]^{2p})}$。若 $B > 0$，定义 $H := H/B$。则 $\|H\|_{C^{k,1}([0,1]^{2p})} \leq 1$，且对每个 $0 < \varepsilon < 1$，定理 *H.9* 给出一个 *ReLU* 网络 $\widetilde{g}_\varepsilon$，使得

$$\|\widetilde{H} - \widetilde{g}_\varepsilon\|_\infty \leq \frac{\varepsilon}{B}.$$

因此，$g_\varepsilon := B\widetilde{g}_\varepsilon$ 满足

$$\|H - g_\varepsilon\|_\infty \leq \varepsilon,$$

其规模/深度界由定理 *H.9* 在精度 ε/B处取值得到。

˜证明。由缩放 $H = H/B$ 和定理 H.9可立即得出。 □

步骤 *5*：将网络拉回到 $\mathcal{M}$。引理 H.14 给出一个网络 g_ε，在 $[0,1]^{2p}$上一致逼近 H。由于 H在 K'上与 h 一致，且 h 是 f 经由 $\varphi_\star$的推前，我们只需通过复合，即可在 $\mathcal{M}$ 上得到一个逼近函数：$\widehat{g}_\varepsilon := g_\varepsilon \circ \varphi_\star$。最后一个引理说明，立方体上的一致逼近误差可直接推出在 $\mathcal{M}$上的一致逼近误差。

引理 H.15（将欧几里得近似函数拉回至 $\mathcal{M}$）。设 $\varphi_\star : \mathcal{M} \to K' \subset [0,1]^{2p}$ 如上，并设$H \in C^{k,\,1}([0,1]^{2p})$ 在 K' 上满足 $H = h$，其中 $h = f \circ \varphi_\star^{-1}$。若 g_ε 满足 $\|H - g_\varepsilon\|_\infty \leq \varepsilon$，则复合函数

$$\widehat{g}_\varepsilon := g_\varepsilon \circ \varphi_\star$$

满足

$$\sup_{x \in \mathcal{M}} |f(x) - \widehat{g}_\varepsilon(x)| \leq \varepsilon.$$

证明。对于 $x \in \mathcal{M}$，$\varphi_\star(x) \in K'$ 且 $H(\varphi_\star(x)) = h(\varphi_\star(x)) = f(x)$。因此

$$|f(x) - \widehat{g}_\varepsilon(x)| = |H(\varphi_\star(x)) - g_\varepsilon(\varphi_\star(x))| \leq \sup_{u \in [0,1]^d} |H(u) - g_\varepsilon(u)| \leq \varepsilon.$$

对 $x \in \mathcal{M}$取上确界即可得到结论。 □

证明的完成。具备这些工具后，定理5.1的证明可通过串联以下引理得到：嵌入与归一化（引理 H.10）、范数控制的拉回（引理 H.11）、范数控制的立方体延拓（引理 H.12–H.13）、立方体上的欧几里得近似（引理 H.14），以及近似函数的拉回（引理 H.15）。

定理*5.1*的证明。令 $\psi \in C^{kp,\,1}(\mathcal{M})$ 并固定 $\varepsilon \in (0,\,1)$。定义

$$M := \|\psi\|_{C^{kp,1}(\mathcal{M})}.$$

若 $M = 0$，则 $\|\psi\|_\infty \leq \|\psi\|_{C^{kp,\,1}(\mathcal{M})} = 0$，因此 $\psi \equiv 0$，结论显然成立。接下来假设 $M > 0$.

若 $\varepsilon \geq M$，则 $\|\psi\|_\infty \leq M \leq \varepsilon$，因此常数网络 $\hat{}g_\varepsilon \equiv 0$ 对任意特征映射 $\varphi_\star$ 都会得到 $\|\psi - \hat{}g_\varepsilon \circ \varphi_\star\|_\infty \leq \varepsilon$。因此我们可以假设

$$0 < \varepsilon < M, \qquad \text{so that} \qquad 0 < \frac{\varepsilon}{M} < 1.$$

定义归一化函数

$$\tilde{\psi} := \frac{1}{M}\psi, \qquad \text{so that} \qquad \|\tilde{\psi}\|_{C^{kp,1}(\mathcal{M})} = 1.$$

根据引理 H.10，选取一个光滑嵌入 $\varphi^\star : \mathcal{M} \to K' \subset [0,1]^{2p}$。定义

$$\widetilde{h} := \tilde{\psi} \circ \varphi_\star^{-1} \quad \text{on } K'.$$

根据引理 H.11，我们有 $\widetilde{h} \in C^{kp,\,1}(K')$，以及

$$\|\widetilde{h}\|_{C^{kp,1}(K')} \le C_{\mathrm{pb}} \|\tilde{\psi}\|_{C^{kp,1}(\mathcal{M})} = C_{\mathrm{pb}}.$$

根据引理 H.13，存在 $\widetilde{H} \in C^{kp,1}([0,1]^{2p})$，使得 $\widetilde{H} = \widetilde{h}$ 在 K' 上且

$$\|\widetilde{H}\|_{C^{kp,1}([0,1]^{2p})} \le C_{\mathrm{ext}} \|\widetilde{h}\|_{C^{kp,1}(K')} \le C_{\mathrm{ext}} C_{\mathrm{pb}}.$$

Set

$$\widetilde{B} := \|\widetilde{H}\|_{C^{kp,1}([0,1]^{2p})}.$$

由于 $M > 0$、$\tilde{\psi} \not\equiv 0$，故 $\widetilde{h} \not\equiv 0$ 在 K'上。因为 $\widetilde{H} = \widetilde{h}$ 在 K'上，$\widetilde{H} \not\equiv 0$ 在 $[0,1]^{2p}$上，故 $\|\widetilde{H}\|_\infty > 0$。由于$\sim$ $\sim\|H\|_{C^{kp,1}([0,1]^{2p})} \ge \|H\|_\infty$，由此可得

$$\widetilde{B} > 0.$$

再次归一化：

$$\overline{H} := \frac{1}{\widetilde{B}} \widetilde{H}, \quad \text{so that} \quad \|\overline{H}\|_{C^{kp,1}([0,1]^{2p})} = 1,$$

定义精度参数

$$\delta := \frac{\varepsilon / M}{\widetilde{B}}.$$

若 $\delta \ge 1$，则零网络已在误差 1 内逼近 H，因而也在误差 δ内，故逼近步骤是平凡的。因此我们可以假设 $\delta \in (0,1)$。

现在，我们采用以下选择应用定理 H.9：

$$n := 2p, \qquad r := kp + 1, \qquad \beta := \frac{3r}{2n} = \frac{3(kp+1)}{4p}.$$

这一选择满足

$$\frac{r}{n} < \frac{kp+1}{2p} < \frac{3(kp+1)}{4p} = \beta \le \frac{kp+1}{p} = \frac{2r}{n},$$

因此，定理 H.9 适用。于是，存在 W_ε，L_ε 以及一个网络

$$\bar{g}_\varepsilon \in \mathcal{NN}^{\sigma}_{2p, W_\varepsilon, L_\varepsilon}$$

使得

$$\|\overline{H} - \bar{g}_\varepsilon\|_\infty \le \delta,$$

and

$$W_\varepsilon \le \left(\frac{\tilde{C}_{r,2p,\beta,\sigma}}{\delta} \right)^{1/\beta}, \qquad L_\varepsilon \le C_{r,2p,\sigma} W^{\beta(2p)/r - 1}.$$

通过如下定义进行缩放还原

$$\tilde{g}_\varepsilon := \widetilde{B} \bar{g}_\varepsilon.$$

Then

$$\|\widetilde{H} - \tilde{g}_\varepsilon\|_\infty = \widetilde{B} \|\overline{H} - \bar{g}_\varepsilon\|_\infty \le \widetilde{B}\delta = \frac{\varepsilon}{M}.$$

最后定义

$$g_\varepsilon := M \tilde{g}_\varepsilon.$$

$\tilde{}$将输出缩放 M（此前还通过 B进行缩放）是通过缩放最终仿射层来实现的，因此它不会改变宽度 W_ε 或深度 L_ε。此外，

$$\|M\widetilde{H} - g_\varepsilon\|_\infty = M \|\widetilde{H} - \tilde{g}_\varepsilon\|_\infty \le M \cdot \frac{\varepsilon}{M} = \varepsilon.$$

~ ~设置 $H := MH$。在 K' 上，我们有 $H = Mh = h$，其中

$$h := \psi \circ \varphi_\star^{-1},$$

~因为 $h = (1/M)h$。

在 $\mathcal{M}$ 上定义拉回近似函数为

$$\hat{\psi}_\varepsilon := \hat{g}_\varepsilon \circ \varphi_\star.$$

对于任意 $x \in \mathcal{M}$，令 $u := \varphi_\star(x) \in K'$。则 $\psi(x) = h(u) = H(u)$，且 $\hat{\psi}_\varepsilon(x) = \hat{g}_\varepsilon(u)$，因此

$$|\psi(x) - \hat{\psi}_\varepsilon(x)| = |H(u) - g_\varepsilon(u)| \le \|H - g_\varepsilon\|_\infty \le \varepsilon.$$

对 $x \in \mathcal{M}$ 取上确界，得到

$$\|\psi - \hat{\psi}_\varepsilon\|_\infty \le \varepsilon.$$

~还需推导出所述速率。由于 $\delta = (\varepsilon/M)/B$，

$$W_\varepsilon \le \left(\frac{\widetilde{C}_{r,2p,\beta,\sigma}}{\delta}\right)^{1/\beta} = \left(\frac{\widetilde{C}_{r,2p,\beta,\sigma}\widetilde{B}M}{\delta}\right)^{1/\beta}.$$

~利用 $B \le C_{\mathrm{ext}}C_{\mathrm{pb}}$，我们得到

$$W_\varepsilon \le C\left(\frac{M}{\varepsilon}\right)^{1/\beta}, \qquad C := \left(\widetilde{C}_{r,2p,\beta,\sigma}C_{\mathrm{ext}}C_{\mathrm{pb}}\right)^{1/\beta}.$$

由 $\beta = {}^{3}\!\!\!

步骤 1：预势的一致逼近 $\psi_\star$。 将定理 5.1应用于函数 $\psi_\star \in C^{kp,\ 1}(\mathcal{M})$。于是存在一个特征映射 $\varphi_\star$： $\mathcal{M} \to \mathbb{R}^{2p}$，满足假设2.2，以及整数 W_ε， $L_\varepsilon \in \mathbb{N}$和一个网络$\hat{}g_\varepsilon \in \mathcal{NN}^\sigma_{2p,\ W_\varepsilon,\ L_\varepsilon}$，使得拉回网络

$$\hat{\psi}_\varepsilon := \hat{g}_\varepsilon \circ \varphi_\star$$

满足

$$\|\psi_\star - \hat{\psi}_\varepsilon\|_\infty < \varepsilon, \tag{26}$$

且 W_ε， L_ε 满足定理 5.1中给出的界。

步骤 2：通过 c-变换传递误差。定义 c-变换 $\hat{}\phi_\varepsilon := \hat{}\psi^c_\varepsilon$。我们声称

$$\|\phi - \hat{\phi}_\varepsilon\|_\infty \le \|\psi_\star - \hat{\psi}_\varepsilon\|_\infty. \tag{27}$$

为证明这一点，令 $\delta := \|\psi_\star - \hat{}\psi_\varepsilon\|_\infty$。那么对于每个 $y \in \mathcal{M}$,

$$\psi_\star(y) - \delta \le \hat{\psi}_\varepsilon(y) \le \psi_\star(y) + \delta.$$

固定 $x \in \mathcal{M}$。因此，对于每个 $y \in \mathcal{M}$，我们有

$$c(x,y) - \psi_\star(y) - \delta \le c(x,y) - \hat{\psi}_\varepsilon(y) \le c(x,y) - \psi_\star(y) + \delta.$$

对 $y \in \mathcal{M}$ 取下确界，得到

$$\inf_y \big(c(x,y) - \psi_\star(y)\big) - \delta \le \inf_y \big(c(x,y) - \hat{\psi}_\varepsilon(y)\big) \le \inf_y \big(c(x,y) - \psi_\star(y)\big) + \delta.$$

By d根据 c-变换的定义（约定 $\eta^c(x) = \inf_y(c(x,\ y) - \eta(y))$），这恰好是

$$\phi(x) - \delta \le \hat{\phi}_\varepsilon(x) \le \phi(x) + \delta.$$

$\hat{}$因此 $|\phi(x) - \phi_\varepsilon(x)| \le \delta$ 对所有 $x \in \mathcal{M}$成立，对 x 取上确界即可证明（27）。

将（27）与（26）结合，得到

$$\|\phi - \hat{\phi}_\varepsilon\|_\infty \le \|\psi_\star - \hat{\psi}_\varepsilon\|_\infty < \varepsilon.$$

回顾 $\phi = \phi_\star$，我们得到

$$\|\phi_\star - \hat{\phi}_\varepsilon\|_\infty < \varepsilon.$$

步骤 3：复杂度界。唯一的神经网络逼近步骤是步骤 1，其中 $\hat{}\psi_\varepsilon = \hat{}g_\varepsilon \circ \varphi_\star$ 是利用定理 5.1构造的。步骤 2 应用确定性算子 $\eta \mapsto \eta^c$，且不会改变 $\hat{}g_\varepsilon$的架构。因此 $\hat{}g_\varepsilon$满足与定理 5.1中完全相同的宽度/深度界。证明完毕。□

H.6. 定理5.3 的证明

在本附录中，为简化记号并提升可读性，我们用 ϕ 代替 $\phi_\star$，并用 ψ 代替 $\psi_\star$。在证明之前，我们先回顾一些来自Cordero-Erausquin 等人(2001) 的技术结果。关于 c-凹函数在 $\mathcal{M}$上的刻画如下。

引理 H.16. *(Cordero-Erausquin 等人，2001，引理 3.3)* 设 ϕ 是定义在 c上的 -凹函数，并定义 $\mathcal{M}$ $T(x) = \exp_x(-\nabla\phi(x))$。

- 函数 ϕ 是利普希茨连续的，因此在 $\mathcal{M}$上几乎处处可微；
- 任取一个满足 ϕ 可微的 $x \in \mathcal{M}$。那么， $y = T(x)$ 当且仅当 y 是以下最小化问题的解

$$c(x,y) - \phi(x) - \phi^c(y) \tag{28}$$

over $y' \in \mathcal{M}$。后一种情况下，有 $\nabla\phi(x) = \nabla d^2_y(x)$。

固定 $U \subseteq \mathcal{M}$ 开集。若函数 $f\colon U \to \mathbb{R}$ 在 U 上满足半凹，则对任意 $x_0 \in U$，都存在一个凸嵌入球 $B_r(x_0)$和一个光滑函数 $V\colon B_r(x_0) \to \mathbb{R}$，使得 $f+V$在整个 $B_r(x_0)$上测地凹。半凹函数的海森矩阵定义如下。

定义 H.17。 (Cordero-Erausquin 等人，2001，定义 3.9) 设 $f\colon U \to \mathbb{R}$ 是一个半凹函数。那么 f 在 $x \in U$ 处有海森矩阵，条件是 f 在 x 处可微，且存在一个自伴算子 海森矩阵$_x f\colon T_x\mathcal{M} \to T_x\mathcal{M}$ 使得

$$\sup_{v \in \partial f(\exp_x(u))} \|\Pi_{x,u} v - \nabla f(x) - \mathrm{Hess} f u\|_x = o(\|u\|) \tag{29}$$

作为 $u \to 0$ 在 T 中 $_x\mathcal{M}$。这里$\Pi_{x,\ u}\colon T_{\exp_x(u)} \to T_x\mathcal{M}$ 表示沿着 $\gamma(t) = \exp_x(tu$ 到 x 的平行移动).

若 $f\colon U \to \mathbb{R}$ 是半凹函数，则亚历山德罗夫–邦盖特定理指出，海森矩阵$_x f$ 相对于体积测度，在 U 上几乎处处存在（Bangert，1979）。事实证明，对于紧致的 $\mathcal{M}$，其上的任意 c-凹函数 $\mathcal{M}$ 都是半凹函数，因此在 $\mathcal{M}$(Cordero-Erausquin 等人，2001，命题 3.14) 上几乎处处有海森矩阵。

令 ϕ 为二次代价函数 $c(x,\ y) = \frac{1}{2}d(x,\ y)$ 的一个 c-凹 Kantorovich势函数[2]，并令 $\psi := \phi^c$。在 ϕ（分别为 ψ）可微的点处，定义

$$\overrightarrow{T}(x) := \exp_x(-\nabla\phi(x)), \qquad \overleftarrow{T}(y) := \exp_y(-\nabla\psi(y)).$$

并引入集合

$$E_\phi := \{x \in \mathcal{M}\colon \mathrm{Hess}_x\phi \text{ exists in the sense of (29)}\},$$
$$E_\psi := \{y \in \mathcal{M}\colon \mathrm{Hess}_y\psi \text{ exists in the sense of (29)}\}.$$

对于 $x \in E_\phi$，令 $y = \overrightarrow{T}(x)$；对于 $y \in E_\psi$，令 $x = \overleftarrow{T}(y)$。定义严格正集

$$\begin{aligned}\widetilde{E}_\phi &:= \Big\{x \in E_\phi\colon \mathrm{Hess}_x\big(\tfrac{1}{2}d(\cdot, \overrightarrow{T}(x))^2 - \phi\big) > 0\Big\},\\ \widetilde{E}_\psi &:= \Big\{y \in E_\psi\colon \mathrm{Hess}_y\big(\tfrac{1}{2}d(\overleftarrow{T}(y), \cdot)^2 - \psi\big) > 0\Big\},\end{aligned} \tag{30}$$

以及良好点集合

$$\Omega := \{x \in \widetilde{E}_\phi\colon \overrightarrow{T}(x) \in \widetilde{E}_\psi\}.$$

根据（Cordero-Erausquin 等人，2001，命题 4.4），$\mu(\Omega) = 1$。

现在我们开始证明定理5.3。

定理 *5.3*的证明。 固定 $x \in \Omega$并令

$$y^\star := \overrightarrow{T}(x).$$

令 $\varepsilon \in (0,1)$，并令 $\psi_\varepsilon \in C(\mathcal{M})$ 满足

$$\|\psi_\varepsilon - \psi\|_\infty \le \varepsilon.$$

定义未扰动目标函数和扰动目标函数

$$F_x(y) := \tfrac{1}{2}d(x,y)^2 - \psi(y), \qquad F_{\varepsilon,x}(y) := \tfrac{1}{2}d(x,y)^2 - \psi_\varepsilon(y), \qquad y \in \mathcal{M},$$

并任取一个极小化元

$$y_\varepsilon^\star \in \operatorname*{arg\,min}_{y \in \mathcal{M}} F_{\varepsilon,x}(y).$$

根据定义

$$y_\varepsilon^\star = T_\varepsilon(x)$$

对于某个可容许选择 T_ε。

~步骤 1：$y^\star$ 是 F_x 的唯一极小值点。由于 $x \in \Omega \subset E_\phi \subset E_\phi$，势函数 ϕ 在 x 处存在海森矩阵，因此特别地，它在 x处可微。根据引理 H.16，在 ϕ的可微点处，$y^\star = \overrightarrow{T}(x)$ 当且仅当 $y^\star$ 是 (28) 的极小化元。由于 $\phi(x)$ 关于 y为常数，极小化 H.16 等价于极小化 $y \mapsto c(x,y) - \psi(y) = F_x(y)$。此外，引理 H.16 表明，任何极小化元都必须与 $\overrightarrow{T}(x)$ 重合，因此 $y^\star$ 是唯一极小化元。

步骤 2：Hess$F_x(y^\star)$ 的严格正性及 $C(x)$ 的定义。由于 $x \in \Omega$，我们有 $y^\star = \overrightarrow{T}(x) \in \widetilde{E}_\psi$。特别地，$y^\star \in E_\psi$，因此 ψ 在 $y^\star$ 处可微，且 $\overrightarrow{T}(y^\star)$ 定义良好。由于 $y^\star \in \partial^c\phi(x)$（由步骤 1），(Cordero-Erausquin 等人，2001，第 3.1 节) 中的 c-次微分对称性 $y \in \partial^c\phi(x) \iff x \in \partial^c\psi(y)$ 推出 $x \in \partial^c\psi(y^\star)$；由于 ψ 在 $y^\star$处可微，引理 H.16 表明 $\partial^c\psi(y^\star) = \{\overleftarrow{T}(y^\star)\}$，因此

$$\overleftarrow{T}(y^\star) = x.$$

根据 E_ψ 在（30）中的定义，这意味着

$$y \longmapsto \tfrac{1}{2}d\big(\overleftarrow{T}(y^\star), y\big)^2 - \psi(y) = \tfrac{1}{2}d(x,y)^2 - \psi(y) = F_x(y)$$

存在于 $y^\star$，且为正定。因此存在 $C(x) > 0$，使得

$$\langle \mathrm{Hess}F_x(y^\star)u, u\rangle \geq 2C(x)\|u\|^2, \qquad \forall u \in T_{y^\star}\mathcal{M}.$$

步骤 3：F_x 在 $y^\star$ 附近的局部二次增长。由于 Ω 和 $\psi(y)$，函数 $F_x(y) = \frac{1}{2}d(x,y)^2 - \psi(y)$ 在 $y^\star$ 处存在（Aleksandrov–Bangert）海森矩阵，其意义如 (29) 所述。特别地，该函数在 $y^\star$ 处可微，且 F_x 蕴含如下二阶展开：当 $u \to 0$属于 $T_{y^\star}\mathcal{M}$时，

$$F_x(\exp_{y^\star}(u)) = F_x(y^\star) + \langle \nabla F_x(y^\star), u\rangle + \frac{1}{2}\langle \mathrm{Hess}F_x(y^\star)u, u\rangle + o(\|u\|^2). \tag{31}$$

由于 $y^\star$ 是 F_x 的极小化元（步骤 1），且 F_x 在 $y^\star$处可微，因此有 $\nabla F_x(y^\star) = 0$，故（31）变为

$$F_x(\exp_{y^\star}(u)) - F_x(y^\star) = \frac{1}{2}\langle \mathrm{Hess}F_x(y^\star)u, u\rangle + o(\|u\|^2). \tag{32}$$

此外，海森矩阵的严格正定性（步骤 2）给出 $C(x) > 0$，使得

$$\frac{1}{2}\langle \mathrm{Hess}F_x(y^\star)u, u\rangle \geq C(x)\|u\|^2 \qquad \forall u \in T_{y^\star}\mathcal{M}. \tag{33}$$

结合 (32)–(33) 可得

$$F_x(\exp_{y^\star}(u)) - F_x(y^\star) \geq C(x)\|u\|^2 + o(\|u\|^2). \tag{34}$$

根据朗道符号 $o(\|u\|^2)$ 的定义，我们有 $o(\|u\|^2)/\|u\|^2 \to 0$，当 $u \to 0$ 时。因此，取 $\eta := C(x)/2 > 0$，存在 $\rho_x > 0$，使得只要 $0 < \|u\| < \rho_x$，

$$\left|\frac{o(\|u\|^2)}{\|u\|^2}\right| \leq \frac{C(x)}{2} \qquad \Longrightarrow \qquad o(\|u\|^2) \geq -\frac{C(x)}{2}\|u\|^2. \tag{35}$$

将 (35) 代入 (34)，可得对所有 $\|u\| < \rho_x$，

$$F_x(\exp_{y^\star}(u)) - F_x(y^\star) \geq C(x)\|u\|^2 - \frac{C(x)}{2}\|u\|^2 = \frac{C(x)}{2}\|u\|^2.$$

选择 $r_x > 0$使得

$$0 < r_x < \min\{\rho_x, \mathrm{inj}(y^\star)\},$$

其中 inj$(y^\star)$ 是 $y^\star$处的内射半径。于是，对每个 $y \in B_{r_x}(y^\star)$，都存在唯一的 $u \in T_{y^\star}\mathcal{M}$，满足 $\|u\| < r_x$，使得 $y = \exp_{y^\star}(u)$。此外，由于 $\|u\| < \mathrm{inj}(y^\star)$，测地线 $\gamma(t) := \exp_{y^\star}(tu)$，$t \in [0,1]$，是从 $y^\star$ 到 y 的唯一最短测地线；其速度为常数 $\|\dot\gamma(t)\| = \|u\|$，故长度为 $L(\gamma) = \int_0^1 \|\dot\gamma(t)\|\,dt = \|u\|$。因此

$$d(y, y^\star) = L(\gamma) = \|u\|. \tag{36}$$

利用 $y = \exp_{y^\star}(u)$和（36），我们得到局部二次增长：对于所有 $y \in B_{r_x}(y^\star)$，

$$F_x(y) \geq F_x(y^\star) + \frac{C(x)}{2}\,d(y, y^\star)^2. \tag{37}$$

Step 4：一致扰动界。 由 $\|\psi_\varepsilon - \psi\|_\infty \le \varepsilon$ 我们立即得到，对于所有 $y \in \mathcal{M}$ ，

$$F_x(y) - \varepsilon \;\le\; F_{\varepsilon,x}(y) \;\le\; F_x(y) + \varepsilon. \tag{38}$$

步骤 5：扰动后的极小值点处的目标函数间隙。 由 $y_\varepsilon^\star$ 对 $F_{\varepsilon\text{的极小性可知，}x}$，我们有 $F_{\varepsilon,\,x}(y_\varepsilon^\star) \le F_{\varepsilon,\,x}(y^\star)$ 。在38处应用（ $y_\varepsilon^\star$ 并结合 $y^\star$ 可得

$$F_x(y_\varepsilon^\star) - F_x(y^\star) \le 2\varepsilon. \tag{39}$$

步骤 6：在 $y_\varepsilon^\star$ 附近定位 $y^\star$。 由步骤 1， $y^\star$ 是紧流形 $\mathcal{M}$ 上连续函数 F_x 的唯一极小值点。因此，对于步骤 3 中固定的半径 r_x， $F_x - F_x(y^\star)$ 在紧集 $\mathcal{M}\setminus B_{r_x}(y^\star)$ 上取得最小值，且该最小值严格为正：

$$\delta_x := \min_{y\in\mathcal{M}\setminus B_{r_x}(y^\star)} \big(F_x(y) - F_x(y^\star)\big) \;>\; 0.$$

若 $y_\varepsilon^\star \notin B_{r_x}(y^\star)$ ，则 $F_x(y_\varepsilon^\star) - F_x(y^\star) \ge \delta_x$，这与（39）矛盾，只要 $2\varepsilon < \delta_x$。因此，令 $\varepsilon_x := \delta_x/4$，我们得出结论：若 $\varepsilon \le \varepsilon_x$，则

$$y_\varepsilon^\star \in B_{r_x}(y^\star).$$

步骤 7：得出距离界。 假设 $\varepsilon \in (0,\min\{1,\varepsilon_x\})$，则 $y_\varepsilon^\star \in B_{r_x}(y^\star)$。在 $y = y_\varepsilon^\star$ 处应用局部二次增长（39），并结合（37）可得

$$\frac{C(x)}{2}\, d(y_\varepsilon^\star, y^\star)^2 \le F_x(y_\varepsilon^\star) - F_x(y^\star) \le 2\varepsilon,$$

因此

$$d(y_\varepsilon^\star, y^\star) \le 2\sqrt{\frac{\varepsilon}{C(x)}}.$$

注意到 $y^\star = \overrightarrow{T}(x)$ 和 $y_\varepsilon^\star = T_\varepsilon(x)$，将 $\overrightarrow{T}$ 与 $T_\star$ 对应起来后即可证明我们的结论。 □

H.7. 引理H.3的证明

证明。 令 $D := \operatorname{diam}(\mathcal{M}) < \infty$。则对于所有 $x \in \mathcal{M}$，均有 $d(x, x_0) \le D$，因此

$$\int_{\mathcal{M}} d(x,x_0)^{r+\delta}\,\mathrm{d}\mu(x) \le \int_{\mathcal{M}} D^{r+\delta}\,\mathrm{d}\mu(x) = D^{r+\delta} < \infty,$$

这里我们用到了 μ 是一个概率测度。

还需控制（22）中的第二项。回顾一下，对于 $\rho \ge 0$,

$$A_{x_0}(\rho) := \sup_{v\in\mathbb{S}_\rho^{p-1}} \sup_{\substack{w\in T_v\mathbb{S}_\rho^{p-1}\\ |w|_{x_0}=\rho}} \big\| d_v \exp_{x_0}[w] \big\|_{\exp_{x_0}(v)}.$$

步骤 1：被积函数 $(v,\ w) \mapsto \|d_v \exp_{x_0}[w]\|_{\exp_{x_0}(v)}$ 的连续性。 由于 $\mathcal{M}$ 是紧致的，因此作为度量空间它是完备的。根据 Hopf–Rinow 定理， $\mathcal{M}$ 于是测地完备；特别地，对于每个 $v \in T_{x_0}\mathcal{M}$ ，具有 $\gamma_v(0) = x_0$ 和 $\dot\gamma_v(0) = v$ 的测地线 γ_v 至少在时间 $t = 1$之前有定义，并且

$$\exp_{x_0}(v) = \gamma_v(1)$$

对所有 $v \in T_{x_0}\mathcal{M}$ (Lee，2018年，推论 6.22）均有定义。此外，$\exp^{x_0}$ 是一个光滑映射（Lee，2018年，命题 5.19）。将$T_{x_0}\mathcal{M} \simeq \mathbb{R}^p$ 视为光滑流形，光滑性意味着微分

$$d_v \exp_{x_0} :\; T_v(T_{x_0}\mathcal{M}) \simeq T_{x_0}\mathcal{M} \;\longrightarrow\; T_{\exp_{x_0}(v)}\mathcal{M}$$

关于 v光滑依赖，因此相应的求值映射

$$T_{x_0}\mathcal{M} \times T_{x_0}\mathcal{M} \to T\mathcal{M}, \qquad (v,w) \longmapsto d_v \exp_{x_0}[w],$$

是光滑的（因而是连续的）（Lee，2012，例 3.4）。

接下来，由于黎曼度量是内积的光滑场（Lee，2018年，第 2 节），切丛上的诱导范数连续依赖于基点：在局部坐标 (U,φ)中，记 $\xi=\sum_i \xi^i\partial_i|_x \in T_p\mathcal{M}$，则有

$$\|\xi\|_x^2 = g_{ij}(x)\,\xi^i\xi^j,$$

系数 $x\mapsto g_{ij}(x)$ 是光滑的，因此也是连续的。由此可知映射

$$(x,\xi)\ \longmapsto\ \|\xi\|_x \qquad (x\in\mathcal{M},\ \xi\in T_x\mathcal{M})$$

is 在 $T\mathcal{M}$ 上连续。结合 $(v,w)\mapsto d(\exp_{x_0})_v[w]$ 的连续性可知 t

$$F: T_{x_0}\mathcal{M}\times T_{x_0}\mathcal{M}\to\mathbb{R}, \qquad F(v,w):=\big\|d_v\exp_{x_0}[w]\big\|_{\exp_{x_0}(v)},$$

是连续的。

步骤 2：给出 $A_{x_0}(\rho)$ 在 $\rho\in[0,D]$ 上的一致界。 在欧几里得空间 $(T_{x_0}\mathcal{M},g_{x_0})$ 中考虑。球面 $\mathbb{S}_\rho^{p-1}=\{u:\|u\|_{x_0}=\rho\}$ 是光滑函数 $u\mapsto\|u\|_{x_0}^2$ 的水平集，因此当 $\rho>0$ 时，它是一个光滑超曲面。其在 $v\in\mathbb{S}_\rho^{p-1}$ 处的切空间为

$$T_v\mathbb{S}_\rho^{p-1}=\{w\in T_{x_0}\mathcal{M}:\ \langle v,w\rangle_{x_0}=0\},$$

因为对 $t\mapsto\|v+tw\|_{x_0}^2$ 在 $t=0$ 处求导得到 $\frac{d}{dt}|_{t=0}\|v+tw\|_{x_0}^2=2\langle v,\ w\rangle_{x_0}$（Lee，2012，第 5 章）。

固定 $\rho\in[0,D]$，并考虑 $A_{x_0}(\rho)$ 定义中的任意可容许对 $(v,\ w)$。根据定义，$v\in\mathbb{S}_\rho^{p-1}$且 $\|w\|_{x_0}=\rho$，因此

$$\|v\|_{x_0}=\rho\le D, \qquad \|w\|_{x_0}=\rho\le D.$$

此外 $w\in T_v\mathbb{S}_\rho^{p-1}$，这等价于 $\langle v,\ w\rangle_{x_0}=0$。因此，每个可容许的 $(v,\ w)$ 都属于 t o

$$E:=\Big\{(v,w)\in T_{x_0}\mathcal{M}\times T_{x_0}\mathcal{M}:\ \|v\|_{x_0}\le D,\ \|w\|_{x_0}\le D,\ \langle v,w\rangle_{x_0}=0\Big\}.$$

我们声称 E 是紧致的。事实上，集合 $\{(v,\ w):\ \|v\|_{x_0}\le D\}$ 和 $\{(v,\ w):\ \|w\|_{x_0}\le D\}$ 是有限维空间 $T_{x_0}\mathcal{M}\times T_{x_0}\mathcal{M}\simeq\mathbb{R}^{2p}$ 中的闭球，因此是紧致的。约束 $\langle v,\ w\rangle_{x_0}=0$ 定义了一个闭子集，因为 $(v,\ w)\mapsto\langle v,\ w\rangle_{x_0}$ 是连续的，且 $\{0\}$ 是闭的。因此，E 是紧集的闭子集，故为紧致集。

由于 F 是连续的，且 E 是紧致的，根据极值定理，F 在 E 上取得最大值。定义

$$M:=\max_{(v,w)\in E}F(v,w)<\infty.$$

由于 $A_{x_0}(\rho)$的可容许集包含于 E中，对于每个 $\rho\in[0,D]$，我们得到

$$A_{x_0}(\rho)=\sup_{\text{admissible }(v,w)}F(v,w)\le\sup_{(v,w)\in E}F(v,w)=M.$$

步骤 3：的可积性$(d(x,\ x_0))^r$。 对于任意 $x\in\mathcal{M}$，我们有 $d(x,\ x_0)\le D$，因此 $d(x,\ x_0)\in[0,D]$，从而

$$A_{x_0}(d(x,x_0))\le M \qquad\Longrightarrow\qquad A_{x_0}(d(x,x_0))^r\le M^r.$$

对概率测度 μ积分可得

$$\int_{\mathcal{M}}A_{x_0}(d(x,x_0))^r\,\mathrm{d}\mu(x)\le\int_{\mathcal{M}}M^r\,\mathrm{d}\mu(x)=M^r<\infty.$$

将此结果与第一项的界结合起来，即可证明（22）。 □

H.8. 引理的证明H.4

证明。 令 $\alpha := \mathrm{supp}(\eta)$。根据假设， α 有限且 $|\alpha| \le m$。根据 2-瓦瑟斯坦距离的定义，

$$W_2(\rho,\eta)^2 = \inf_{\gamma\in\Pi(\rho,\eta)} \int_{\mathcal{M}\times\mathcal{M}} d(y,z)^2 \,\mathrm{d}\gamma(y,z),$$

其中 $\Pi(\rho,\ \eta)$ 表示 ρ 和 η 的耦合集合。

步骤 1：对任意耦合进行析取。 任取 $\gamma \in \Pi(\rho,\ \eta)$。由于 $\mathcal{M}$ 是完备可分度量空间，我们可以对Figalli (2010, 定理 1.4）中的析取定理应用于 γ ，相对于投影$\mathrm{pr}_1 : \mathcal{M}\times\mathcal{M} \to \mathcal{M}$。由于$(\mathrm{pr}_1)_{\#}\gamma = \rho$，存在一个 ρ-可测的概率测度族 $\{\gamma_y\}_{y\in\mathcal{M}} \subset \mathcal{P}(\mathcal{M})$，使得

$$\gamma(\mathrm{d}y,\mathrm{d}z) = \rho(\mathrm{d}y)\,\gamma_y(\mathrm{d}z), \tag{40}$$

即对每个有界博雷尔函数 $\varphi :\ \mathcal{M}\times\mathcal{M} \to \mathbb{R}$,

$$\int_{\mathcal{M}\times\mathcal{M}} \varphi(y,z)\,\mathrm{d}\gamma(y,z) = \int_{\mathcal{M}} \left(\int_{\mathcal{M}} \varphi(y,z)\,\mathrm{d}\gamma_y(z) \right) \mathrm{d}\rho(y). \tag{41}$$

由于 $\mathcal{M}$ 是紧致的，函数 $(y,\ z) \mapsto d(y,\ z)^2$ 在 $\mathcal{M}\times\mathcal{M}$上有界。因此，我们可以在（41）中取 $\varphi(y,\ z) = d(y,\ z)^2$ 以得到

$$\int_{\mathcal{M}\times\mathcal{M}} d(y,z)^2\,\mathrm{d}\gamma(y,z) = \int_{\mathcal{M}} \left(\int_{\mathcal{M}} d(y,z)^2\,\mathrm{d}\gamma_y(z) \right) \mathrm{d}\rho(y). \tag{42}$$

步骤 2：利用 γ_y 支撑于 α。 由于 $\gamma \in \Pi(\rho,\eta)$ 的第二边缘分布为 η ，且 η 支撑于 α，因此必有：由于 $\gamma \in \Pi(\rho,\eta)$ 的第二边缘分布为 η，即

$$(\mathrm{pr}_2)_{\#}\gamma = \eta,$$

其中，pr_2：$\mathcal{M}\times\mathcal{M} \to \mathcal{M}$ 表示第二投影 $\mathrm{pr}_2(y,z) = z$。因此，根据推前测度的定义，对于任意博雷尔集 $B \subset \mathcal{M}$,

$$\eta(B) = ((\mathrm{pr}_2)_{\#}\gamma)(B) = \gamma(\mathrm{pr}_2^{-1}(B)).$$

取 $B = \mathcal{M}\setminus\alpha$ ，并利用 η 支撑于 α （因此 $\eta(\mathcal{M}\setminus\alpha) = 0$） ，可得

$$\gamma\big(\mathrm{pr}_2^{-1}(\mathcal{M}\setminus\alpha)\big) = 0.$$

最后，$\mathrm{pr}_2^{-1}(\mathcal{M}\setminus\alpha) = \mathcal{M}\times(\mathcal{M}\setminus\alpha)$，因此

$$\gamma(\mathcal{M}\times(\mathcal{M}\setminus\alpha)) = \eta(\mathcal{M}\setminus\alpha) = 0.$$

利用（40）对此恒等映射作分解，得到

$$0 = \gamma(\mathcal{M}\times(\mathcal{M}\setminus\alpha)) = \int_{\mathcal{M}} \gamma_y(\mathcal{M}\setminus\alpha)\,\mathrm{d}\rho(y),$$

因此，对 ρ-几乎处处的 y，有 $\gamma_y(\mathcal{M}\setminus\alpha) = 0$。等价地,

$$\mathrm{supp}(\gamma_y) \subseteq \alpha \qquad \text{for } \rho\text{-a.e. } y.$$

步骤 3：用最近站点距离给条件代价函数设下界。固定 y 使得 $\mathrm{supp}(\gamma_y) \subseteq \alpha$。那么，由于 γ_y 是一个支撑集位于 α上的概率测度,

$$\int_{\mathcal{M}} d(y,z)^2\,\mathrm{d}\gamma_y(z) = \int_{\alpha} d(y,z)^2\,\mathrm{d}\gamma_y(z) \ge \min_{a\in\alpha} d(y,a)^2 \int_{\alpha} \mathrm{d}\gamma_y(z) = \min_{a\in\alpha} d(y,a)^2.$$

将其代入（42），得到

$$\int_{\mathcal{M}\times\mathcal{M}} d(y,z)^2\,\mathrm{d}\gamma(y,z) \ge \int_{\mathcal{M}} \min_{a\in\alpha} d(y,a)^2\,\mathrm{d}\rho(y). \tag{43}$$

步骤 4：对耦合取下确界，并与 $V_{m,\,2}(\rho)$。 由于（43）对每个 $\gamma \in \Pi(\rho,\ \eta)$ 都成立，对 γ 取下确界可得

$$W_2(\rho,\eta)^2 = \inf_{\gamma\in\Pi(\rho,\eta)} \int d(y,z)^2 \,\mathrm{d}\gamma \geq \int_{\mathcal{M}} \min_{a\in\alpha} d(y,a)^2 \,\mathrm{d}\rho(y).$$

最后，根据定义H.1，$V_{m,2}$ 是所有规模至多为 m的集合上的下确界，并且由于 $|\alpha| \leq m$，我们有

$$\int_{\mathcal{M}} \min_{a\in\alpha} d(y,a)^2 \,\mathrm{d}\rho(y) \ \geq \ V_{m,2}(\rho).$$

结合最后两个式子即可证明该结论。 □

H.9. 引理的证明H.5

证明。 若 $\mathcal{M}$ 是紧致的，则每个概率测度都具有有限的二阶矩，因此 $\nu_t, \nu_s \in \mathcal{P}_2(\mathcal{M})$与 $W_2(\nu_t, \nu_s)$是良定义的。

定义可测映射(t, s)：$\mathcal{M} \to \mathcal{M} \times \mathcal{M}$，由 $x \mapsto (t(x)\ \ s(x))$给出，并令

$$\pi := (t, s)_{\#}\mu.$$

设 $\mathrm{pr}_1, \mathrm{pr}_2 \colon \mathcal{M}\times\mathcal{M} \to \mathcal{M}$ 为附录 B中定义的坐标投影。由于 $\mathrm{pr}_1 \circ (t, s) = t$ 且 $\mathrm{pr}_2 \circ (t, s) = s$，推前测度的复合规则（Teschl，2025年，§2.3，例 2.12）给出

$$(\mathrm{pr}_1)_{\#}\pi = (\mathrm{pr}_1)_{\#}(t, s)_{\#}\mu = (\mathrm{pr}_1 \circ (t, s))_{\#}\mu = t_{\#}\mu = \nu_t, \qquad (\mathrm{pr}_2)_{\#}\pi = \nu_s.$$

因此 $\pi \in \Pi(\nu_t,\ \nu_s)$。根据 W_2 的定义，

$$W_2(\nu_t, \nu_s)^2 = \inf_{\gamma\in\Pi(\nu_t,\nu_s)} \int_{\mathcal{M}\times\mathcal{M}} d(y, z)^2 \,\mathrm{d}\gamma(y, z) \leq \int_{\mathcal{M}\times\mathcal{M}} d(y, z)^2 \,\mathrm{d}\pi(y, z).$$

由于 $\mathcal{M}$ 是紧致的，$g(y, z) = d(y, z)^2$ 有界，因此关于 $\mathcal{M} \times \mathcal{M}$上的任意概率测度都可积。因此，我们可以应用推前测度的变量替换公式（Teschl，2025年，定理 2.16），

$$\int_{\mathcal{M}\times\mathcal{M}} d(y, z)^2 \,\mathrm{d}\pi(y, z) = \int_{\mathcal{M}} d\big(t(x), s(x)\big)^2 \,\mathrm{d}\mu(x).$$

开平方得到（23）。 □

H.10. 引理 H.6 的证明

引理 H.18 (极小值点的稳定性)。 设 K 为一个紧致度量空间，且 f_k、$f \colon K \to \mathbb{R}$ 连续并满足 $f_k \to f$一致。假设 f 有唯一极小值点 $z^\star \in K$。若对每个 k，$z_k \in \operatorname{argmin}_K f_k$，则 $z_k \to z^\star$。

引理的证明H.18推迟到附录H.11。现在我们准备证明引理H.6。

引理*H.6*的证明。令 $\phi \in \Psi_c(\mathcal{M})$ 并设 $\psi := \phi^c \in C(\mathcal{M}, \mathbb{R})$，从而 $\phi = \psi^c$。

步骤 1：构造 $\psi_k \in \varphi^*\mathcal{F}$ 的近似 $\psi = \phi^c$。 由于 $\mathcal{F}$ 在 $C(\mathbb{R}^n, \mathbb{R})$ 中关于ucc拓扑稠密，且 φ 满足假设2.2，因此可得（例如，参见（Kratsios & Bilokopytov，2020年，定理 3.3））， 拉回类 $\varphi^*\mathcal{F}$ 在$C(\mathcal{M}, \mathbb{R})$ 中关于一致收敛稠密。因此存在一个序列 $\psi_k \in \varphi^*\mathcal{F}$ 使得

$$\|\psi_k - \psi\|_\infty \xrightarrow[k\to\infty]{} 0.$$

定义 $\phi_k := \psi_k^c \in \mathfrak{C}(\varphi^*\mathcal{F})$。

步骤 2 $\phi_k \to \phi$ $x \in \mathcal{M}$ $y \in \mathcal{M}$ y $y \le$**：一致收敛。** 对于任意和任意，$\psi\ y\ + \|\psi_k - \psi\|_\infty$
我们有 $\psi() - \|\psi_k - \psi\|_\infty \le \psi_k() ()$，因此

$$c(x,y) - \psi(y) - \|\psi_k - \psi\|_\infty \le c(x,y) - \psi_k(y) \le c(x,y) - \psi(y) + \|\psi_k - \psi\|_\infty.$$

取 $\inf_{y\in\mathcal{M}}$ 可得

$$\phi(x) - \|\psi_k - \psi\|_\infty \le \phi_k(x) \le \phi(x) + \|\psi_k - \psi\|_\infty,$$

因此

$$\|\phi_k - \phi\|_\infty \le \|\psi_k - \psi\|_\infty \xrightarrow[k\to\infty]{} 0.$$

步骤 3：在一个满 μ-测度集上的可微性与梯度表示。每个 ϕ_k 和 ϕ 都是 c-凹的，因此在 $\mathcal{M}$ 上是利普希茨连续的（McCann, 2001，引理 2）。根据黎曼流形上的拉德马赫定理（McCann，2001，引理 4），ϕ 和每个 ϕ_k 都几乎处处（相对于 $\mathrm{vol}_{\mathcal{M}}$）可微，因此也几乎处处（相对于 μ）可微，因为 $\mu \ll \mathrm{vol}_{\mathcal{M}}$。用 $\mathrm{Diff}(\phi) \subset \mathcal{M}$ 和 $\mathrm{Diff}(\phi_k) \subset \mathcal{M}$ 表示 ϕ 和 ϕ_k 可微的点集。因此，对每个 k，$\mu(\mathrm{Diff}(\phi)) = 1$ 且 $\mu(\mathrm{Diff}(\phi_k)) = 1$。
此外，对于二次代价函数 $c(x,\ y) = \frac{1}{2}d(x,\ y)^2$, McCann (2001，引理 7）表明，在每个点 $x \in \mathrm{Diff}(\phi)$，其$c$-次微分 $\partial^c\phi(x)$ 均为单元素集。等价地，极小化元

$$y^\star(x) \in \arg\min_{y\in\mathcal{M}} \big(c(x,y) - \psi(y)\big)$$

是唯一的，且梯度满足一阶条件

$$\nabla\phi(x) = -\nabla_x c\big(x, y^\star(x)\big) = -\log_x\big(y^\star(x)\big).$$

对每个 ϕ_k，同样的结论也成立：对于任意 $x \in \mathrm{Diff}(\phi_k)$，极小化元 $y_k(x) \in \arg\min_{y\in\mathcal{M}}\big(c(x,y) - \psi_k(y)\big)$ 是唯一的，且

$$\nabla\phi_k(x) = -\nabla_x c\big(x, y_k(x)\big) = -\log_x\big(y_k(x)\big).$$

我们定义

$$N_\star := \mathrm{Diff}(\phi) \cap \bigcap_{k=1}^{\infty} \mathrm{Diff}(\phi_k).$$

那么 $N_\star$ 是可测的，并且满足 $\mu(N_\star) = 1$，因为它是一个由满 μ-测度的可测集组成的可数交。特别地，对于每个 $x \in N_\star$，梯度 $\nabla\phi(x)$ 和 $\nabla\phi_k(x)$ 都有定义，且唯一极小值点 $y^\star(x)$ 和 $y_k(x)$ 满足

$$\nabla\phi(x) = -\log_x\big(y^\star(x)\big), \qquad \nabla\phi_k(x) = -\log_x\big(y_k(x)\big).$$

在余下的证明中固定 $x \in N_\star$。

步骤 5：极小值点的稳定性 $y_k(x) \to y^\star(x)$。在紧集 $\mathcal{M}$ 上定义连续函数如下

$$h_x(y) := c(x,y) - \psi(y), \qquad h_{k,x}(y) := c(x,y) - \psi_k(y).$$

因此，$y^\star(x) = \arg\min h_x$ 和 $y_k(x) = \arg\min h_{k,\ x}$ 由 $N_\star$的构造可知是唯一的。此外，

$$\sup_{y\in\mathcal{M}} |h_{k,x}(y) - h_x(y)| = \sup_{y\in\mathcal{M}} |\psi_k(y) - \psi(y)| = \|\psi_k - \psi\|_\infty \to 0.$$

由于$h_{k,\ x} \to h_x$ 在紧空间上一致 $\mathcal{M}$ 且 h_x 具有唯一极小值点，引理H.18 表明

$$y_k(x) \to y^\star(x).$$

步骤 6：梯度的收敛。根据步骤 3，我们有恒等映射

$$\nabla\phi(x) = -\nabla_x c\big(x, y^\star(x)\big) = -\log_x\big(y^\star(x)\big), \qquad c(x,y) = \tfrac{1}{2}d(x,y)^2.$$

特别地，黎曼对数映射在 $y^\star(x)$ 处定义良好。

因此，$y^\star(x) \notin$ 割迹(x)。由于 $\mathcal{M} \setminus$割迹(x) 是开集，且 $\log_x$ 在其上光滑，因此存在一个以 $y^\star(x)$ 为中心的开邻域 $V_x \subset \mathcal{M} \setminus$割迹$(x)$，在该邻域上 $\log_x$ 光滑。等价地，存在 $r_x > 0$，使得

$$B_{r_x}\big(y^\star(x)\big) \subset V_x \subset \mathcal{M} \setminus \mathrm{Cut}(x).$$

根据步骤 5，我们已经知道 $y_k(x)$ 在 $\mathcal{M}$ 中收敛于 $\to y^\star(x)$。因此，根据度量空间中收敛的定义，对于上述半径 r_x，存在 $k_0(x)$，使得对于所有 $k \geq k_0(x)$，

$$d\big(y_k(x), y^\star(x)\big) < r_x, \qquad \text{so that} \qquad y_k(x) \in B_{r_x}\big(y^\star(x)\big) \subset V_x.$$

特别地，对于所有 $k \geq k_0(x)$，点 $y_k(x)$ 都位于 $\log_x$ 光滑（因而连续）的区域内。因此，

$$\log_x\big(y_k(x)\big) \longrightarrow \log_x\big(y^\star(x)\big),$$

并利用梯度表示，

$$\nabla\phi_k(x) = -\log_x\big(y_k(x)\big) \longrightarrow -\log_x\big(y^\star(x)\big) = \nabla\phi(x).$$

由于 $x \in N_\star$ wa s 任意且 $\mu(N_\star) = 1$，因此对 μ-几乎处处的 $x \in \mathcal{M}$，收敛性成立，从而证明完成 f.

□

H.11. 引理 H.18的证明

证明。 我们采用反证法证明所需结果。具体而言，假设 $z_k \to z^\star$ 不成立。那么

$$\exists \varepsilon_0 > 0 \quad \text{s.t.} \quad \forall N \geq \mathbb{N},\ \exists n \geq N \quad \text{s.t.} \quad d(z_n, z^\star) \geq \varepsilon. \tag{44}$$

首先，我们证明可以从(z_k)中提取一个子序列，使其始终与 $z^\star$至少相距 ε_0。然后，通过紧致性，我们证明该子序列存在一个收敛子序列，其收敛到不同于 $z^\star$的点 $\bar{z}$。

步骤 1：构造收敛子序列。令 $N_1 := 1$。则根据（44），存在 $k_1 \geq N_1$ 使得

$$d(z_{k_1}, z^\star) \geq \varepsilon_0. \tag{45}$$

设 $N_2 := k_1 + 1$。同样，存在 $k_2 \geq N_2$ 使得

$$d(z_{k_2}, z^\star) \geq \varepsilon_0.$$

我们继续归纳地进行：在已选定 k_j 后，令 $N_{j+1} := k_j + 1$。于是存在 $k_{j+1} \geq N_{j+1}$ 使得 t

$$d(z_{k_{j+1}}, z^\star) \geq \varepsilon_0.$$

根据构造，这些指标严格递增 $k_1 < k_2 < k_3 < \ldots$，因此 (z_{k_j}) 是 (z_k) 的一个子序列。此外，该子序列中的每一项与 $z^\star$ 至少相距 ε_0，即

$$d(z_{k_j}, z^\star) \geq \varepsilon_0, \quad \text{for all } j.$$

现在，(z_{k_j}) 是 K 中的一个序列，而 K 是紧致的。因此，存在 (z_{k_j}) 的一个更进一步的子序列（稍微滥用一下记号，我们仍将其记为 (z_{k_j})），以及一个点 $\bar{z} \in K$，使得

$$z_{k_j} \to \bar{z} \quad \text{as} \quad j \to \infty.$$

由于该子序列中的每一项都满足 $d(z_{k_j}, z^\star) \geq \varepsilon_0$，因此可得 $\bar{z} \neq z^\star$；否则，距离将收敛到零。

步骤 2：比较 $f(\bar{z})$ 与 $f(z^\star)$。 由于一致收敛蕴含逐点收敛，因此对任意固定点 $z \in K$，

$$f_k(z) \to f(z).$$

由于每个 su 收敛序列的 bsequence 本身也收敛，并且与原序列具有相同的极限 ，

$$f_{k_j}(z^\star) \to f(z^\star).$$

另一方面，考虑如下不等式

$$|f_{k_j}(z_{k_j}) - f(\bar{z})| \leq |f_{k_j}(z_{k_j}) - f(z_{k_j})| + |f(z_{k_j}) - f(\bar{z})|.$$

第一项因一致收敛而趋于零，而第二项因 f 与 $z_{k_j} \to \bar{z}$ 的连续性而趋于零。因此，

$$f_{k_j}(z_{k_j}) \to f(\bar{z}). \tag{46}$$

由于每个 z_{k_j} 都是 f_{k_j} 的极小化元，因此

$$f_{k_j}(z_{k_j}) \leq f_{k_j}(z^\star) \quad \text{for all } j. \tag{47}$$

现在我们令 $j \to \infty$ 趋于极限。由（46），左侧得到

$$f_{k_j}(z_{k_j}) \to f(\bar{z}).$$

由一致收敛，右侧得到

$$f_{k_j}(z^\star) \to f(z^\star).$$

由于对所有 j，$f_{k_j}(z_{k_j}) \leq f_{k_j}(z^\star)$，且 $f_{k_j}(z_{k_j}) \to f(\bar{z})$ 和 $f_{k_j}(z^\star) \to f(z^\star)$，因此

$$f(\bar{z}) \leq f(z^\star). \tag{48}$$

步骤 3：导出矛盾。 我们知道，$z^\star$ 是 K 上 f 的唯一极小值点。因此，

$$f(z^\star) \leq f(z) \quad \text{for all } z \in K \tag{49}$$

并且如果 $f(z) = f(z^\star)$，那么 $z = z^\star$。由 $z^\star$ 的极小性可知

$$f(z^\star) \leq f(\bar{z}). \tag{50}$$

结合（48）和（50），我们得到

$$f(\bar{z}) = f(z^\star). \tag{51}$$

因此，$\bar{z} = z^\star$，由极小化元的唯一性可知。因此我们得到了一个矛盾：子序列（z_{k_j}）始终与 $z^\star$相距至少ε_0，因此其极限 $\bar{z}$ 不可能等于 $z^\star$。由此我们得出结论，极小化元序列 z_k 必定收敛到唯一极小值点 $z^\star$：$z_k \to z^\star$。 □

H.12. 定理H.9的证明

我们的证明基 于Yarotsky & Zhevnerchuk （2020年，定理 3.3），为方便起见，我们在下文陈述该定理 .

定理 H.19 （Yarotsky & Zhevnerchuk （2020年）中的定理 3.3）。对于任意 $r > 0$，任意速率 $\beta \in (\frac{r}{n}, \frac{2r}{n}]$ 均可通过具有$L \leq c_{r个n} W^{\beta n/r-1}$ 层的深层 *ReLU* 网络实现。

定理 *H.9* 的证明。我们说明定理 H.9 如何由定理 H.19 在定义 H.7-H.8 的复杂度约定下得出。

步骤 1：来自 Yarotsky & Zhevnerchuk (2020年) 的定量结论。根据定理 H.19，对于每个 $r > 0$ 以及每个 $\beta \in (r/n, 2r/n]$，都存在常数 $C_{r,\, n,\, \beta} > 0$ 和 $C_{r,\, n} > 0$，使得对于每个整数 $W \geq 1$，都存在一个

ReLU 近似方案（由一个架构 η_W 和一个参数选择映射G_W组成，其中权重不超过 W），满足

$$\sup_{f\in\mathcal{H}_{r,n}} \left\|f-\tilde{f}_{\eta_W,G_W}(f)\right\|_\infty \le C_{r,n,\beta}\, W^{-\beta},$$

且其深度（层数，即我们的定义 H.8）满足

$$L_W \le C_{r,n}\, W^{\beta n/r-1}.$$

固定$f\in\mathcal{H}_{r,\ n}$并定义相应的网络

$$\hat{f}_W := \tilde{f}_{\eta_W,G_W}(f).$$

于是$\hat{}\ f_W\in\mathcal{NN}_{\mathrm{ReLU}n,\ L_W,\ W}$并且

$$\|f-\hat{f}_W\|_\infty \le C_{r,n,\beta}\, W^{-\beta}.$$

步骤 2：选择 W 作为 ε的函数。为方便起见，设

$$C'_{r,n,\beta} := \max\{1, C_{r,n,\beta}\}, \qquad \widetilde{C}_{r,n,\beta} := 2^\beta C'_{r,n,\beta}.$$

固定 $\varepsilon\in(0,\ 1)$，并定义阈值

$$\bar{W}_\varepsilon := \left(\frac{C'_{r,n,\beta}}{\varepsilon}\right)^{1/\beta}.$$

任选一个整数 $\bar{}$使 $W_\varepsilon\in\mathbb{N}$ 满足 $W_\varepsilon\ge\bar{W}_\varepsilon$（例如，$W_\varepsilon=\lceil\bar{W}_\varepsilon\rceil$）。由于$C_{r,\ n,\ \beta}\le C'_{r,\ n,\ \beta}$，我们得到

$$C_{r,n,\beta}W_\varepsilon^{-\beta} \le C'_{r,n,\beta}W_\varepsilon^{-\beta} \le C'_{r,n,\beta}\bar{W}_\varepsilon^{-\beta} = \varepsilon.$$

$\hat{}\hat{}$取 $f_\varepsilon :=f_{W_\varepsilon}$，因此我们有

$$\|f-\hat{f}_\varepsilon\|_\infty \le \varepsilon.$$

$\bar{}$**步骤 3：权重界**。取 $W_\varepsilon := \lceil\bar{W}_\varepsilon\rceil$。由于 $\varepsilon\in(0,\ 1)$ 且$C'_{r,\ n,\ \beta}\ge 1$，我们有 $W_\varepsilon\ge 1$，因此对所有$x\ge 1$，$\lceil x\rceil\le 2x$。于是，

$$W_\varepsilon \le 2\bar{W}_\varepsilon = 2\left(\frac{C'_{r,n,\beta}}{\varepsilon}\right)^{1/\beta} = \left(\frac{\widetilde{C}_{r,n,\beta}}{\varepsilon}\right)^{1/\beta}.$$

步骤 4：深度界。根据步骤 1 以及 W_ε的选择，

$$L_\varepsilon := L_{W_\varepsilon} \le C_{r,n}\, W_\varepsilon^{\beta n/r-1}.$$

这恰好是按定义H.8.中的约定给出的深度估计。

步骤 5：将 ReLU 替换为一般的分段线性激活函数。设 $\sigma:\ \mathbb{R}\to\mathbb{R}$ 是任意非仿射分段$\hat{}$线性激活函数。根据Yarotsky (2017 年，命题 1(b))，在有界域上（此处为 $[0,1]^n$），任意 ReLU 神经网络 f_ε都可以转换为一个激活函数为 σ 的神经网络，使其在 $[0,1]^n$上计算相同的函数。更准确地说，存在一个常数 $K_\sigma\ge 1$（仅依赖于 σ）以及一个激活函数为$\hat{}\ f_\varepsilon^{(\sigma)}$的神经网络 σ，使得

$$\hat{f}_\varepsilon^{(\sigma)}(x) = \hat{f}_\varepsilon(x) \quad \text{for all } x\in[0,1]^n,$$

且其深度 $L_\varepsilon^{(\sigma)}$和参数数量 $W_\varepsilon^{(\sigma)}$满足

$$L_\varepsilon^{(\sigma)} = L_\varepsilon, \qquad W_\varepsilon^{(\sigma)} \le K_\sigma\, W_\varepsilon.$$

特别地，

$$\|f-\hat{f}_\varepsilon^{(\sigma)}\|_\infty = \|f-\hat{f}_\varepsilon\|_\infty \le \varepsilon.$$

因此，调整常数 $\sim C_{r,\ n,\ \beta}$（调整因子仅依赖于 σ，且不改变深度阶数）后，激活函数为 σ 的神经网络也满足相同的逼近结论。

□