arXiv:2301.01766v1 [math.ST] 4 Jan 2023

使用 Wasserstein-Fisher-Rao 梯度流学习高斯混合

Yuling Yan*† Kaizheng Wang*‡ Philippe Rigollet§

2023年1月5日

摘要

高斯混合模型构成了一类灵活且表现力丰富的分布参数族，已广泛应用于各个领域。然而，从计算角度来看，将这些模型拟合到数据上是出了名的难题。目前，只有基于矩的方法具有理论保证，而基于似然的方法则依赖诸如期望最大化之类的启发式算法；众所周知，这些算法在简单示例中也会失效。在这项工作中，我们提出一种新算法，用于计算高斯混合模型中的非参数最大似然估计量（NPMLE）。我们的方法基于在配备 Wasserstein-Fisher-Rao 几何的概率测度空间上进行梯度下降，并为此建立了收敛保证。在实践中，该方法可以通过相互作用粒子系统进行近似，其中粒子的权重和位置交替更新。我们进行了广泛的数值实验，以验证所提出算法的有效性，并将其与经典基准方法以及相对于更简单几何结构的类似梯度下降算法进行比较。具体而言，这些模拟展示了同时更新相互作用粒子的权重和位置所带来的优势。

关键词： 高斯混合模型，非参数最大似然估计，Wasserstein-Fisher-Rao几何，最优传输，Wasserstein梯度流，过参数化

1 引言

由于灵活多变，自皮尔逊于十九世纪引入以来，混合模型便成为统计建模的核心对象。然而，这种灵活性往往伴随着计算成本：这类模型通常难以拟合，而且直到最近，混合模型的计算问题一直未受到重视。时至今日，理论与实践仍存在分歧：前者主要关注矩方法，后者则以变分方法为主，尤其是最大似然。本文旨在通过开发具有理论保证的最大似然估计算法，缩小这一差距。

考虑独立同分布样本 $\{X_i\}_{1\le i\le N} \in \mathbb{R}^d$，它们来自各向同性高斯分布[1]混合分布 $\rho^\star * \mathcal{N}(0,\ I_d)$，其密度函数为

$$(\rho^\star * \phi)(x) = \int_{\mathbb{R}^d} \phi(x-y)\,\rho^\star(\mathrm{d}y),$$

其中 $\rho^\star$ 是定义在 $\mathbb{R}^d$上的混合分布，而 $\phi(x) = (2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$ 是各向同性高斯分布 $\mathcal{N}(0, I_d)$ 的密度函数。目标是根据 N个样本学习高斯混合分布 $\rho^\star * \mathcal{N}(0, I_d)$。

*前两位作者贡献相同。

†普林斯顿大学运筹学与金融工程系，美国新泽西州普林斯顿，邮编 08544；电子邮箱：yulingy@princeton.edu。

‡哥伦比亚大学工业工程与运筹学系，美国纽约州纽约市，邮编 10027；电子邮箱：kaizheng.wang@columbia.edu。§麻省理工学院数学系，美国马萨诸塞州剑桥市，邮编 02139；电子邮箱：rigollet@math.mit.edu。[1]所有这些工作都可以推广到一般混合模型。一般性处理请参见附录。

该问题的负对数似然定义为

$$\ell_N(\rho) = -\frac{1}{N}\sum_{i=1}^{N}\log\left[(\rho * \phi)(X_i)\right].$$

虽然 ℓ_N 本身是一个平凡的凸泛函 ρ，但对其进行最小化的测度类往往并非凸集。例如，在有限混合模型中，ρ 被限制为支撑在至多 k 个原子上的测度。约束缺乏凸性是这一问题计算困难的主要来源。为克服这一局限，Kiefer and Wolfowitz (1956年)提出了非参数最大似然估计（NPMLE），其做法是在集合 ℓ_N 上对 $\mathcal{P}(\mathbb{R}^d)$ 中所有 $\mathbb{R}^d$上的概率分布进行最小化：

$$\widehat{\rho} \in \underset{\rho\in\mathcal{P}(\mathbb{R}^d)}{\operatorname{argmin}}\ \ell_N(\rho). \tag{1.1}$$

虽然 $\mathcal{P}(\mathbb{R}^d)$ 是凸的，但它是无限维的。事实上，非参数最大似然估计可以看作过参数化的极端情形；这一现象正对深度学习理论中的传统统计学观点构成挑战（Allen-Zhu等人，2019年；Chizat and Bach, 2018年）。事实上，当 $d = 1$ 时，(1.1) 的解具有有趣的结构性质。具体而言，在这种情况下，优化问题 (1.1) 存在唯一解 $\widehat{\rho}$ （Jewell，1982；Lindsay和Roeder，1993年），并且一般而言，该解的支撑集至多包含 N 个原子（Lindsay，1983年）。当 $\rho^\star$ 为次高斯时，Polyanskiy和Wu（2020年）将这一上界改进为 $O_{\mathbb{P}}(\log N)$。颇为引人注目的是，人们对维度为 $d \geq 2$ 时非参数最大似然估计的结构所知甚少。据我们所知，以下定理 1 首次证明了在任意维度下，(1.1) 都存在一个解。

经典统计学结果为以下估计量提供了海灵格风险界： $\widehat{\rho} * \phi$ ，用于估计 $\rho^\star * \phi$ (Dicker和Zhao，2016；Saha和Guntuboyina，2020；Zhang，2009）。即使考虑更大的 C^∞ 密度函数类，这些收敛率在对数因子范围内仍与极小极大最优性相当。

尽管已有显著进展，非参数最大似然估计的计算方面仍未得到充分探索。大多数相关工作采用离散化方案，预先设置细网格，并求解（1.1），同时附加约束 ρ 的支撑集位于该网格上（Jiang和Zhang，2009年；Koenker和Mizera，2014；Lindsay，1983年；Zhang等人，2022）。尽管这些方法在理论上已有充分认识，却受到维数灾难的困扰，其复杂度随维度 d呈指数增长。为克服这一局限，Zhang等人(2022)提出了一种启发式方法，使用期望最大化（EM）算法交替更新权重和支撑集，但该方法缺乏理论保证。另一项值得关注的工作是Groeneboom等人(2008)提出的支撑缩减算法；该算法在计算上同样效率低下，因为它要求在每次迭代中计算 $\mathbb{R}^d$ 处一个非凸函数的极小化子。

在这项工作中，我们提出在赋予 Wasserstein-Fisher-Rao（WFR）几何1.1的概率测度空间中，使用梯度下降求解（Chizat 等人，2018年；GallouëtandMonsaingeon，2017；Kondratyev 等人，2016；Liero 等人，2018年）。具体而言，我们引入负对数似然的 WFR 梯度流 ℓ_N ，并证明在温和条件下，它会收敛至非参数最大似然估计。进而，我们分别在时间上采用欧拉离散化、在空间上采用粒子离散化来实现这一 WFR 梯度流，从而得到一个加权相互作用粒子系统。本质上，所得算法 1交替更新位置和权重，与上文所述的期望最大化算法类似；但 WFR 梯度流所产生的更新在本质上有所不同。

顾名思义，WFR 几何是 Wasserstein 几何（Ambrosio 等人，2008；Otto，2001）与 Fisher-Rao 几何（Bauer 等人，2016）的复合。前者决定支撑集的更新，后者决定权重的更新。这两种几何各自都导出一种梯度下降算法，但我们的数值结果表明，将二者结合起来是实现快速收敛的关键；参见第4节。事实上，我们证明，既有文献中的固定位置 EM 算法（例如参见Jiang和Zhang，2009年）实现了关于 Fisher-Rao 几何的梯度下降。作为分析学的一个副产品，我们还证明，在一定条件下，该算法在无限粒子情形下收敛到非参数最大似然估计。

2 非参数最大似然估计量（NPMLE）

本节考察优化问题（1.1）的最优性条件。为此，记 $\delta\ell_N(\rho)$ 为 ℓ_N 在测度 ρ 处的一阶变分[2]，并注意到它可表示为

$$\delta\ell_N(\rho): x \mapsto -\frac{1}{N}\sum_{i=1}^{N}\frac{\phi(x-X_i)}{(\rho*\phi)(X_i)}. \tag{2.1}$$

下面的定理证明了 NPMLE 在一般维度下的存在性，并给出了其最优性条件。证明见附录 B.1。

定理 1. *NPMLE* 满足以下性质：

1.（存在性）优化问题 (1.1) 存在极小化子。

2.（最优性条件）当且仅当（*i*） $\delta\ell_N(\widehat{\rho})(x) \geq -1$ 对所有 $x \in \mathbb{R}^d$成立，且（*ii*） $\delta\ell_N(\widehat{\rho})(x) = -1$ 对 $\widehat{\rho}$-几乎处处的 x 成立时，分布 $\widehat{\rho} \in \mathcal{P}(\mathbb{R}^d)$ 是 *NPMLE*。

注 1. 我们在附录 B.1 中证明，对任意 $\rho \in \mathcal{P}(\mathbb{R}^d)$， $\int \delta\ell_N(\rho)\mathrm{d}\rho = -1$ 总是成立。因此，定理 1 中的最优性条件（ii）可由（i）推出，这意味着仅需满足 $\delta\ell_N(\widehat{\rho})(x) \geq -1$ 对所有 $x \in \mathbb{R}$成立，就已经是 $\widehat{\rho}$ 成为 NPMLE 的充要条件。不过，我们在定理中保留这两个条件，因为它们各自揭示了 NPMLE 的重要结构信息。

定理1 断言 NPMLE 的存在性，但当 $d \geq 2$ 时，其唯一性仍是一个开放问题。尽管唯一性问题尚未解决，本文的收敛理论仍然成立：在这种情况下，NPMLE 指的是（1.1）的任意极小化子。

记号。我们用 $\mathcal{P}(\mathbb{R}^d)$ 表示 $\mathbb{R}^d$ 上概率测度的空间，用 $\mathcal{P}_2(\mathbb{R}^d)$ 表示 $\mathbb{R}^d$ 上具有有限二阶矩的概率测度的空间，并用 $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ 表示相对于 $\mathbb{R}^d$ 上勒贝格测度绝对连续的概率测度空间。令 Δ^{m-1} 为 $m-1$ 维概率单纯形。对于任意 $\rho \in \mathcal{P}(\mathbb{R}^d)$，$\mathrm{supp}(\rho)$ 表示其支撑集，即满足 $\rho(C) = 1$ 的最小闭集 $C \subseteq \mathbb{R}^d$。对于任意映射 $T: \mathbb{R}^d \to \mathbb{R}^d$ 和任意分布 $\rho \in \mathcal{P}(\mathbb{R}^d)$，令 $T_{\#}\rho$ 表示 ρ 经 T 的推前（或像测度），其定义为：对于 $\mathbb{R}^d$ 中任意博雷尔集 A，有 $T_{\#}\rho(A) = \rho(T^{-1}(A))$。对于任意 $x \in \mathbb{R}^d$，我们用 δ_x 表示点 x 处的狄拉克质量。对于两个概率测度 μ、$\nu \in \mathcal{P}(\mathbb{R}^d)$，我们用 $\mu \ll \nu$ 表示 μ 相对于 ν 绝对连续。对于 $\mathcal{P}(\mathbb{R}^d)$ 中的序列 $\{\rho_n\}_{n=0}^{\infty}$ 和 $\rho \in \mathcal{P}(\mathbb{R}^d)$，若 ρ_n 弱收敛于 ρ，即对每个有界连续函数 f: $\mathbb{R}^d \to \mathbb{R}$，均有 $\int_{\mathbb{R}^d} f(x)\rho_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} f(x)\rho(\mathrm{d}x)$，则记作 $\rho_n \xrightarrow{\mathrm{w}} \rho$。令 $C_c^\infty(\mathbb{R}^d)$ 为 $\mathbb{R}^d$ 上具有紧支撑的光滑函数集合。若对于任意 $\varphi \in C_c^\infty(\mathbb{R}^d)$，均有以下关系成立，则称 $(\rho_t)_{t\geq 0}$ 是偏微分方程（PDE） $\partial_t\rho_t = -\mathrm{div}(\rho_t v_t) + \rho_t\alpha_t$ 的分布解，其中 $v_t: \mathbb{R}^d \to \mathbb{R}^d$ 且 $\alpha_t: \mathbb{R}^d \to \mathbb{R}$。

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(x)\,\rho_t(\mathrm{d}x) = \int_{\mathbb{R}^d}\left[\langle\nabla\varphi(x), v_t(x)\rangle + \varphi(x)\alpha_t(x)\right]\rho_t(\mathrm{d}x).$$

最后，我们用 ODE 作为常微分方程的简称。

3 Wasserstein-Fisher-Rao 梯度下降

概率测度上的梯度流是开发采样算法的有用工具，其目标是从目标测度中生成样本。一个经典例子是 π，例如只知道归一化常数的贝叶斯后验。在这一背景下，Wasserstein-Fisher-Rao（WFR）梯度流近来已成为标准 Wasserstein 梯度流的一种有力替代方案。事实上，它们是Lu 等人(2019a)提出的生灭采样算法，以及Lu 等人(2022年)提出的基于粒子的方法的基础。关于采样以及 Wasserstein 梯度流在这一背景下的作用，读者可参阅Chewi (2022年)近期发表的手稿，了解更多详情。

[2]具体计算可由变分法中的标准论证得出，故放在附录C.1。

本文关注的似然最大化问题与采样问题不同，因为它旨在优化另一个目标。不过，它仍然是一个优化问题，因此可以在此背景下运用概率测度空间上的梯度流工具。据我们所知，本文首次尝试了这种应用。具体而言，我们用于求解（1.1）的主要算法依赖于 Wasserstein-Fisher-Rao 梯度流的一种特定离散化。我们先简要介绍概率测度度量空间上的梯度流，专家读者可以放心略过这一部分。

3.1 概率测度度量空间上的梯度流 s

概率测度度量空间上的梯度流是变分法的核心主题，它已应用于从分析学、几何学到概率论和统计学等众多领域。我们简要介绍这一强大工具背后的主要思想，并请读者参阅Ambrosio 等人(2008年的权威著作，以了解这一深奥问题的更多详情。

回顾一下，我们的目标是在概率测度空间上为泛函 ℓ_N 导出一个梯度流。这个梯度流本质上是一条曲线 $(\rho_t)_{t\geq 0}$，满足 $\partial_t \rho_t = -\nabla \ell_N(\rho_t)$，其中梯度的概念 ∇ 将在下文定义。在其开创性工作中，Jordan 等人(1998)通过类比欧几里得梯度流，在无需实际定义梯度的情况下，成功地在 Wasserstein 空间上定义了梯度流。我们沿用他们的方法，定义泛函 ℓ_N 相对于适当几何结构的梯度流；该几何结构的测地距离为 $d(\cdot,\cdot)$，其定义在概率测度空间上，如下所示：

$$\partial_t \rho_t = \lim_{\eta \to 0} \frac{\rho_t^\eta - \rho_t}{\eta},$$

其中

$$\rho_t^\eta := \operatorname*{arg\,min}_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \delta \ell_N(\rho_t) \, \mathrm{d}(\rho - \rho_t) + \frac{1}{2\eta} \mathsf{d}^2(\rho, \rho_t) \right\}$$

给定一个距离，是否存在极限绝对连续曲线 $(\rho_t)_{t\geq 0}$，是一个重要且核心的问题；本综述不作讨论。

本文主要关注 Wasserstein-Fisher-Rao 距离，它是 Fisher-Rao 距离与（二次）Wasserstein 距离的组合。现在我们来介绍这三种距离。回顾一下，两点之间的测地距离衡量连接这两点的最短曲线。这三种距离的差异取决于赋予概率分布的微分结构，大致而言，这对应于允许采用的曲线类型。随后，使用黎曼度量来衡量这些曲线的长度；所有情况下的度量都相当直接，因此我们将重点讨论曲线。反过来，这些曲线及其长度定义了概率测度空间上的几何结构。

Fishe r-Rao 距离。 Fisher-Rao 距离与以下形式的反应方程相关：

$$\partial_t \rho_t = \rho_t \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \Big), \tag{3.1}$$

其中 $\alpha_t(x) \in \mathbb{R}$ 是一个标量，用于控制在 $x \in \mathbb{R}^d$ 和时间 t处生成的质量。很容易看出，这些动力学保持概率分布的总质量为 1。在所有连接 ρ_0与 ρ_1的此类曲线中，Fisher-Rao 测地线是总长度最短的一条。更具体地说，Fisher-Rao 距离 d_{FR} 定义如下（Bauer 等人，2016）：

$$d_{\mathsf{FR}}^2(\rho_0, \rho_1) = \inf \Bigg\{ \int_0^1 \int \Big[\Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \Big)^2 \Big] \mathrm{d}\rho_t \mathrm{d}t : \ (\rho_t, \alpha_t)_{t\in[0,1]} \text{ solves} \\ \partial_t \rho_t = \rho_t \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \Big) \Bigg\}. \tag{3.2}$$

虽然这对我们的问题没有用处，但值得注意的是

$$d_{\mathsf{FR}}^2(\rho_0, \rho_1) = 4 \int \Big| \sqrt{\frac{\mathrm{d}\rho_0}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\rho_1}{\mathrm{d}\lambda}} \Big|^2 \mathrm{d}\lambda,$$

其中， λ 是任意满足 ρ_0， $\rho_1 \ll \lambda$ 的正测度（例如， $\lambda = \rho_0 + \rho_1$。）。因此，除一个常数因子外，Fisher-Rao 距离其实就是统计学家熟知的海林格距离。

Wasserstein 距离。 给定两个概率测度 ρ_0， $\rho_1 \in \mathcal{P}(\mathbb{R}^d)$，与 ρ_1 之间的（二次）Wasserstein 距离定义为（Villani，2009年）

$$d_{\mathsf{W}}^2(\rho_0, \rho_1) = \inf_{\pi \in \Pi(\rho_0, \rho_1)} \int \|x - y\|_2^2 \, \pi(\mathrm{d}x, \mathrm{d}y),$$

其中，下确界是在 ρ_0 和 ρ_1的所有耦合上取得的。

它还 可以通过 Benamou-Brenier公式解释为测地距离 :

$$d_{\mathsf{W}}^2(\rho_0, \rho_1) = \inf \left\{ \int_0^1 \int \|v_t\|^2 \, \mathrm{d}\rho_t \mathrm{d}t : (\rho_t, v_t)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathsf{div}(\rho_t v_t) \right\}. \tag{3.3}$$

其中，可容许曲线由连续性方程给出

$$\partial_t \rho_t = -\mathsf{div}(\rho_t v_t), \tag{3.4}$$

该方程描述了粒子密度在 $\mathbb{R}^d$ 中的演化，其遵循时变向量场 $v_t : \mathbb{R}^d \to \mathbb{R}^d$。

Wasserstein-Fisher-Rao 距离。 Wasserstein-Fisher-Rao 距离背后的几何结构，是通过满足以下演化方程的曲线构建的:

$$\partial_t \rho_t = -\mathsf{div}(\rho_t v_t) + \rho_t \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \Big).$$

注意，右侧恰好是反应方程（3.1）和连续性方程（3.4）右侧之和，它们分别支配 Fisher-Rao 几何和 Wasserstein 几何。因此，它是这两种几何结构的复合。

接着，Wasserstein-Fisher-Rao 距离 d_{WFR} 定义为（Chizat 等，2018年；Gallouët 和 Mon- saingeon，2017；Kondratyev 等，2016；Liero 等，2018年）

$$d_{\mathsf{WFR}}^2(\rho_0, \rho_1) = \inf \left\{ \int_0^1 \int \Big[\|v_t\|^2 + \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \Big)^2 \Big] \mathrm{d}\rho_t \mathrm{d}t : (\rho_t, v_t, \alpha_t)_{t\in[0,1]} \text{ solves } \right.$$
$$\left. \partial_t \rho_t = -\mathsf{div}(\rho_t v_t) + \rho_t \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \Big) \right\}. \tag{3.5}$$

有了这个距离，我们便可以定义 WFR 梯度流及其时间离散化形式——WFR 梯度下降。

3.2 Wasserstein-Fisher-Rao 梯度下降

In t本节介绍我们的主要算法：Wasserstein-Fisher-Rao 梯度下降 nt.

负对数似然 $\ell_N(\rho)$ 在 $\mathcal{P}_2(\mathbb{R}^d)$ 下关于 Wasserstein-Fisher-Rao 距离 $d_{\mathsf{WFR}}(\cdot,\cdot)$ 的梯度流 $\{\rho_t\}_{t\geq 0}$ 为

$$\partial_t \rho_t = -\left[1 + \delta \ell_N(\rho_t)\right] \rho_t + \mathsf{div}(\rho_t \nabla \delta \ell_N(\rho_t)). \tag{3.6}$$

式（3.6的形式推导基于 Wasserstein-Fisher-Rao度量所蕴含的黎曼结构，详见附录 C.4。

WFR 梯度流难以直接实现，原因有两个。首先，它是在连续时间中描述的。其次，它需要处理完整的概率测度 ρ_t ，而这类对象定义在 $\mathbb{R}^d$ 上，且具有无限维。

为克服第一个障碍，我们采用一种简单的时间离散化方案，得到 WFR 梯度下降算法。该算法会生成一列概率测度 $\{\rho_n\}_{n\geq 0}$，并交替执行以下两个步骤：

$$\frac{\mathrm{d}\widetilde{\rho}_n}{\mathrm{d}\rho_n} = 1 - \eta\left[1 + \delta\ell_N\left(\rho_n\right)\right] \qquad \text{(Fisher-Rao gradient update)} \tag{3.7a}$$

$$\rho_{n+1} = \left[\mathrm{id} - \eta\nabla\delta\ell_N\left(\widetilde{\rho}_n\right)\right]_{\#}\widetilde{\rho}_n \qquad \text{(Wasserstein gradient update)} \tag{3.7b}$$

对于 $n = 0, 1, \ldots\ldots$，其中 $\eta > 0$ 是步长。注意，每一项都对应于（3.6）右侧的一个加项。事实上，（3.7a）是相对于 Fisher-Rao 几何的一步梯度下降更新，而（3.7b）则对应于相对于 Wasserstein 几何的一步梯度更新。还值得一提的是，（3.7）与 Gallouëtand Monsaingeon（2017）提出的分裂格式有关：（3.7）可视为对（3.6）进行数值近似的前向欧拉格式（显式格式），而 Gallouët and Monsaingeon（2017）采用后向欧拉格式（隐式格式）。后者的实现要求在每次迭代中对概率测度进行优化，这是一个难题。

下面的定理表明，当初始分布在整个空间上均赋予权重时，Wasserstein-Fisher-Rao 梯度下降会收敛。证明见附录 D。事实上，只需考虑一些额外的技术细节，便可得出关于连续时间 Wasserstein-Fisher-Rao 梯度流的类似结果。

定理 2（收敛至非参数最大似然估计）。假设初始化 $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ 满足$\mathrm{supp}(\rho_0) = \mathbb{R}^d$。考虑（3.7）中定义的 *Wasserstein-Fisher-Rao* 梯度下降 $\{\rho_n\}_{n\geq 0}$。存在由样本 $\{X_i\}_{1\leq i\leq N}$ 决定的 $\eta_0 > 0$，使得如果 $0 < \eta \leq \eta_0$ 且当 $n \to \infty$ 时 $\rho_n \xrightarrow{\mathrm{w}} \widehat{\rho}$，那么 $\widehat{\rho}$ 就是非参数最大似然估计。

定理 2 中的收敛结果是有条件的：我们只证明，如果 WFR 梯度下降弱收敛到某个极限分布 $\widehat{\rho}$，那么 $\widehat{\rho}$ 就是非参数最大似然估计。这与 Chizat and Bach （2018年）在研究浅层神经网络的训练动力学时所建立的收敛理论类似。由于缺乏测地凸性，我们无法建立无条件的全局收敛保证。

为克服第二个障碍，我们还需要在空间上对 ρ_n 进行离散化。得益于瓦瑟斯坦更新（3.7b），我们无需像以往用于非参数最大似然估计的算法那样使用固定网格。相反，注意，如果我们将 WFR 梯度下降初始化为一个支撑在 m 个原子上的分布，那么 ρ_n 在所有m 时刻都仍支撑在 n个原子上。在这种情况下，ρ_n 描述了带有权重的 m 相互作用粒子的演化。

更具体地说，考虑如下初始化

$$\rho_0 = \sum_{l=1}^{m}\omega_0^{(l)}\delta_{\mu_0^{(l)}}, \quad \text{where } \mu_0^{(1)}, \ldots, \mu_0^{(m)} \in \mathbb{R}^d \text{ and } \omega_0 = (\omega_0^{(1)}, \ldots, \omega_0^{(m)}) \in \Delta^{m-1},$$

其中，粒子的位置独立地从数据点 $\{X_i\}_{1\leq i\leq N}$中采样。下述定理精确刻画了从 ρ_0初始化的 Wasserstein-Fisher-Rao 梯度流。证明见附录 E.3。

定理 3 (粒子 Wasserstein-Fisher-Rao 梯度流)。耦合常微分方程组

$$\dot{\mu}_t^{(j)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i - \mu_t^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(X_i - \mu_t^{(l)}\right)}\left(X_i - \mu_t^{(j)}\right), \tag{3.8a}$$

$$\dot{\omega}_t^{(j)} = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i - \mu_t^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(X_i - \mu_t^{(l)}\right)} - 1\right]\omega_t^{(j)}, \tag{3.8b}$$

以初始化 $\mu_0^{(1)}, \ldots \mu_0^{(m)} \overset{\text{独立同分布}}{\sim}$ 均匀分布$(\{X_i\}_{1\leq i\leq N})$ 开始，且 $\omega_0 = [\omega_0^{(j)}]_{1\leq j\leq m} \in \Delta^{m-1}$ 在任意时间区间 $[0, T]$上都有唯一解。此外，定义为$(\rho_t)_{t\geq 0}$ 的流为

$$\rho_t := \sum_{l=1}^{m}\omega_t^{(l)}\delta_{\mu_t^{(l)}} \tag{3.9}$$

is t Wasserstein-Fisher-Rao 梯度流，即偏微分方程的一个分布解 (3.6).

在实践中，我们可以通过离散化常微分方程组（3.8）获得 Wasserstein-Fisher-Rao 梯度流的时间离散化，从而得到 Wasserstein-Fisher-Rao 梯度下降算法，如算法 1所示。

算法 1 Wasserstein-Fisher-Rao 梯度下降。

输入：数据 $\{X_i\}_{1\le i\le n}$，粒子数量 m，步长 $\eta>0$，最大迭代次数 t_0。**初始化**：从均匀分布 $\{X_i\}_{1\le i\le n}$ 中抽取 $\mu_0^{(1)}$、…、$\mu_0^{(m)}$ 独立同分布∼，并设置 $\omega_0^{(1)}=\cdots=\omega_0^{(m)}=1/m$。**更新**：**对于** $t=0$，**1**、…、t_0，**执行**

$$\mu_{t+1}^{(j)}=\mu_t^{(j)}+\eta\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(X_i-\mu_t^{(l)}\right)}\left(X_i-\mu_t^{(j)}\right),$$

$$\omega_{t+1}^{(j)}=\omega_t^{(j)}+\eta\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_{t+1}^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(X_i-\mu_{t+1}^{(l)}\right)}-1\right]\omega_t^{(j)},$$

对所有 $j=1$、…、m 均成立。这里，$\phi(x)=(2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$ 是 $\mathcal{N}(0,\ I_d)$ 的概率密度函数。**输出** $\rho=\sum_{j=1}^{m}\omega_{t_0}^{(j)}\delta_{\mu_{t_0}^{(j)}}$ 作为（近似）非参数最大似然估计。

用于证明定理2 的证明技巧遗憾地并不适用于算法1 ，因为初始测度并非支撑于整个空间。不过，由于已知在某些情况下，非参数最大似然估计（NPMLE）的支撑集只包含少量原子（Polyanskiy和Wu, 2020年），因此，只要取 m足够大，就很可能足以建立收敛结果。第4节的数值实验支持了这一猜想。

3.3 替代几何

如上所述，Wasserstein-Fisher-Rao（WFR）几何由 Wasserstein 几何与 Fisher-Rao 几何复合而成。相应地，WFR 梯度下降交替进行 Fisher-Rao 梯度步和 Wasserstein 梯度步。这一观察引出了一个迫切的问题：WFR 梯度下降的复合性质是否是取得良好效果的必要条件，还是这两个构成部分中的任意一个单独使用就足够？在本小节中，我们将探讨这两种自然替代方案的性质与局限性。

Fisher-Rao 梯度下降。我们在附录C.2.1 中证明，Fisher-Rao梯度流$(\rho_t)_{t\ge0}$ 函数 $\ell_N(\rho)$ 由以下偏微分方程定义：

$$\partial_t\rho_t=-\left[1+\delta\ell_N\left(\rho_t\right)\right]\rho_t. \tag{3.11}$$

通过时间离散化，可以轻松得到 Fisher-Rao 梯度下降更新规则 $\{\rho_n\}_{n\ge0}$：

$$\frac{\mathrm{d}\rho_{n+1}}{\mathrm{d}\rho_n}=1-\gamma\left[1+\delta\ell_N\left(\rho_n\right)\right],\qquad n=0,1,\ldots, \tag{3.12}$$

其中 $\gamma>0$ 是步长。尽管 Fisher-Rao梯度流/梯度下降是借助黎曼几何以抽象方式推导出来的，但它与一些众所周知的算法有着密切联系。

1. 将 *Fisher-Rao* 梯度流视为近端梯度流。Fisher-Rao 梯度流（3.11）可视为 Fisher-Rao 度量下近端梯度算法的连续时间极限：

$$\partial_t\rho_t=\lim_{\eta\to0+}\frac{\rho_t^\eta-\rho_t}{\eta},\quad\text{where}\quad\rho_t^\eta:=\operatorname*{arg\,min}_{\rho\in\mathcal{P}(\mathbb{R}^d)}\left\{\int_{\mathbb{R}^d}\delta\ell_N\left(\rho_t\right)\mathrm{d}\left(\rho-\rho_t\right)+\frac{1}{2\eta}d_{\mathsf{FR}}^2\left(\rho,\rho_t\right)\right\}. \tag{3.13}$$

2. 将 *Fisher-Rao* 梯度流视为镜像流。Fisher-Rao 梯度流（3.11）也可视为采用 Kullback-Leibler（KL）散度的 $\ell_N(\rho)$ 镜像下降算法的连续时间极限

$$\partial_t \rho_t = \lim_{\eta \to 0+} \frac{\rho_t^\eta - \rho_t}{\eta}, \quad \text{where} \quad \rho_t^\eta := \arg\min_{\rho \ll \rho_t} \left\{ \int_{\mathbb{R}^d} \delta \ell_N(\rho_t) \, \mathrm{d}(\rho - \rho_t) + \frac{1}{\eta} \mathsf{KL}(\rho \,\|\, \rho_t) \right\}. \tag{3.14}$$

3. 将 *Fisher-Rao* 梯度下降视为固定位置 *EM* 算法。当 ρ_0 为离散分布时，步长为 $\gamma = 1$ 的 Fisher-Rao 梯度下降（3.12）与例如 Jiang和Zhang (2009年）中针对高斯混合模型的固定位置 EM 算法一致。

上述三种联系的形式化论证见附录C.2.2。下述定理表明，当 ρ_0 经过扩散时，Fisher-Rao 梯度下降具有良好的收敛性质。证明见附录 D。

定理 4（收敛至非参数最大似然估计）。 假设初始化 $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ 满足 $\mathrm{supp}(\rho_0) = \mathbb{R}^d$。考虑定义于(3.12)中的 *Fisher-Rao* 梯度下降 $\{\rho_n\}_{n \geq 0}$ 。存在由样本 $\{X_i\}_{1 \leq i \leq N}$确定的 η_0 ，使得若 $0 < \eta \leq \eta_0$ 且当 $n \to \infty$时 $\rho_n \xrightarrow{\mathrm{w}} \widehat{\rho}$，则 $\widehat{\rho}$ 为非参数最大似然估计。

我们还可以证明连续时间Fisher-Rao梯度流 $(\rho_t)_{t \geq 0}$ （即偏微分方程（3.11）的分布解）也有类似的结果：若 $\rho_t \xrightarrow{\mathrm{w}} \widehat{\rho}$ 满足 $t \to \infty$，则 $\widehat{\rho}$ 是一个非参数最大似然估计。

定理4 表明，当从充分分散的分布进行初始化时，Fisher-Rao梯度流/下降具有收敛保证。然而在实践中，我们只能从含有 m个粒子的离散分布进行初始化

$$\rho_0 = \sum_{l=1}^m \omega_0^{(l)} \delta_{\mu^{(l)}}, \text{ where } \mu^{(1)}, \ldots, \mu^{(m)} \in \mathbb{R}^d \text{ and } \omega_0 = (\omega_0^{(1)}, \ldots, \omega_0^{(m)}) \in \Delta^{m-1}.$$

下面的定理刻画了从具有 m 个质量点的离散分布ρ_0 初始化时的 Fisher-Rao梯度流。

定理 5（粒子Fisher-Rao梯度流）。常微分方程组

$$\dot{\omega}_t^{(j)} = -\omega_t^{(j)} \left[1 - \frac{1}{N} \sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^m \omega_t^{(l)} \phi\left(X_i - \mu^{(l)}\right)} \right], \qquad 1 \leq j \leq m \tag{3.15}$$

以 $\omega_0 = [\omega_0^{(j)}]_{1 \leq j \leq m} \in \Delta^{m-1}$ 初始化时，在任意时间区间 $[0, T]$上都有唯一解。此外，该流 $(\rho_t)_{t \geq 0}$ 定义为

$$\rho_t := \sum_{l=1}^m \omega_t^{(l)} \delta_{\mu^{(l)}} \tag{3.16}$$

是 *Fisher-Rao*梯度流，即 (3.16) 是偏微分方程 (3.11)的分布解。

证明见附录 E.1。为便于实现，我们还可以对梯度流（3.16）进行时间离散化，从而得到 Fisher-Rao 梯度下降；参见算法 2。

快速查看算法 2 中给出的伪代码，便会发现一个致命缺陷：粒子的位置在初始化时便固定下来，只有粒子的权重会更新。这会引入系统性的近似误差，其规模可能随维度呈指数级增长 d。第 4节的数值实验也印证了这一诊断。

Wasserstein 梯度下降。 近年来，Wasserstein梯度流无论在理论层面（Ambrosio 等人，2008；Santambrogio，2017）还是作为一种实用工具，在从采样（Chewi 等人，2020年）到变分推断（Lambert 等人，2022年）以及神经网络理论（Chizat and Bach，2018年；Sander et al.，2022）在内的多种应用中，都受到了广泛关注。Wasserstein梯度流在实践中取得的成功，很大程度上得益于它们可以通过相互作用粒子系统来实现。

事实上，与此前关于浅层神经网络的研究（Chizatand Bach，2018年；Mei 等人，2018年）类似，负对数似然的Wasserstein梯度流准确描述了梯度下降的动态过程

算法2 Fisher-Rao 梯度下降。

输入：数据 $\{X_i\}_{1\le i\le n}$、粒子数量 m、步长 $\eta>0$、最大迭代次数 t_0。**初始化**：从均匀分布($\{X_i\}_{1\le i\le n}$)中独立同分布地抽取 $\mu^{(1)}$、……、$\mu^{(m)}\sim$，并抽取 $\omega_0^{(1)}=\cdots=\omega_0^{(m)}=1/m$。**更新**：**对于** $t=0$，1、……、t_0 **执行**

$$\omega_{t+1}^{(j)}=\omega_t^{(j)}+\eta\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(X_i-\mu^{(l)}\right)}-1\right]\omega_t^{(j)},$$

对所有 $j=1$、……、m。这里，$\phi(x)=(2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$ 是 $\mathcal{N}(0$、$I_d)$ 的概率密度函数。**输出** $\rho=\sum_{j=1}^{m}\omega_{t_0}^{(j)}\delta_{\mu^{(j)}}$ 作为（近似）非参数最大似然估计。

关于拟合混合分布的位置参数。感兴趣的读者可参阅附录8 中的定理F ，其中给出了严格表述。

在附录C.3 中，我们推导了 $\ell_N(\rho)$ 在 Wasserstein 几何下的梯度流，其演化满足偏微分方程

$$\partial_t\rho_t=\operatorname{div}\left(\rho_t\nabla\delta\ell_N\left(\rho_t\right)\right) \tag{3.18}$$

相应的离散时间算法为

$$\rho_{t+1}=\left[\mathrm{id}-\eta\nabla\delta\ell_N\left(\rho_t\right)\right]_{\#}\rho_t,\qquad t=0,1,\ldots$$

其中步长为 $\eta>0$。当从离散分布初始化时

$$\rho_0=\frac{1}{m}\sum_{l=1}^{m}\delta_{\mu_0^{(l)}},\ \text{where}\ \mu_0^{(1)},\ldots,\mu_0^{(m)}\in\mathbb{R}^d,$$

下述定理对 Wasserstein梯度流给出了简洁的刻画。证明见附录 E.2。

定理 6 (粒子 Wasserstein 梯度流)。 常微分方程组

$$\dot{\mu}_t^{(j)}=\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(X_i-\mu_t^{(l)}\right)}\left(X_i-\mu_t^{(j)}\right), \tag{3.19}$$

以 $\mu_0^{(1)},\ldots\mu_0^{(m)}$ 独立同分布$\sim$均匀分布$(\{X_i\}_{1\le i\le N})$ 为初始化，在任意时间区间 $[0,T]$ 上均有唯一解。此外，将流 $(\rho_t)^{t\ge0}$ 定义为

$$\rho_t:=\frac{1}{m}\sum_{l=1}^{m}\delta_{\mu_t^{(l)}} \tag{3.20}$$

即为 *Wasserstein*梯度流，也就是说，(3.20) 是偏微分方程 (3.18) 的一个分布解。

通过时间离散 化，我们还可得到 Wasserstein 梯度下降算法；参见算法3。

与 Fisher-Rao 梯度下降算法不同，Wasserstein 流允许粒子在空间中演化，因此可以减轻近似误差。然而，要将质量从初始化时空间中的某个区域移动到相距较远的另一个区域，这种移动可能需要很长时间。相比之下，Wasserstein-Fisher-Rao 梯度下降不仅允许粒子在空间中演化，还允许权重发生变化，因此与朴素的 Wasserstein 梯度下降相比，性能大幅提升。数值实验再次证明了这一优势。特别是，得益于定理 8，这些实验表明，WFR 梯度下降在混合分布的位置参数上优于经典梯度下降。

读者会注意到，这里没有类似于定理7 针对 WFR 以及定理 4针对 Fisher-Rao 梯度下降的收敛结果。遗憾的是，我们未能推导出此类收敛结果。Wasserstein梯度流的动力学相当复杂，附录 F 专门用于建立部分结果，以帮助理解这些动力学。

算法 3 Wasserstein 梯度下降。

输入：数据 $\{X_i\}_{1\le i\le n}$，粒子数 m，步长 $\eta>0$，最大迭代次数 t_0。**初始化**：抽取 $\mu_0^{(1)},\dots,\mu_0^{(m)}$ 独立同分布的$\sim$ 均匀分布($\{X_i\}_{1\le i\le n}$)。**更新**：**对于** $t=0,\ 1,\dots,t_0$ **执行**

$$\mu_{t+1}^{(j)}=\mu_t^{(j)}+\eta\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\left(X_i-\mu_t^{(j)}\right),$$

对所有 $j=1,\dots,m$ 均成立。这里，$\phi(x)=(2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$ 是 $\mathcal{N}(0,I_d)$ 的概率密度函数。**输出** $\rho=m^{-1}\sum_{j=1}^{m}\delta_{\mu_{t_0}^{(j)}}$ 作为（近似）非参数最大似然估计。

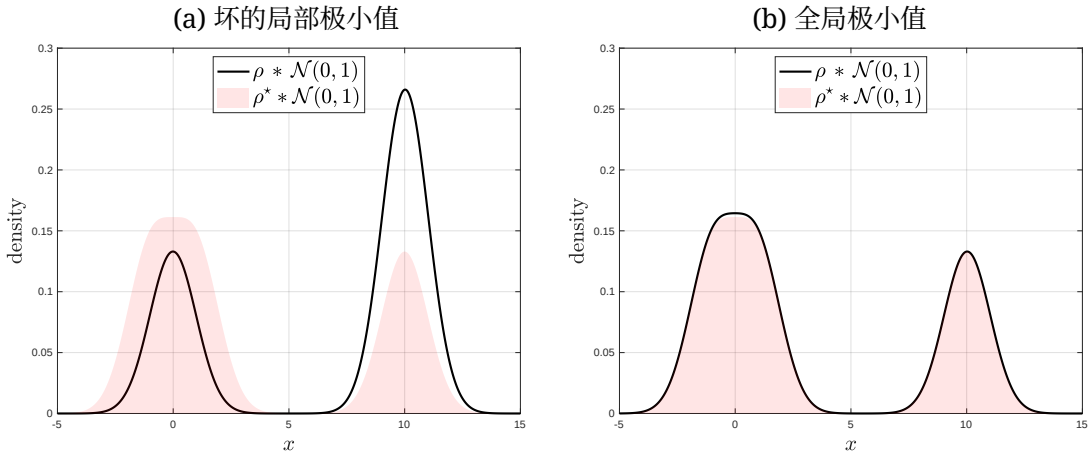


图1：$\rho*\mathcal{N}(0,1)$（学习得到的高斯混合分布）和 $\rho^\star*\mathcal{N}(0,1)$（真实的高斯混合分布）的密度图，其中 $\rho=\frac{1}{3}\delta_{\mu_1}+\frac{1}{3}\delta_{\mu_2}+\frac{1}{3}\delta_{\mu_3}$，$\rho^\star=\rho_{\mathsf{d}}$ 和 (μ_1,μ_2,μ_3) 是期望最大化和梯度下降算法返回的输出。图1(a)展示了当 (μ_1,μ_2,μ_3) 是 ℓ 的一个坏的局部极小值时学习得到的高斯混合分布；图1(b)则对应于 (μ_1,μ_2,μ_3) 是 ℓ 的一个全局极小值的情况。

4 数值实验

本节中，我们进行了一系列数值实验，以验证并补充我们的理论。我们考虑 $\mathbb{R}^d$ 中的两种混合分布：第一种是连续的各向同性高斯分布

$$\rho_{\mathsf{c}}=\mathcal{N}(0,I_d),$$

另一种是受 Jin 等人 （2016） 的启发而构造的离散分布

$$\rho_{\mathsf{d}}=\left(\frac{1}{3}\delta_{-1}+\frac{1}{3}\delta_1+\frac{1}{3}\delta_{10}\right)\otimes(\delta_0)^{\otimes(d-1)}.$$

第二种混合分布 ρ_{d} 是一个乘积分布，其第一个边缘分布是在 $\{-1,1,10\}$ 上的均匀分布，其余 $d-1$ 边缘分布则是取常数零的退化分布。根据 Jin 等人 （2016），即使拥有无限样本和已知权重，经典期望最大化算法和梯度下降算法也无法学习该高斯混合分布中各分量的位置。

4.1 经典算法的不稳定性

本节比较学习高斯混合分布的两种经典算法，其中混合分布为 $\rho^\star=\rho_{\mathsf{d}}$：(i) 期望最大化（EM）算法；(ii) 梯度下降（GD）

算法，以及本文提出的 Wasserstein-Fisher-Rao 梯度下降算法（参见算法 1）。对于前两种算法，我们假设混合成分的数量 $k = 3$ 和权重$\omega_1^\star = \omega_2^\star = \omega_3^\star = 1/3$ 事先已知，并采用标准的 EM 和梯度下降算法来求解最大似然估计（MLE）

$$\min_{\mu_1,\mu_2,\mu_3} \ell(\mu_1,\mu_2,\mu_3) := -\frac{1}{N}\sum_{i=1}^{N}\log\left\{\frac{1}{3}\sum_{j=1}^{3}\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}\|X_i-\mu_j\|_2^2\right]\right\}.$$

期望最大化算法的更新规则为

$$\mu_j^{t+1} = \frac{\sum_{i=1}^N \omega_{i,j}^t X_i}{\sum_{i=1}^N \omega_{i,j}^t} \qquad \text{where} \qquad \omega_{i,j}^t = \frac{\omega_j^\star \phi\left(X_i;\mu_j^t,I_d\right)}{\sum_{l=1}^3 \omega_l^\star \phi\left(X_i;\mu_l^t,I_d\right)} \quad \forall\, i\in[N]\,. \tag{4.1}$$

对所有 $j = 1$，2，3 以及 $t \geq 0$，使用样本 μ_1^0、 μ_2^0、 μ_3^0 独立同分布地$\sim$ 均匀分布$(\{X_i\}_{1\leq i\leq N})$进行随机初始化。梯度下降算法与 Wasserstein 梯度下降（参见算法 3）一致，其中有 $m = 3$ 个粒子。

为便于可视化，我们考虑一维情形（即 $d = 1$），事实证明，这足以展示期望最大化和梯度下降的不稳定性。我们生成 $N = 1500$ 个样本 $\{X_i\}_{1\leq i\leq N}$，取自$\rho^\star * \mathcal{N}(0, 1)$。随后固定这些样本，并对这三种算法分别进行 100 次独立试验。对于期望最大化，我们运行 200 次迭代。对于梯度下降和 Wasserstein-Fisher-Rao 梯度下降，我们将所有步长设为 $\eta = 0.1$，并运行 $t_0 = 1000$ 次迭代。期望最大化和梯度下降各有两种可能的输出：（i）第一种为$\mu_1 \approx \mu_2 \approx 10$ 和 $\mu_3 \approx 0$（忽略排列顺序），这是 ℓ 的一个坏的局部极小值；（ii）第二种为$\mu_1 \approx -1$、$\mu_2 \approx 1$ 和 $\mu_3 \approx 10$，这是 ℓ 的全局极小值。图1展示了与这两种输出相对应的学习得到的高斯混合分布；显然，当两种算法收敛到坏的局部极小值时，都无法学到真实的高斯混合分布。在 100 次独立试验中，期望最大化有 23 次收敛到坏的局部极小值，而梯度下降有 32 次，两者都表现出对随机初始化的不稳定性。相比之下，正如我们将在第4.3节中所展示的，当粒子数为 $m = 500$ 时，Wasserstein-Fisher-Rao 梯度下降能够稳定地收敛到非参数最大似然估计，并学到图4(c)所示的高斯混合分布。

4.2 Wasserstein-Fisher-Rao 梯度下降的优越性

本节比较本文所研究的三种梯度下降算法的实证表现：(i) Fisher-Rao梯度下降（参见算法2）；(ii) Wasserstein-Fisher-Rao 梯度下降（参见算法 1）；以及 (iii) Wasserstein 梯度下降（参见算法 3）。我们将样本量设为 $N = 1500$，维度设为 $d = 10$，最大迭代次数设为 $t_0 = 1000$，算法2和算法 3 的步长设为 $\eta = 10^{-1}$，算法 1 的步长设为 $\eta = 10^{-2}$。图2(a)展示了三种算法在20次独立试验中，负对数似然 $\ell_N(\rho)$（带有一条标准差误差线）随粒子数 m 的变化情况。与前一个实验不同，我们在每次独立试验中都生成新的样本 $\{X_i\}_{1\leq i\leq N}$。可以看出，随着粒子数增加，损失逐渐降低；对于所有 m，Wasserstein-Fisher-Rao 梯度下降均能达到最小损失。还可以观察到，当 $m > 500$ 时，对于 Wasserstein-Fisher-Rao 梯度下降，增加粒子数所带来的边际收益变得可以忽略不计。类似地，图2(b)展示了三种算法在20次独立试验中，负对数似然 $\ell_N(\rho_t)$（带有一条标准差误差线）随迭代次数 t 的变化情况。可以看到，在所有迭代次数下，Wasserstein-Fisher-Rao 梯度下降再次实现了最小损失，进一步确认了该算法的优越性。

4.3 在一维情形下验证非参数最大似然估计的最优性条件

本文中的收敛保证仅涵盖 Wasserstein-Fisher-Rao 梯度下降的无限粒子极限（参见算法 1）。本节提供数值证据，表明当使用大量粒子时，算法1 会收敛到（近似的）非参数最大似然估计。

回顾定理 1 及其后的注释可知，$\rho \in \mathcal{P}(\mathbb{R}^d)$ 是非参数最大似然估计，当且仅当 $\delta\ell_N(\rho)(x) \geq -1$对所有 $x \in \mathbb{R}^d$均成立。我们关注一维情形（即 $d = 1$），因为在一维中检查函数 $\delta\ell_N(\rho)(x)$ 的取值在计算上可行（例如，在细网格上）。对于离散分布

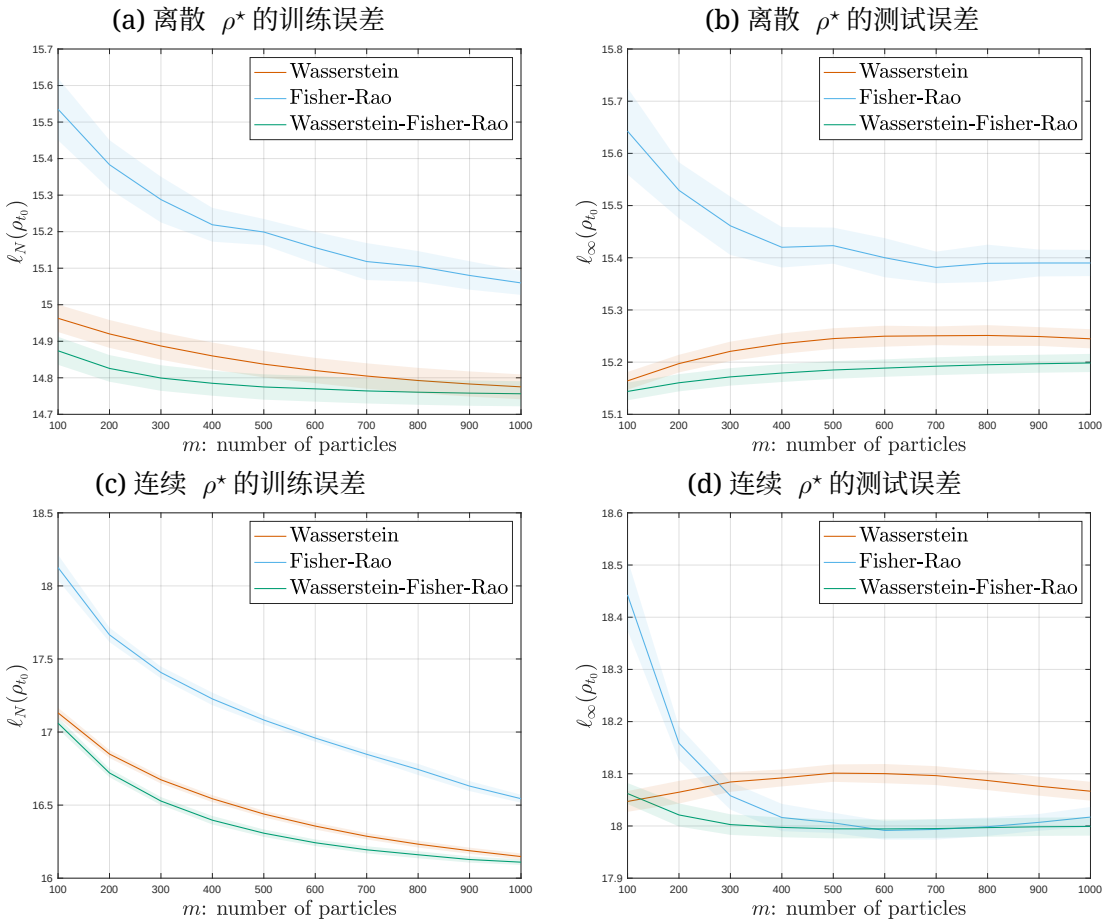


图 2：三种算法的训练或测试误差（含误差条）随粒子数的变化。训练误差和测试误差分别使用 $\ell_N(\rho)$ 和 $\ell_\infty(\rho)$ 进行评估。图 (a) 和 (b) 展示了混合分布 $\rho^\star = \rho_{\mathsf{d}}$ 为离散分布时的训练误差和测试误差，而图 (c) 和 (d) 展示了混合分布 $\rho^\star = \rho_{\mathsf{c}}$ 为连续分布时的训练误差和测试误差。结果基于 $N = 1500$、 $d = 10$ 和 $t_0 = 1000$ 的 20 次独立试验报告。

$\rho = \sum_{j=1}^m \omega_j \delta_{\mu_j}$，我们定义以下次优性间隙：

$$\mathsf{gap}(\rho) := \sup_{x\in\mathbb{R}} \max\left\{-1 - \delta\ell_N(\rho)(x), 0\right\}, \qquad \widehat{\mathsf{gap}}(\rho) := \max_{x\in\mathsf{grid}(\rho)} \max\left\{-1 - \delta\ell_N(\rho)(x), 0\right\}.$$

显然，当第一个最优性差距 $\mathsf{gap}(\rho) = 0$ 时，定理 1 表明 ρ 是非参数最大似然估计。然而，$\mathsf{gap}(\rho)$ 通常难以计算，一种实用方案是用细网格 $\mathsf{grid}(\rho) \subseteq \mathbb{R}$ 上的最大值近似计算 $\mathbb{R}$ 上的上确界，由此得到第二个最优性差距 $\widehat{\mathsf{gap}}$。在我们的实验中，我们将 $\mathsf{grid}(\rho)$ 设为覆盖 $[\min_{1\le j\le m}\mu_j - 1, \max_{1\le j\le m}\mu_j + 1]$ 的 0.01-网。

我们设定样本量 $N = 1500$、维度 $d = 1$，并运行 Wasserstein-Fisher-Rao 梯度下降（参见算法 1），粒子数为 $m = 500$，步长为 $\eta_1 = \eta_2 = 10^{-1}$。图 4展示了 $\widehat{\mathsf{gap}}$ 次优性间隙 在（4.3）中的变化情况（带有一个标准差的误差线），以及算法输出与 $\mathcal{N}(0, 1)$ 卷积后得到的密度图；这些结果是基于 20次独立试验得出的。粗略地说，两个次优性间隙都随迭代次数增加而近似成反比地减小。最后，图 5描绘了一阶变分 $\delta\ell_N(\rho)$，清楚地表明最优性条件以很高的精度近似得到满足；更多详情请参见图 5 的图注。

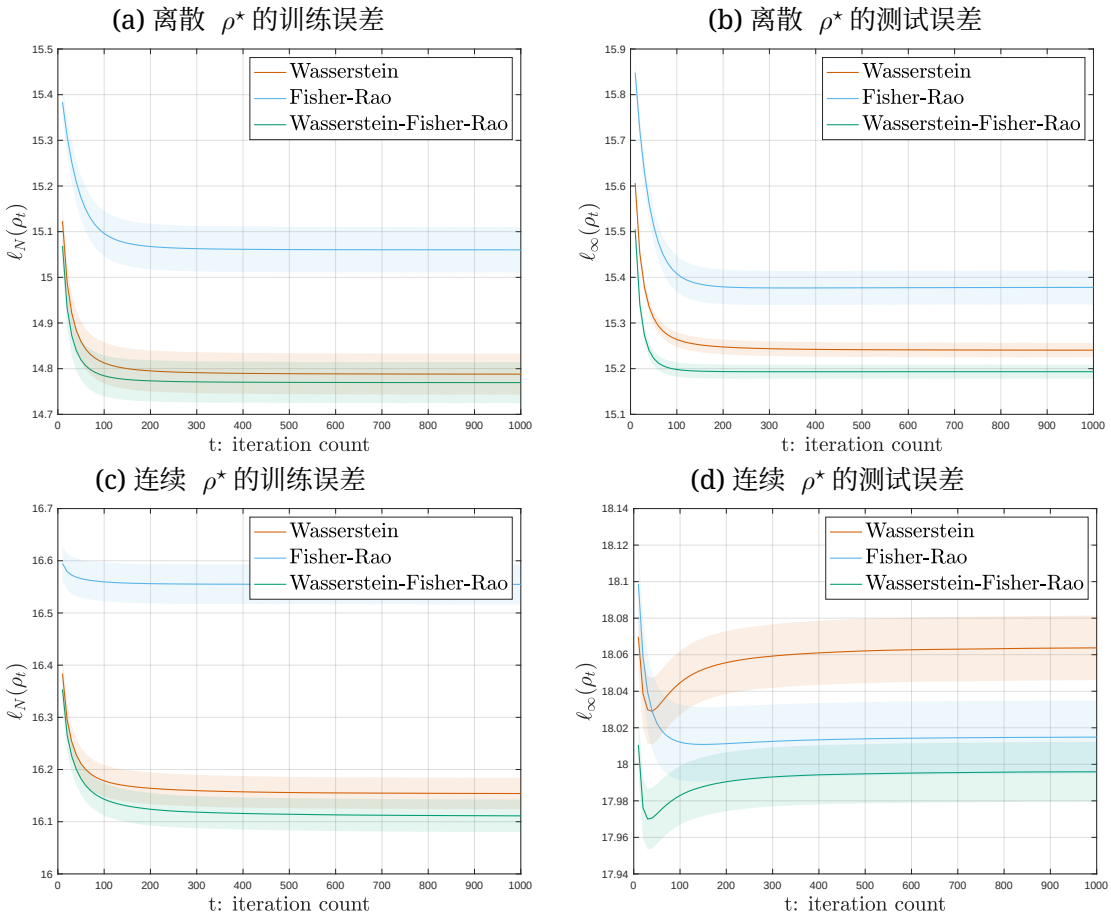


图 3：三种算法的训练或测试误差（含误差条）随迭代次数的变化。训练误差和测试误差分别使用 $\ell_N(\rho_t)$ 和 $\ell_\infty(\rho_t)$ 进行评估。图 (a) 和 (b) 展示了混合分布 $\rho^\star = \rho_{\mathsf{d}}$ 为离散时的训练误差和测试误差，而图 (c) 和 (d) 展示了混合分布 $\rho^\star = \rho_{\mathsf{c}}$ 为连续时的训练误差和测试误差。结果基于 $N = 1500$、$d = 10$ 和 $m = 500$ 的 20 次独立试验报告。

5 讨论

本文提出使用由 Wasserstein-Fisher-Rao 梯度下降驱动的相互作用粒子系统，求解高斯混合分布的非参数最大似然估计。在无限粒子极限下，我们证明，在一定条件下，所提出的算法收敛到非参数最大似然估计。在实践中，我们进行了大量数值实验，展示所提出的算法仅使用有限（甚至少量）粒子即可精确计算非参数最大似然估计的能力，并证明其相较于其他算法的优越性。展望未来，还有许多值得进一步研究的扩展方向。例如，我们的收敛理论仅适用于从支撑于整个空间的分布初始化 Wasserstein-Fisher-Rao 梯度下降；将当前分析扩展到更实用的情形将会很有意义，即从具有有限数量粒子的离散分布初始化算法。另一个有趣的方向是，研究如何在各向同性情形之外学习高斯混合分布（即不假设协方差矩阵为单位矩阵），例如利用 Bures-Wasserstein 空间上的 Wasserstein-Fisher-Rao 梯度流（Lambert 等人，2022年）。

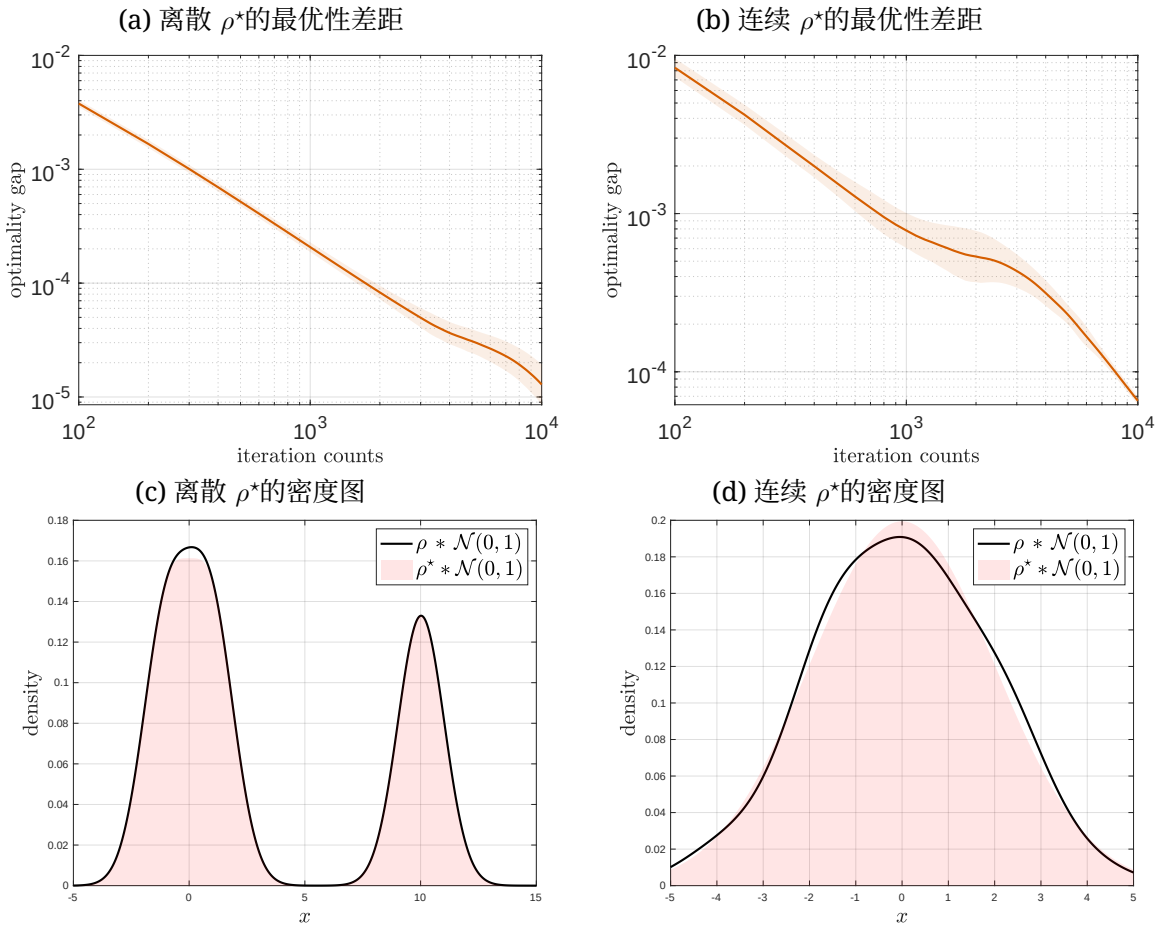
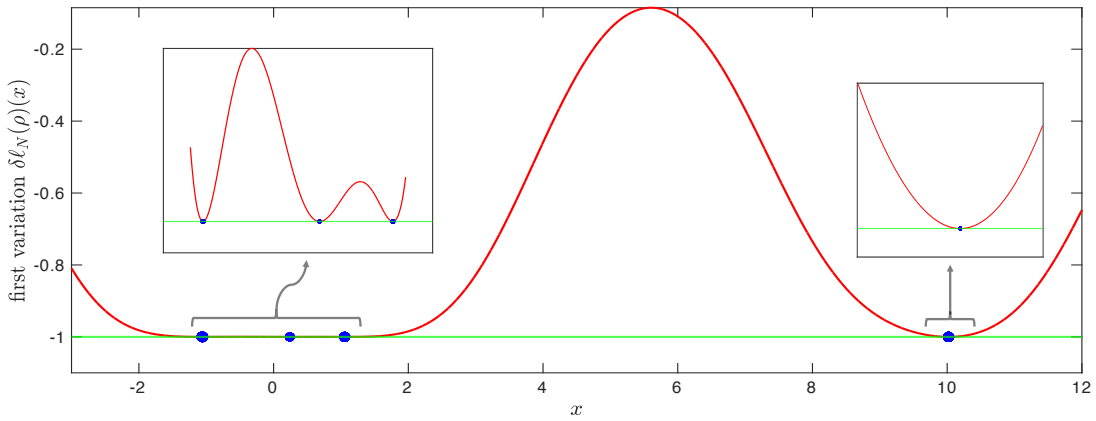


图 4：图 (a)–(b) 展示了使用 Wasserstein-Fisher-Rao 梯度下降时，子最优性差距（参见 (4.3)）随迭代次数的变化，其中采用了离散混合分布 $\rho^\star = \rho_{\mathsf{d}}$（图 (a)）和 $\rho^\star = \rho_{\mathsf{c}}$（图 (b)）。误差条基于20 次独立试验计算得到。图 (c)–(d) 是 ρ 和 $\rho^\star$ 与标准高斯分布进行卷积后的密度图，其中 (c) 采用离散混合分布 $\rho^\star = \rho_{\mathsf{d}}$，(d) 采用连续混合分布$\rho^\star = \rho_{\mathsf{c}}$。结果报告为 $N = 1500$, $d = 1$ 和 $m = 500$。

致谢

作者感谢 Donghao Wang 提供了有益的讨论。Y. Yan 部分获得了普林斯顿大学 Charlotte Elizabeth Procter荣誉奖学金的支持。Y. Yan 于 2022年秋季访问麻省理工学院期间完成了本研究的部分工作。K. Wang 获得了美国国家科学基金会 DMS-2210907 项目资助以及哥伦比亚大学的启动经费。P. Rigollet 获得了美国国家科学基金会 IIS-1838071、DMS-2022448 和 CCF-2106377 项目资助。

(a) 离散 $\rho^\star = \frac{1}{3}\delta_{-1} + \frac{1}{3}\delta_1 + \frac{1}{3}\delta_{10}$的一阶变分 $\delta\ell_N(\rho)$



(b) 连续 $\rho^\star = \mathcal{N}$的一阶变分 $\delta\ell_N(\rho)(0,\ 1)$

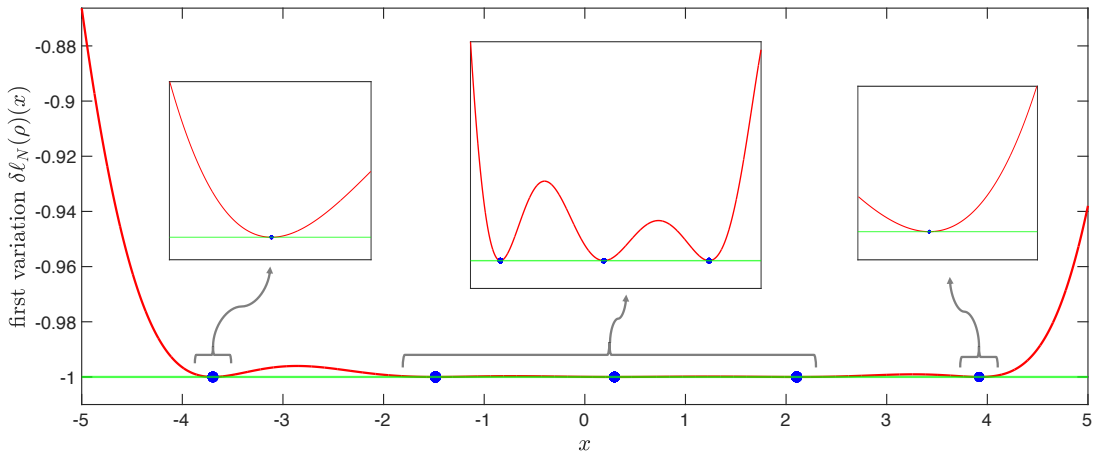


图5：上图中离散 $\rho^\star = \rho_{\mathrm{d}}$ 的一阶变分 $\delta\ell_N(\rho)$（红线），下图中连续$\rho^\star = \rho_{\mathrm{c}}$的一阶变分。蓝点为 $\{(\mu_j, \delta\ell_N(\rho)(\mu_j)):\ 1 \le j \le m\}$，其中点的大小与权重 $\{\omega_j\}_{1\le j\le m}$成正比。绿线为水平线 $y = -1$。上图中的两个子图分别放大了区域 $-1.5 \le x \le 1.5$ 和 $9.8 \le x \le 10.2$，下图中的三个子图分别放大了区域 $-4 \le x \le -3.4$、$-2 \le x \le 3$ 和 $3.8 \le x \le 4.1$。

A 预备知识

在正文中，我们重点研究了高斯混合模型，其中 $\phi(x) = (2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$。事实上，本文中的算法和定理也适用于更一般的一类概率密度ϕ。在附录中，我们只假设 ϕ 满足以下正则性假设。

假设1 (正则性)。 假设密度 $\phi(x) > 0$ 对于任意 $x \in \mathbb{R}^d$成立。此外， $\phi \in C^{\max\{d,2\}}(\mathbb{R}^d)$，$\lim_{\|x\|_2\to\infty}\phi(x) = 0$，$\sup_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2 < \infty$ 以及$\sup_{x\in\mathbb{R}^d}\|\nabla^2\phi(x)\|_2 < \infty$。

显然，高斯核 $\phi(x) = (2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$ 满足假设1。我们还定义以下集合和量，它们将在整个证明过程中发挥作用。

定义1。 令 $\Omega = \mathrm{conv}(\{X_i\}_{1\le i\le N})$。对于 $r \ge 0$，定义 $\Omega_r = \{x \in \mathbb{R}^d : \mathrm{dist}(x, \Omega) \le r\}$，$\bar{\phi}(r) = \sup_{\|x\|_2\ge r}\phi(x)$ 以及 $\underline{\phi}(r) = \inf_{\|x\|_2\le r}\phi(x)$。

下面的引理表明，NPMLE 的支撑集是紧致的。证明见附录B.2。

引理1 (NPMLE的紧支撑)。 设假设 *1* 成立，且 $\widehat{\rho}$ 是 (1.1)的任意最优解。定义 $R_1 =$ 下确界$\{r \geq 0 : \quad \phi(r) \leq \phi[\Omega$ 的直径$]/2\}$ 以及

$$R = \inf\left\{r \geq 0 : \ \bar{\phi}(r) \leq \frac{\bar{\phi}(R_1)\underline{\phi}(R_1 + \operatorname{diam}(\Omega))}{8\bar{\phi}(0)}\right\}.$$

那么，我们有 支撑$(\widehat{\rho}) \subseteq \Omega_R$。

B NPMLE 的结构性结果证明

本节证明 NPMLE 的两个结构性结果，即定理 1 和引理 1 在假设 1下成立。

B.1 定理1的证明

第 1 部分：NPMLE 的存在性。 注意，损失函数 ℓ_N 有下界

$$\ell_N(\rho) = -\frac{1}{N}\sum_{i=1}^{N}\log\left[(\rho * \phi)(X_i)\right] \geq -\log\|\phi\|_\infty, \tag{B.1}$$

其中，最后一个不等式由以下关系得出：对任意 $x \in \mathbb{R}^d$， $\rho * \phi(x) = \int_{\mathbb{R}^d}\phi(y-x)\,\rho(\mathrm{d}y) \leq \|\phi\|_\infty$。因此，存在一个概率分布序列 $\{\rho_n\}$，使得

$$\ell_N(\rho_n) \leq \inf_{\rho\in\mathcal{P}(\mathbb{R}^d)}\ell_N(\rho) + \frac{1}{n}. \tag{B.2}$$

现在我们证明 $\{\rho_n\}$ 具有紧性。为此，我们证明存在 $r > 0$，使得对于任意 $\varepsilon > 0$，当 ρ_n 足够大时，有 $\rho(\Omega \geq 1-\varepsilon)$ n。

对于任意 n 和 $r > 0$，定义

$$\rho_{n,r} \coloneqq \rho_n(\Omega_r)\cdot\rho_n|_{\Omega_r} + \rho_n(\Omega_r^{\mathrm{c}})\cdot\mathsf{Unif}(\Omega),$$

其中， $\rho_n|_{\Omega_r}(\cdot) = \rho_n(\cdot|\Omega_r)$ 是在给定 Ω_r 的条件下 ρ_n 的条件分布，而 $\mathsf{Unif}(\Omega)$ 是 Ω 上的均匀分布。我们有

$$\ell_N(\rho_n) - \ell_N(\rho_{n,r}) = -\frac{1}{N}\sum_{i=1}^{N}\log\left[(\rho_n*\phi)(X_i)\right] + \frac{1}{N}\sum_{i=1}^{N}\log\left[(\rho_{n,r}*\phi)(X_i)\right] = \frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{(\rho_{n,r}*\phi)(X_i)}{(\rho_n*\phi)(X_i)}\right].$$

注意，对于每个 $i \in [N]$

$$\begin{aligned}
\log\left[\frac{(\rho_{n,r}*\phi)(X_i)}{(\rho_n*\phi)(X_i)}\right] &= \log\left[\frac{\int_{\Omega_r}\phi(X_i-y)\,\rho_n(\mathrm{d}y) + \rho_n(\Omega_r^{\mathrm{c}})\int_\Omega\phi(X_i-y)\,\mathrm{d}y}{\int_{\Omega_r}\phi(X_i-y)\,\rho_n(\mathrm{d}y) + \int_{\Omega_r^{\mathrm{c}}}\phi(X_i-y)\,\rho_n(\mathrm{d}y)}\right] \\
&= \log\left[1 + \frac{\rho_n(\Omega_r^{\mathrm{c}})\int_\Omega\phi(X_i-y)\,\mathrm{d}y - \int_{\Omega_r^{\mathrm{c}}}\phi(X_i-y)\,\rho_n(\mathrm{d}y)}{(\rho_n*\phi)(X_i)}\right] \\
&\geq \log\left[1 + \frac{\rho_n(\Omega_r^{\mathrm{c}})\left[\underline{\phi}(\operatorname{diam}(\Omega)) - \bar{\phi}(r)\right]}{\|\phi\|_\infty}\right].
\end{aligned}$$

我们可以选择 $r > 0$ 足够大，使得 $\phi(r) \leq \phi(\Omega$ 的直径$)/2$，因此对于每个 $i \in [N$ $]$

$$\ell_N(\rho_n) - \ell_N(\rho_{n,r}) \geq \log\left[1 + \frac{\rho_n(\Omega_r^{\mathrm{c}})\,\underline{\phi}(\operatorname{diam}(\Omega))}{2\|\phi\|_\infty}\right].$$

鉴于（B.2），我们知道

$$\ell_N(\rho_n) - \ell_N(\rho_{n,r}) \leq \frac{1}{n}.$$

综合以上两个不等式可得

$$\rho_n\left(\Omega_r^{\mathrm{c}}\right) \leq \frac{2\|\phi\|_{\infty}\left[\exp\left(\frac{1}{n}\right)-1\right]}{\underline{\phi}\left(\operatorname{diam}(\Omega)\right)}.$$

因此我们有

$$\rho_n\left(\Omega_r^{\mathrm{c}}\right) \leq \varepsilon \qquad \text{as long as} \qquad n \geq n_\varepsilon := \left\lceil 1/\log\left[1+\frac{\varepsilon\underline{\phi}\left(\operatorname{diam}(\Omega)\right)}{2\|\phi\|_{\infty}}\right]\right\rceil,$$

这意味着 $\{\rho_n\}$ 具有紧性。利用普罗霍罗夫定理，我们得到：存在一个子序列 $\{\rho_{n_k}\}$以及 $\widehat{\rho}\in\mathcal{P}(\mathbb{R}^d)$ 使得 ρ_{n_k} 弱收敛于 $\widehat{\rho}$，且后者必然是（1.1）的极小化子。

第 2 部分：最优性条件。 首先，很容易验证，对于任意 $\rho\in\mathcal{M}(\mathbb{R}^d)$

$$\int_{\mathbb{R}^d}\delta\ell_N(\rho)(x)\,\rho(\mathrm{d}x) = -\frac{1}{N}\sum_{i=1}^N\frac{\int\phi(x-X_i)\,\rho(\mathrm{d}x)}{(\rho*\phi)(X_i)} = -\frac{1}{N}\sum_{i=1}^N\frac{(\rho*\phi)(X_i)}{(\rho*\phi)(X_i)} = -1.$$

若 $\widehat{\rho}\in\mathcal{M}(\mathbb{R}^d)$ 是（1.1）的最优解，则对于任意 $x\in\mathbb{R}^d$ 以及任意 $\varepsilon\in[0,1]$ 我们有

$$\ell_N(\widehat{\rho}) \leq \ell_N\left((1-\varepsilon)\widehat{\rho}+\varepsilon\delta_x\right) = \ell_N\left(\rho+\varepsilon\left(\delta_x-\widehat{\rho}\right)\right).$$

因此我们有

$$\delta\ell_N(\widehat{\rho})(x)+1 = \int_{\mathbb{R}^d}\delta\ell_N(\widehat{\rho})\,\mathrm{d}\left(\delta_x-\widehat{\rho}\right) = \lim_{\varepsilon\to0}\frac{1}{\varepsilon}\left[\ell_N\left(\widehat{\rho}+\varepsilon\left(\delta_x-\widehat{\rho}\right)\right)-\ell_N(\widehat{\rho})\right]\geq 0.$$

由于 x 是任意的，这意味着对于任意 $x\in\mathbb{R}^d$，$\delta\ell_N(\widehat{\rho})(x)\geq -1$。结合 $\int_{\mathbb{R}^d}\delta\ell_N(\widehat{\rho})\mathrm{d}\widehat{\rho}=-1$，即可得到对于 $\widehat{\rho}$-几乎处处的 x，$\delta\ell_N(\widehat{\rho})(x)=-1$。

反过来，如果 $\widehat{\rho}\in\mathcal{M}(\mathbb{R}^d)$ 对所有 $x\in\mathbb{R}^d$ 都满足 $\delta\ell_N(\widehat{\rho})(x)\geq -1$，那么对于任意 $\rho\in\mathcal{M}(\mathbb{R}^d)$，有

$$0 \leq \int_{\mathbb{R}^d}\delta\ell_N(\widehat{\rho})\,\mathrm{d}\rho+1 = \int_{\mathbb{R}^d}\delta\ell_N(\widehat{\rho})\,\mathrm{d}\left(\rho-\widehat{\rho}\right) = \lim_{\varepsilon\to0}\frac{1}{\varepsilon}\left[\ell_N\left(\widehat{\rho}+\varepsilon\left(\rho-\widehat{\rho}\right)\right)-\ell_N(\widehat{\rho})\right] \leq \ell_N(\rho)-\ell_N(\widehat{\rho})$$

其中最后一个不等式由泛函的凸性得出 $\rho\mapsto\ell_N(\rho)$。上述推导表明，$\widehat{\rho}$ 是（1.1）的全局极小化子。

B.2 引理 1的证明

根据定理 1，$\delta\ell_N(\widehat{\rho})=-1$ 在 $\mathsf{supp}(\widehat{\rho})$上。我们将证明，当 $|\delta\ell_N(\widehat{\rho})(y)|<1/2$且 y离Ω足够远时，并由此得出 $\mathsf{supp}(\widehat{\rho})$ 必须保持在Ω附近。为此，我们在下面的引理中给出一些有用的估计，这些估计在后文也会用到。

引理 2. 设假设 *1* 成立。对于任意 $\rho\in\mathcal{P}(\mathbb{R}^d)$，$x\in\Omega$ 和 $r\geq0$，有

$$\rho(\Omega_r)\underline{\phi}(r+\operatorname{diam}(\Omega)) \leq (\rho*\phi)(x) \leq \rho(\Omega_r^{\mathrm{c}})\bar{\phi}(r)+\rho(\Omega_r)\bar{\phi}(0) \leq \bar{\phi}(r)+\rho(\Omega_r)\bar{\phi}(0).$$

因此，

$$-\log\left(\bar{\phi}(r)+\rho(\Omega_r)\bar{\phi}(0)\right) \leq \ell_N(\rho) \leq -\log\left(\rho(\Omega_r)\underline{\phi}(r+\operatorname{diam}(\Omega))\right).$$

任取 $R\geq0$ 使得 $\bar{\phi}(R)\leq e^{-\ell_N(\rho)}/2$。对于任意 $\mu\in\mathcal{P}(\mathbb{R}^d)$ 满足 $\ell_N(\mu)\leq\ell_N(\rho)$，我们有

$$\begin{aligned}
&\mu(\Omega_R) \geq e^{-\ell_N(\rho)}/[2\bar{\phi}(0)],\\
&\inf_{x\in\Omega}(\mu*\phi)(x) \geq e^{-\ell_N(\rho)}\underline{\phi}(R+\operatorname{diam}(\Omega))/[2\bar{\phi}(0)],\\
&\sup_{y\in\Omega_r^c}|\delta\ell_N(\mu)(y)| \leq \frac{2e^{\ell_N(\rho)}\bar{\phi}(0)}{\underline{\phi}(R+\operatorname{diam}(\Omega))}\cdot\bar{\phi}(r), \qquad \forall r\geq0.
\end{aligned}$$

引理 2的证明推迟到本节末尾。现在回过头来证明引理 1。任取一个 $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ 其支撑集为 Ω。根据引理 2，我们有

$$\ell_N(\rho_0) \le -\log\Big(\underline{\phi}[\mathrm{diam}(\Omega)]\Big), \qquad \forall r \ge 0.$$

取 $R_1 = \inf\{r \ge 0 : \ \phi(r) \le \phi[\Omega \text{ 的直径}]/2\}$。由 ϕ 的连续性，

$$\bar{\phi}(R_1) \le \underline{\phi}[\mathrm{diam}(\Omega)]/2 \le e^{-\ell_N(\rho_0)}/2.$$

引理 2 可得

$$\sup_{y\in\Omega_r^{\mathrm{c}}} |\delta\ell_N(\widehat{\rho})(y)| \le \frac{2e^{\ell_N(\rho_0)}\bar{\phi}(0)}{\underline{\phi}(R_1+\mathrm{diam}(\Omega))}\cdot\bar{\phi}(r) \le \frac{4\bar{\phi}(0)}{\bar{\phi}(R_1)\underline{\phi}(R_1+\mathrm{diam}(\Omega))}\cdot\bar{\phi}(r), \qquad \forall r\ge 0.$$

Let

$$R = \inf\left\{r\ge 0 : \ \bar{\phi}(r) \le \frac{\bar{\phi}(R_1)\underline{\phi}(R_1+\mathrm{diam}(\Omega))}{8\bar{\phi}(0)}\right\}.$$

ϕ 的连续性 ϕ推出 $R(\le \phi)$ $R_1(\phi\Omega$ 的直径$R_1+)/[8\phi(0)]$，因此 $\sup_{y\in\Omega_R^{\mathrm{c}}} |\delta\ell_N(\widehat{\rho})(y)| \le 1/2 < 1$。结果是，$\mathrm{supp}(\widehat{\rho}) \cap \Omega_R^{\mathrm{c}} = \varnothing$ 且 $\mathrm{supp}(\widehat{\rho}) \subseteq \Omega_R$。

引理 2的证明。 注意，$\phi(x-y) \le \bar{\phi}(0)$ 对于任意 $x,\ y\in\mathbb{R}^d$。如果 $x\in\Omega$ 和 $y\in\Omega_r^{\mathrm{c}}$，那么 $\|x-y\|_2 \ge r$ 因此 $\phi(x-y) \le \phi(r)$。因此，

$$(\rho*\phi)(x) \le \int_{\Omega_r}\phi(x-y)\rho(\mathrm{d}y) + \int_{\Omega_r^{\mathrm{c}}}\phi(x-y)\rho(\mathrm{d}y) \le \bar{\phi}(0)\rho(\Omega_r) + \bar{\phi}(r)\rho(\Omega_r^{\mathrm{c}}).$$

若 $x\in\Omega$ 且 $y\in\Omega_r$，则 $\|x-y\|_2 \ge r+\Omega$ 的直径，因此 $\phi(x-y) \ge \underline{\phi}(r+\Omega$ 的直径)。因此，

$$(\rho*\phi)(x) \ge \int_{\Omega_r}\phi(x-y)\rho(\mathrm{d}y) \ge \underline{\phi}(r+\mathrm{diam}(\Omega))\rho(\Omega_r).$$

关于 $\ell_N(\cdot)$ 的所需界显而易见。若 $\mu\in\mathcal{P}(\mathbb{R}^d)$ 且 $\ell_N(\mu) \le \ell_N(\rho)$，则我们对 ℓ 的估计意味着

$$-\log\Big(\bar{\phi}(r)+\mu(\Omega_r)\bar{\phi}(0)\Big) \le \ell_N(\mu) \le \ell_N(\rho), \qquad \forall r\ge 0.$$

因此，$\bar{\phi}(r) + \mu(\Omega_r)\,\bar{\phi}(0) \ge e^{-\ell_N(\rho)}$。根据假设 1，$\lim_{r\to\infty}\bar{\phi}(r) = 0$。任取满足 $\bar{\phi}(R) \le e^{-\ell_N(\rho)}/2$ 的 $R\ge 0$。那么，

$$\begin{aligned}
&\mu(\Omega_R) \ge [e^{-\ell_N(\rho)} - \bar{\phi}(r)]/\bar{\phi}(0) \ge e^{-\ell_N(\rho)}/[2\bar{\phi}(0)],\\
&\inf_{x\in\Omega}(\mu*\phi)(x) \ge \mu(\Omega_R)\underline{\phi}(R+\mathrm{diam}(\Omega)) \ge e^{-\ell_N(\rho)}\underline{\phi}(R+\mathrm{diam}(\Omega))/[2\bar{\phi}(0)].
\end{aligned}$$

对于任意 $r\ge 0$，只要 $X\in\mathrm{supp}(\nu)$ 且 $y\in\Omega_r^{\mathrm{c}}$，就有 $\|X-y\|_2 \ge r$。那么，

$$|\delta\ell_N(\mu)(y)| = \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i-y)}{(\mu*\phi)(X_i)} \le \frac{\bar{\phi}(r)}{\inf_{x\in\Omega}(\mu*\phi)(x)}, \qquad \forall r\ge 0,\ \ y\in\Omega_r^{\mathrm{c}}.$$

将此结果与我们上面得到的 $\inf_{x\in\Omega}(\mu*\phi)(x)$ 的下界结合起来，即完成证明。

□

C 关于 $\mathcal{P}_2(\mathbb{R}^d)$ 上梯度流的推导

C.1 一阶变分

回顾一下，总体损失函数和有限样本损失函数分别为

$$\ell_\infty(\rho)=\mathbb{E}_{X\sim(\rho^\star*\phi)}\left\{\log\left[(\rho*\phi)(X)\right]\right\},\qquad \ell_N(\rho)=-\frac{1}{N}\sum_{i=1}^N\log\left[(\rho*\phi)(X_i)\right].$$

ℓ_N 的一阶变分定义为任意可测函数 $\delta\ell(\rho):\mathbb{R}^d\to\mathbb{R}$，满足

$$\lim_{\varepsilon\to0}\frac{\ell(\rho+\varepsilon\mathcal{X})-\ell(\rho)}{\varepsilon}=\int\delta\ell(\rho)\,\mathrm{d}\mathcal{X}$$

对于所有满足 $\int\mathrm{d}\mathcal{X}=0$ 的带符号测度 $\mathcal{X}$。特别地，容易看出，一阶变分只在一个加法常数的意义下确定。

通过直接计算，我们得到

$$\begin{aligned}\lim_{\varepsilon\to0}\frac{\ell_N(\rho+\varepsilon\mathcal{X})-\ell_N(\rho)}{\varepsilon}&=-\frac{1}{N}\sum_{i=1}^N\lim_{\varepsilon\to0}\frac{\log\left[[(\rho+\varepsilon\mathcal{X})*\phi](X_i)\right]-\log\left[(\rho*\phi)(X_i)\right]}{\varepsilon}\\&=-\frac{1}{N}\sum_{i=1}^N\lim_{\varepsilon\to0}\frac{1}{\varepsilon}\log\left[1+\varepsilon\frac{(\mathcal{X}*\phi)(X_i)}{(\rho*\phi)(X_i)}\right]=-\frac{1}{N}\sum_{i=1}^N\frac{(\mathcal{X}*\phi)(X_i)}{(\rho*\phi)(X_i)}\\&=-\frac{1}{N}\sum_{i=1}^N\int\frac{\phi(x-X_i)}{(\rho*\phi)(X_i)}\mathcal{X}(\mathrm{d}x),\end{aligned}$$

因此，我们得到

$$\delta\ell_N(\rho):x\to-\frac{1}{N}\sum_{i=1}^N\frac{\phi(x-X_i)}{(\rho*\phi)(X_i)}.$$

类似地，我们也可以计算

$$\begin{aligned}\lim_{\varepsilon\to0}\frac{\ell_\infty(\rho+\varepsilon\mathcal{X})-\ell_\infty(\rho)}{\varepsilon}&=\lim_{\varepsilon\to0}\frac{1}{\varepsilon}\left[-\int\log\left[\frac{(\rho+\varepsilon\mathcal{X})*\phi(x)}{(\rho*\phi)(x)}\right](\rho^\star*\phi)(x)\,\mathrm{d}x\right]\\&=-\int\frac{(\mathcal{X}*\phi)(x)}{(\rho*\phi)(x)}(\rho^\star*\phi)(x)\,\mathrm{d}x=-\int\frac{(\rho^\star*\phi)(x)}{(\rho*\phi)(x)}\left[\int\phi(x-y)\,\mathcal{X}(\mathrm{d}y)\right]\mathrm{d}x\\&=-\int\left[\int\frac{(\rho^\star*\phi)(x)}{(\rho*\phi)(x)}\phi(x-y)\,\mathrm{d}x\right]\mathcal{X}(\mathrm{d}y).\end{aligned}$$

由此可得

$$\delta\ell_\infty(\rho):x\to-\int\frac{(\rho^\star*\phi)(y)}{(\rho*\phi)(y)}\phi(x-y)\,\mathrm{d}y.$$

C.2 Fisher-Rao梯度流

C.2.1 使用黎曼几何学对梯度流的形式化推导

我们首先在 $\mathcal{P}_2(\mathbb{R}^d)$ 上引入一种黎曼结构，该结构蕴含 Fisher-Rao度量。将 $\rho\in\mathcal{P}_2(\mathbb{R}^d)$ 处的切空间定义为

$$\mathrm{Tan}_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d)\coloneqq\left\{\zeta:\zeta=\rho\left(\alpha-\int\alpha\mathrm{d}\rho\right)\text{ for some }\alpha\text{ satisfying }\int\alpha^2\mathrm{d}\rho<\infty\right\}.$$

我们在切空间 $T_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d)$ 上赋予如下黎曼度量张量 $g_\rho^{\mathsf{FR}}(\cdot,\cdot):\ \mathsf{Tan}_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d)\times\mathsf{Tan}_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$ as

$$g_\rho^{\mathsf{FR}}(\zeta_1,\zeta_2)\coloneqq\int\frac{\zeta_1\cdot\zeta_2}{\rho^2}\,\mathrm{d}\rho$$

$$= \int_{\mathbb{R}^d} \left[\alpha_1(x) - \int_{\mathbb{R}^d} \alpha_1 \mathrm{d}\rho\right] \left[\alpha_2(x) - \int_{\mathbb{R}^d} \alpha_2 \mathrm{d}\rho\right] \rho(\mathrm{d}x)$$
$$= \int_{\mathbb{R}^d} \alpha_1(x)\alpha_2(x)\rho(\mathrm{d}x) - \left(\int_{\mathbb{R}^d} \alpha_1 \mathrm{d}\rho\right)\left(\int_{\mathbb{R}^d} \alpha_2 \mathrm{d}\rho\right),$$

对于任意 $\zeta_1 = \rho(\alpha_1 - \int \alpha_1 \mathrm{d}\rho)$ 和 $\zeta_2 = \rho(\alpha_2 - \int \alpha_2 \mathrm{d}\rho)$。由此黎曼结构诱导的度量，即 Fisher-Rao度量 $d_{\mathsf{FR}}(\cdot,\cdot)$，满足以下性质：

$$d^2_{\mathsf{FR}}(\rho_0,\rho_1) = \inf\left\{\int_0^1 \int \left[\left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t\right)^2\right]\mathrm{d}\rho_t \mathrm{d}t : (\rho_t,\alpha_t)_{t\in[0,1]} \text{ solves } \partial_t\rho_t = \rho_t\alpha_t\right\}.$$

接下来我们遵循 Gallouët and Monsaingeon (2017);Lu 等人 (2019b) 推导关于泛函 ℓ_N 的 Fisher-Rao 梯度流。令 $(\rho_t)^{t\geq 0}$ 为一条 C^1 曲线，满足 $\rho_0 = \rho$ 且初始速度为

$$\partial_t\rho_t|_{t=0} = \zeta = \rho\left(\alpha - \int \alpha \mathrm{d}\rho\right).$$

F isher-Rao 梯度在 ρ 处对 ℓ_N 的定义为函数 $\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho) \in L^2(\rho)$，使得 t

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0} = g^{\mathsf{FR}}_\rho\left(\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho), \zeta\right).$$

为计算它，注意到上述恒等式的右侧为

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0} &= \int \delta\ell_N(\rho)\cdot\partial_t\rho_t\Big|_{t=0} \\ &= \int \delta\ell_N(\rho), \zeta = \int \delta\ell_N(\rho)\left(\alpha - \int \alpha \mathrm{d}\rho\right)\mathrm{d}\rho \\ &= \int \left(\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\mathrm{d}\rho\right)\left(\alpha - \int \alpha \mathrm{d}\rho\right)\mathrm{d}\rho \\ &= g^{\mathsf{FR}}_\rho\left(\rho\left(\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\mathrm{d}\rho\right), \zeta\right).\end{aligned}$$

因此

$$g^{\mathsf{FR}}_\rho\left(\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho)\cdot\zeta\right) = g^{\mathsf{FR}}_\rho\left(\rho\left(\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\mathrm{d}\rho\right), \zeta\right)$$

对任意 $\zeta \in \mathsf{Tan}^{\mathsf{FR}}_\rho\mathcal{P}_2(\mathbb{R}^d)$ 均成立，因此

$$\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho) = \rho\left[\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\mathrm{d}\rho\right] = \rho\left[\delta\ell_N(\rho) + 1\right],$$

其中我们利用了以下事实： $\int \delta\ell_N(\rho)\mathrm{d}\rho = -1$。因此，关于 Fisher-Rao 度量 d_{FR}， ℓ_N 的梯度流为

$$\partial_t\rho_t = -\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho_t) = -\rho_t\left[\delta\ell_N(\rho_t) + 1\right].$$

C.2.2 Fisher-Rao 梯度流的其他视角

本节将正式说明 Fisher-Rao梯度流（3.11）与近端梯度下降和镜像下降之间的联系。为简洁起见，我们重点讨论 ρ_t 为连续的情形； ρ_t 为离散的情形与之类似。我们还将说明粒子Fisher-Rao梯度下降（2）与期望最大化算法之间的联系。

Fisher-Rao梯度流作为近端梯度流。 考虑（3.13）中的近端梯度更新。回顾一下，对于 μ， $\nu \in \mathcal{P}_{\mathsf{ac}}(\mathbb{R}^d)$，Fisher-Rao 距离可表示为

$$d_{\mathsf{FR}}^2(\mu,\nu) = 4\int \left|\sqrt{\mu(x)} - \sqrt{\nu(x)}\right|^2 \mathrm{d}x.$$

请注意，就定义梯度流而言，度量只需考虑其二阶局部展开

$$d_{\mathsf{FR}}^2(\mu,\nu) = \int \frac{\Delta^2(x)}{\nu(x)}\mathrm{d}x + \text{higher-order terms},$$

其中 $\Delta = \mu - \nu$。因此，我们得到一个渐近等价的问题（当 $\eta \to 0$ 时）

$$\rho_t^\eta := \operatorname*{arg\,min}_{\rho\in\mathcal{P}_{\mathsf{ac}}(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \delta\ell_N(\rho_t)\,\mathrm{d}(\rho-\rho_t) + \frac{1}{2\eta}\int \frac{[\rho(x)-\rho_t(x)]^2}{\rho_t(x)}\mathrm{d}x \right\}.$$

一阶最优性条件为

$$\delta\ell_N(\rho_t)(x) + \frac{1}{\eta}\cdot\frac{\rho(x)-\rho_t(x)}{\rho_t(x)} = c$$

对于某个常数 $c \in \mathbb{R}$。由此可得

$$\rho_t^\eta(x) = \rho_t(x)\left[1 + x\eta - \eta\delta\ell_N(\rho_t)(x)\right].$$

由于 ρ_t^η 是概率密度，我们有

$$1 = \int_{\mathbb{R}^d}\rho_t^\eta(x)\,\mathrm{d}x = \int_{\mathbb{R}^d}\rho_t(x)\left[1+c\eta-\eta\delta\ell_N(\rho_t)(x)\right]\mathrm{d}x = 1 + c\eta - \eta,$$

其中我们利用了如下事实：对于任意 $\rho \in \mathcal{P}_2(\mathbb{R}^d)$， $\int_{\mathbb{R}^d}\delta\ell_N(\rho)\,\mathrm{d}\rho = -1$。由此可得 $c = -1$。因此

$$\rho_t^\eta(x) = \rho_t(x)\left[1-\eta-\eta\delta\ell_N(\rho_t)(x)\right],$$

因此，

$$\partial_t\rho_t = \lim_{\eta\to 0+}\frac{\rho_t^\eta - \rho_t}{\eta} = -\left[1+\delta\ell_N(\rho_t)\right],$$

这就恢复了 Fisher-Rao 梯度流（3.11）。

Fisher-Rao 梯度流作为镜像流。请回想，镜像下降更新定义为 s

$$\rho_t^\eta := \operatorname*{arg\,min}_{\rho\in\mathcal{P}_{\mathsf{ac}}(\mathbb{R}^d)} \int_{\mathbb{R}^d}\delta\ell_N(\rho_t)\,\mathrm{d}(\rho-\rho_t) + \frac{1}{\eta}\mathsf{KL}(\rho\,\|\,\rho_t).$$

$f(\cdot) := \mathsf{KL}(\cdot\,\|\,\rho_t)$ 的一阶变分为

$$\delta f(\rho)(x) = \log\left[\frac{\rho(x)}{\rho_t(x)}\right],$$

因此，一阶最优性条件为

$$\delta\ell_N(\rho_t)(x) + \frac{1}{\eta}\log\left[\frac{\rho(x)}{\rho_t(x)}\right] = c$$

对于某个常数 $c > 0$。这给出

$$\frac{\rho(x)}{\rho_t(x)} = \exp\left\{\eta\left[c - \delta\ell_N(\rho_t)(x)\right]\right\} \propto \exp\left[-\eta\delta\ell_N(\rho_t)(x)\right].$$

由于 $\int_{\mathbb{R}^d} \rho_t^\eta(x)\mathrm{d}x = 1$，我们知道闭式解为

$$\rho_t^\eta(x) = \frac{\rho_t(x)\exp\left[-\eta\delta\ell_N(\rho_t)(x)\right]}{\int \rho_t(y)\exp\left[-\eta\delta\ell_N(\rho_t)(y)\right]\mathrm{d}y}. \tag{C.1}$$

于是，当 $\eta \to 0$ 时，我们可以计算

$$\begin{aligned}\rho_t^\eta(x) &= \frac{\rho_t(x)\left[1-\eta\delta\ell_N(\rho_t)(x)+O\left(\eta^2\right)\right]}{\int \rho_t(y)\left[1-\eta\delta\ell_N(\rho_t)(y)+O\left(\eta^2\right)\right]\mathrm{d}y} = \frac{\rho_t(x)\left[1-\eta\delta\ell_N(\rho_t)(x)+O\left(\eta^2\right)\right]}{1+\eta+O\left(\eta^2\right)} \\ &= \rho_t(x)\left\{1-\eta\left[1+\delta\ell_N(\rho_t)(x)\right]+O\left(\eta^2\right)\right\},\end{aligned}$$

其中，我们利用了以下事实：对于任意 $\rho \in \mathcal{P}_2(\mathbb{R}^d)$，$\int \delta\ell_N(\rho)\mathrm{d}\rho = -1$。因此，镜像下降的连续时间极限为

$$\partial_t\rho_t = \lim_{\eta\to 0+}\frac{\rho_t^\eta(x)-\rho_t(x)}{\eta} = -\left[1+\delta\ell_N(\rho_t)(x)\right],$$

这就得到了 Fisher-Rao梯度流（3.11）。

将 Fisher-Rao 梯度下降视为期望最大化算法。现在，我们考虑拟合一个具有未知权重 $\{\omega^{(j)}\}_{1\le j\le m}$、已知位置参数 $\{\mu_j\}_{1\le j\le m}$ 和各向同性协方差的 m 分量高斯混合模型。给定数据 $\{X_i\}_{1\le i\le N}$，最大似然估计为

$$\arg\max_{\omega\in\Delta^{m-1}}\ell(\omega) = \frac{1}{N}\sum_{i=1}^N \log\left[\sum_{j=1}^m \omega^{(j)}\phi(X_i-\mu_j)\right].$$

用于求解上述最大似然估计的期望最大化算法如下。首先，我们引入服从 $\mathbb{P}(J_i=j)=\omega^{(j)}$（$1\le j\le m$）分布的潜在独立同分布随机变量 $\{J_i\}_{1\le i\le N}$，那么观测样本的分布为 $X_i|J_i=j\sim\mathcal{N}(\mu_j,\ I_d)$。$(X_i,\ J_i)$ 的联合分布为

$$p_\omega(x,j) = \phi(X_i-\mu_j)\,\omega^{(j)},$$

并且在给定 $X_i=x$ 的条件下，J_i 的条件分布为

$$p_\omega(j|x) = \frac{p_\omega(x,j)}{\sum_{l=1}^m p_\omega(x,l)} = \frac{\phi(X_i-\mu_j)\,\omega^{(j)}}{\sum_{l=1}^m \phi(X_i-\mu_l)\,\omega^{(l)}}.$$

给定当前估计 ω_t 后，E 步相当于计算

$$\begin{aligned}Q(\omega|\omega_t) &= \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^m p_{\omega_t}(j|X_i)\log p_\omega(X_i,j) \\ &= \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^m \frac{\phi(X_i-\mu_j)\,\omega_t^{(j)}}{\sum_{l=1}^m\phi(X_i-\mu_l)\,\omega_t^{(l)}}\log\left[\phi(X_i-\mu_j)\,\omega^{(j)}\right].\end{aligned}$$

M 步是更新

$$\omega_{t+1} := \arg\max_{\omega\in\Delta^{m-1}} Q(\omega|\omega_t),$$

其表达式为

$$\omega_{t+1}^{(j)} = \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i-\mu_j)\,\omega_t^{(j)}}{\sum_{l=1}^m\phi(X_i-\mu_l)\,\omega_t^{(l)}} \qquad \forall\, 1\le j\le m.$$

这等价于算法2，步长为 $\eta=1$。

C.3 Wasserstein梯度流

我们在 $\mathcal{P}_2(\mathbb{R}^d)$ 上引入了二次Wasserstein距离所对应的黎曼结构。我们将 $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ 处的切空间定义为

$$\mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}_2(\mathbb{R}^d) = \left\{\zeta : \zeta = -\mathsf{div}\left(\rho\nabla u\right) \text{ for some } u \text{ satisfying } \int \|\nabla u\|_2^2 \mathrm{d}\rho < \infty\right\}.$$

我们在此切空间上赋予 $L^2(\rho)$ 度量，即定义黎曼度量张量 $g_\rho^{\mathsf{W}}(\cdot,\ \cdot) : \mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}_2(\mathbb{R}^d) \times \mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ 如下：

$$g_\rho^{\mathsf{W}}\left(\zeta_1, \zeta_2\right) \coloneqq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left\langle \nabla u_1, \nabla u_2 \right\rangle \rho\left(\mathrm{d}x\right)$$

对于任意 $\zeta_1 = -\mathsf{div}(\rho\nabla u_1)$ 和 $\zeta_2 = -\mathsf{div}(\rho\nabla u_2)$。由此黎曼结构诱导的度量可还原二次Wasserstein距离，即

$$\begin{aligned} d_{\mathsf{W}}^2\left(\rho_0, \rho_1\right) &= \inf\left\{\int_0^1 \int \|v_t\|_2^2 \,\mathrm{d}\rho_t \mathrm{d}t : \left(\rho_t, v_t\right)_{t\in[0,1]} \text{ solves } \partial_t\rho_t + \mathsf{div}(\rho_t v_t) = 0\right\} \\ &= \inf_{\pi\in\Pi(\rho_0,\rho_1)} \int \|x-y\|_2^2 \,\pi\left(\mathrm{d}x, \mathrm{d}y\right), \end{aligned}$$

其中 $\Pi(\rho_0,\ \rho_1)$ 是 ρ_0 与 ρ_1的耦合集合。这称为 Wasserstein 距离的 Benamou-Brenier公式。接下来，我们推导关于泛函 ℓ_N的 Wasserstein梯度流。感兴趣的读者可参阅Ambrosio 等人(2008，了解 Wasserstein 梯度流的详细介绍。设$(\rho_t)^{t\geq 0}$ 是一条 C^1 满足 $\rho_0 = \rho$ 且初始速度为

$$\partial_t\rho_t|_{t=0} = \zeta = -\mathsf{div}\left(\rho\nabla u\right).$$

于是应有

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_N\left(\rho_t\right)\Big|_{t=0} = g_\rho^{\mathsf{W}}\left(\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho\right), \zeta\right).$$

上述等式左侧等于

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\ell_N\left(\rho_t\right)\Big|_{t=0} &= \int \delta\ell_N\left(\rho\right)\zeta \mathrm{d}x = -\int \delta\ell_N\left(\rho\right)\mathsf{div}\left(\rho\nabla u\right)\mathrm{d}x \\ &= -\int \left\langle \nabla\delta\ell_N\left(\rho\right), \nabla u\right\rangle \mathrm{d}\rho \\ &= g_\rho^{\mathsf{W}}\left(-\mathsf{div}\left(\nabla\delta\ell_N\left(\rho\right)\rho\right), \zeta\right). \end{aligned}$$

因此

$$g_\rho^{\mathsf{W}}\left(\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho\right), \zeta\right) = g_\rho^{\mathsf{W}}\left(-\mathsf{div}\left(\nabla\delta\ell_N\left(\rho\right)\rho\right), \zeta\right)$$

对任意 $\zeta \in \mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}_2(\mathbb{R}^d)$，均成立，因此

$$\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho\right) = -\mathsf{div}\left(\nabla\delta\ell_N\left(\rho\right)\rho\right).$$

这表明，关于二次Wasserstein距离 d_{W}， ℓ_N 的梯度流为

$$\partial_t\rho_t = -\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho_t\right) = \mathsf{div}\left(\nabla\delta\ell_N\left(\rho_t\right)\rho_t\right).$$

C.4 Wasserstein-Fisher-Rao 梯度流

我们引入 Wasserstein-Fisher-Rao度量所依托的 $\mathcal{P}_2(\mathbb{R}^d)$上的黎曼结构。将 $\rho \in \mathcal{P}_2(\mathbb{R}^d)$处的切空间定义为

$$\mathrm{Tan}_\rho^{\mathsf{WFR}}\mathcal{P}_2(\mathbb{R}^d) = \left\{\zeta : \zeta = -\mathsf{div}\left(\rho\nabla u\right) + \rho\left(\alpha - \int \alpha \mathrm{d}\rho\right) \text{ for some } u, \alpha : \mathbb{R}^d \to \mathbb{R}\right.$$

$$\text{satisfying } \int(\alpha^2+\|\nabla u\|_2^2)\mathrm{d}\rho<\infty\Big\}.$$

我们在该切空间上赋予黎曼度量张量 $g_\rho^{\mathsf{WFR}}(\cdot,\ \cdot):\mathsf{Tan}_\rho^{\mathsf{WFR}}\mathcal{P}_2(\mathbb{R}^d)\times\mathsf{Tan}_\rho^{\mathsf{WFR}}\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$，定义如下

$$\begin{aligned}g_\rho^{\mathsf{WFR}}(\zeta_1,\zeta_2)&:=\int_{\mathbb{R}^d}\langle\nabla u_1,\nabla u_2\rangle\rho(\mathrm{d}x)+\int_{\mathbb{R}^d}\left[\alpha_1(x)-\int_{\mathbb{R}^d}\alpha_1\mathrm{d}\rho\right]\left[\alpha_2(x)-\int_{\mathbb{R}^d}\alpha_2\mathrm{d}\rho\right]\rho(\mathrm{d}x)\\&=\int_{\mathbb{R}^d}\langle\nabla u_1,\nabla u_2\rangle\rho(\mathrm{d}x)+\int_{\mathbb{R}^d}\alpha_1(x)\alpha_2(x)\rho(\mathrm{d}x)-\left(\int_{\mathbb{R}^d}\alpha_1\mathrm{d}\rho\right)\left(\int_{\mathbb{R}^d}\alpha_2\mathrm{d}\rho\right)\end{aligned}$$

对于任意 $\zeta_1=-\mathsf{div}(\rho\nabla u_1)+\rho(\alpha_1-\int\alpha_1\mathrm{d}\rho)$ 和 $\zeta_2=-\mathsf{div}(\rho\nabla u_2)+\rho(\alpha_2-\int\alpha_2\mathrm{d}\rho)$。上述黎曼结构诱导的度量，即 Wasserstein-Fisher-Rao度量 $d_{\mathsf{WFR}}(\cdot,\ \cdot)$，定义为

$$\begin{aligned}d_{\mathsf{WFR}}^2(\rho_0,\rho_1)=\inf\Bigg\{\int_0^1\int\left[\|v_t\|^2+\left(\alpha_t-\int\alpha_t\mathrm{d}\rho_t\right)^2\right]\mathrm{d}\rho_t\mathrm{d}t:(\rho_t,v_t,\alpha_t)_{0\le t\le1}\\\text{solves }\partial_t\rho_t=-\mathsf{div}(\rho_t v_t)+\rho_t\alpha_t\Bigg\}.\end{aligned}$$

接下来我们遵循 Gallouët and Monsaingeon （2017）；Lu et al. （2019b）推导关于泛函 ℓ_N 的 Wasserstein-Fisher-Rao 梯度流。令 $(\rho_t)_{t\ge0}$ 为满足 C^1 的曲线，且其 $\rho_0=\rho$ 初始速度为

$$\partial_t\rho_t|_{t=0}=\zeta=-\mathsf{div}(\rho\nabla u)+\rho\left(\alpha-\int\alpha\mathrm{d}\rho\right).$$

那么应有

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0}=g_\rho^{\mathsf{WFR}}(\mathrm{grad}_{\mathsf{WFR}}\ell_N(\rho),\zeta).$$

上述方程的左端等于

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0}&=\int\delta\ell_N(\rho)\zeta\mathrm{d}x=\int\delta\ell_N(\rho)\left[-\mathsf{div}(\rho\nabla u)+\rho\left(\alpha-\int\alpha\mathrm{d}\rho\right)\right]\mathrm{d}x\\&=-\int\langle\nabla\delta\ell_N(\rho),\nabla u\rangle\mathrm{d}\rho+\int\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\mathrm{d}\rho\right)\left(\alpha-\int\alpha\mathrm{d}\rho\right)\mathrm{d}\rho\\&=g_\rho^{\mathsf{WFR}}\left(-\mathsf{div}(\nabla\delta\ell_N(\rho)\rho)+\rho\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\mathrm{d}\rho\right),\zeta\right).\end{aligned}$$

因此

$$g_\rho^{\mathsf{WFR}}(\mathrm{grad}_{\mathsf{W}}\ell_N(\rho),\zeta)=g_\rho^{\mathsf{WFR}}\left(-\mathsf{div}(\nabla\delta\ell_N(\rho)\rho)+\rho\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\mathrm{d}\rho\right),\zeta\right)$$

对于任意 $\zeta\in\mathsf{Tan}_\rho^{\mathsf{WFR}}\mathcal{P}_2(\mathbb{R}^d)$ 都成立，因此

$$\begin{aligned}\mathrm{grad}_{\mathsf{WFR}}\ell_N(\rho)&=-\mathsf{div}(\nabla\delta\ell_N(\rho)\rho)+\rho\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\mathrm{d}\rho\right)\\&=-\mathsf{div}(\nabla\delta\ell_N(\rho)\rho)+\rho[1+\delta\ell_N(\rho)],\end{aligned}$$

其中我们用到了 $\int\delta\ell_N(\rho)\mathrm{d}\rho=-1$ 这一事实。这表明，关于 Wasserstein-Fisher-Rao度量 d_{WFR} 的 ℓ_N 梯度流为

$$\partial_t\rho_t=-\mathrm{grad}_{\mathsf{WFR}}\ell_N(\rho_t)=\mathsf{div}(\nabla\delta\ell_N(\rho_t)\rho_t)-\rho_t[1+\delta\ell_N(\rho_t)].$$

D 收敛理论（定理 4 和 2的证明）

本节将系统地讨论 Fisher-Rao 梯度下降和 Wasserstein-Fisher-Rao 梯度下降的收敛性。我们不分别证明定理 4 和2 ，而是在定理 7中证明一个更一般的收敛结果，其中定理 4 和 2 是其特例。

首先，我们定义 Wasserstein 梯度下降更新。

定义 2 （Wasserstein 梯度下降）。 对于任意 $\rho \in \mathcal{P}(\mathbb{R}^d)$ 和 $\eta \geq 0$，我们定义 $\rho^{\mathrm{W},\eta} = (\mathrm{id} - \eta\nabla\delta\ell_N(\rho))_{\#}\rho$。给定初始分布 $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ ，用于求解 （1.1）的 *Wasserstein* 梯度下降通过 $\rho_{n+1} = \rho^{\mathrm{W},\eta}_n$, $\forall n \geq 0$递归定义。

换言之， $\rho^{\mathrm{W},\eta}$是通过映射 $x \mapsto x - \eta\nabla\delta\ell_N(\rho)(x)$将 ρ推前得到的分布。对于任意 $p \geq 1$，定义 p-Wasserstein 距离，即 ρ_0 与 ρ_1 在 $\mathcal{P}(\mathbb{R}^d)$ 中的距离，为

$$W_p(\rho_0, \rho_1) = \left[\inf_{\pi \in \Pi(\rho_0,\rho_1)} \int \|x - y\|_p^p \, \pi(\mathrm{d}x, \mathrm{d}y)\right]^{1/p}.$$

下面的引理刻画了执行一步 Wasserstein 梯度下降后损失函数 ℓ_N 的下降幅度，该幅度可以由两个迭代结果之间的 Wasserstein 距离平方给出下界。

引理 3 (瓦 瑟斯坦梯度下降)。 设假设 *1* 成立。任取 $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ 并定义 *ne*

$$c_0 = \frac{e^{-\ell_N(\rho_0)}\underline{\phi}(\inf\{r \geq 0 : \bar{\phi}(r) \leq e^{-\ell_N(\rho_0)}/2\} + \mathrm{diam}(\Omega))}{2\bar{\phi}(0)}.$$

假设

$$0 \leq \eta < \frac{c_0}{\sup_{x\in\mathbb{R}^d}\|\nabla^2\phi(x)\|_2 + \sup_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2^2/c_0}.$$

定义 $\rho_{n+1} = \rho^{\mathrm{W},\eta}_n$对于 $n \geq 0$。那么，对于任意 $n \geq 0$,

$$\ell_N(\rho_{n+1}) - \ell_N(\rho_n) \leq -\frac{\eta}{2}\mathbb{E}_{Y\sim\rho_n}\|\nabla\delta\ell_N(\rho_n)(Y)\|_2^2 \leq -\frac{1}{2\eta}W_2^2(\rho_{n+1}, \rho_n).$$

此外，如果 支撑$(\rho_0) = \mathbb{R}^d$，那么对于所有 $n \geq 0$，支撑$(\rho_n) = \mathbb{R}^d$ 。

证明。 见附录 D.1。 □

接下来定义 Fisher-Rao 梯度下降的更新规则。

定义 3 （Fisher-Rao 梯度下降）。 对于任意 $\rho \in \mathcal{P}(\mathbb{R}^d)$ 和 $\gamma \in [0, 1]$，我们通过 $\rho^{\mathrm{FR},\gamma} \in \mathcal{P}(\mathbb{R}^d)$ 定义

$$\frac{\mathrm{d}\rho^{\mathrm{FR},\gamma}}{\mathrm{d}\rho} = 1 - \gamma\left[\delta\ell_N(\rho) + 1\right].$$

给定初始分布 $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ ，求解 （1.1）的 *Fisher-Rao* 梯度下降按递推方式定义为 $\rho_{n+1} = \rho^{\mathrm{FR},\gamma}_n$, $\forall n \geq 0$。

容易看出 $\rho^{\mathrm{FR},\gamma} = (1-\gamma)\rho + \gamma\rho^{\mathrm{FR},1}$以及

$$\frac{\mathrm{d}\rho^{\mathrm{FR},1}}{\mathrm{d}\rho}(x) = -\delta\ell_N(\rho)(x) = \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - x)}{(\rho * \phi)(X_i)}, \qquad \forall x \in \mathbb{R}^d.$$

根据附录 C.2.1 可知，带有 $\gamma = 1$ 的 Fisher-Rao 梯度下降可视为固定位置 EM 算法。下面的引理表明，执行一步 Fisher-Rao 梯度下降后，损失函数的下降量可以用两个迭代点之间的 KL 散度作为下界。

Lemma 4 （Fisher-Rao 梯度下降）。 任取 $\rho \in \mathcal{P}(\mathbb{R}^d)$ 和 $\gamma \in [0, 1]$。我们有

$$\ell_N(\mu) \leq \ell_N(\rho) + \mathsf{KL}(\rho^{\mathrm{FR},1}\|\mu) - \mathsf{KL}(\rho^{\mathrm{FR},1}\|\rho), \qquad \forall\, \mu \ll \rho,$$

and

$$\ell_N(\rho^{\mathrm{FR},\gamma}) - \ell_N(\rho) \leq -\mathsf{KL}(\rho^{\mathrm{FR},\gamma}\|\rho).$$

证明。 参见附录 D.3。 □

最后，我们定义 Wasserstein-Fisher-Rao 梯度下降，它可以视为迭代地先执行一步 Wasserstein 梯度下降（参见定义 2），再执行一步 Fisher-Rao 梯度下降（参见定义 3）。

定义 4 (Wasserstein-Fisher-Rao 梯度下降)。 令 $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$、 $\eta \geq 0$ 和 $\gamma \in [0,1]$。 *Wasserstein-Fisher-Rao* 梯度下降定义为 $\widetilde{\rho}_n = \rho^{\mathrm{W}, \eta}_n$和 $\rho_{n+1} = \widetilde{\rho}^{\mathrm{FR}, \gamma}_n$，对所有 $n \geq 0$均成立。

Wasserstein 梯度下降和 Fisher-Rao 梯度下降分别是取 $\gamma = 0$ 和 $\eta = 0$时 Wasserstein-Fisher-Rao 梯度下降的特例。下面的定理表明，如果 Wasserstein-Fisher-Rao 梯度下降弱收敛到某个极限分布，那么该弱极限就是非参数最大似然估计。

定理 7。 假设假设条件 *1*成立。令 $\{\rho_n\}_{n=0}^{\infty}$ 为 *Wasserstein-Fisher-Rao* 梯度下降的迭代序列，并且

$$c_0 = \frac{e^{-\ell_N(\rho_0)}\underline{\phi}(\inf\{r \geq 0 : \bar{\phi}(r) \leq e^{-\ell_N(\rho_0)}/2\} + \mathrm{diam}(\Omega))}{2\bar{\phi}(0)}.$$

若 支撑集$(\rho_0) = \mathbb{R}^d$, $\gamma \in (0,1]$, 和

$$0 \leq \eta < \frac{c_0}{\sup_{x\in\mathbb{R}^d} \|\nabla^2\phi(x)\|_2 + \sup_{x\in\mathbb{R}^d} \|\nabla\phi(x)\|_2^2/c_0},$$

则对所有 $n \geq 0$， $\ell_N(\rho_{n+1}) \leq \ell_N(\rho_n)$ 均成立，且 支撑集$(\rho_n) = \mathbb{R}^d$。此外，若 $\{\rho_n\}_{n=0}^{\infty}$ 弱收敛到 $\rho_\infty \in \mathcal{P}(\mathbb{R}^d)$，则此极限 ρ_∞ 是非参数最大似然估计，即式 (1.1)的一个最优解。

证明。 见附录 D.4。 □

定理 7 直接推出定理 2 成立。取 $\eta = 0$，即可推出定理4 成立。注意，定理 7 要求 $\gamma > 0$，因此无法为 Wasserstein 梯度下降提供收敛保证。

以下是证明弱极限 ρ_∞ 最优性的一些关键思路。如果 ρ_∞ 不是最优解，那么定理 1 意味着对于某些 $x_0 \in \mathbb{R}^d$ 和 $\varepsilon > 0$, $\delta\ell_N(\rho_\infty)(x_0) < -1-\varepsilon$ 成立。我们可以找到合适的 $y \in (-1-\varepsilon, -1-\varepsilon/2)$，并研究次水平集 $\bar{S} = \{x \in \mathbb{R}^d : \delta\ell_N(\rho_\infty) \leq y\}$。注意，当 n 足够大时，我们有 $\nabla\delta\ell_N(\rho_n) \approx \nabla\delta\ell_N(\rho_\infty)$。我们分别分析 Wasserstein 步骤和 Fisher-Rao 步骤:

- 对于 Wasserstein 步骤，我们证明对于任意 $x \in \bar{S}$，梯度下降步 $x - \eta\nabla\delta\ell_N(\rho_n)(x)$ 仍处于 S 中；这由梯度步的定义以及步长选得足够小这一事实可知。其中一个关键步骤是选择 y，使得梯度 $\nabla\delta\ell_N(\rho_\infty)$ 在水平集 $\{x \in \mathbb{R}^d$ 上不为零： $\delta\ell_N(\rho_\infty) = y\}$。由 Sard 引理可直接得出这样的 y 存在。由于 $\widetilde{\rho}_n = (\mathrm{id} - \nabla\delta\ell_N(\rho_n))_{\#}\rho_n$，我们得到 $\widetilde{\rho}_n(S) \geq \rho_n(S)$。
- 对于 Fisher-Rao 步骤，我们首先证明，对于所有 $x \in S$， $\delta\ell_N(\widetilde{\rho}_n)(x) < -1-\varepsilon/4$。回顾 that

$$\frac{\mathrm{d}\rho_{n+1}}{\mathrm{d}\widetilde{\rho}_n}(\cdot) = (1-\gamma) + \gamma\cdot[-\delta\ell(\widetilde{\rho}_n)(\cdot)],$$

即可得出$\rho_{n+1}(S) \geq (1+\varepsilon/4)\widetilde{\rho}_n(S)$。

将这两个步骤结合起来，我们得到存在 N使得 $\rho_{n+1}(S) \geq (1+\varepsilon/4)\rho_n(S)$ 对所有 $n > N$ 均成立。注意，这一几何改进完全由证明中的 Fisher-Rao 部分驱动。最后，我们利用 $\mathrm{supp}(\rho_N) = \mathbb{R}^d$得到 $\widetilde{\rho}_N(\bar{S}) > 0$，从而 $\lim_{n\to\infty}\widetilde{\rho}_n(\bar{S}) = \infty$，这与假设矛盾，因此 ρ_∞ 必定是 (1.1) 的一个最优解。

D.1 引理3 的证明

我们将用到一个 Wasserstein 梯度下降的下降引理，其证明推迟至附录 D.2。这类引理在凸优化中是标准结果（例如参见 Bubeck，2015，式 (3.5)）。对于 Wasserstein 空间上的优化，Salim 等人（2020年）也曾在泛函沿广义测地线凸这一假设下给出该引理；然而，负对数似然并不满足这一假设 ℓ_N。

引理 5 (一个下降引理)。 设假设 *1* 成立。任取 $\rho \in \mathcal{P}(\mathbb{R}^d)$。定义 $c = \text{下确界}_{x\in\Omega}(\rho * \phi)(x)$，$G = \text{上确界}_{x\in\mathbb{R}^d} \|\nabla\phi(x)\|_2$ 以及 $H = \text{上确界}_{x\in\mathbb{R}^d} \|\nabla^2\phi(x)\|_2$。

- 我们有

$$\ell_N(\rho^{\mathrm{W},\eta}) - \ell_N(\rho) \le -\eta\left[1 - \frac{\eta}{2c}\left(H + \frac{G^2}{c}\right)\right]\mathbb{E}_{Y\sim\rho}\|\nabla\delta\ell_N(\rho)(Y)\|_2^2.$$

  此外，我们有 $\text{上确界}_{x\in\mathbb{R}^d} \|\nabla^2\delta\ell_N(\rho)(x)\|_2 \le H/c$。

- 如果 $0 \le \eta < c/H$ 且 $\mathrm{supp}(\rho) = \mathbb{R}^d$，那么 $\mathrm{supp}(\rho^{\mathrm{W}^\eta}) = \mathbb{R}^d$。

现在回到引理 3。令 $R = \text{下确界}\{r \ge 0 : \bar{\phi}(r) \le e^{-\ell_N(\rho_0)}/2\}$。引理 2 意味着对于任意$\mu \in \mathcal{P}(\mathbb{R}^d)$ 满足 $\ell_N(\mu) \le \ell_N(\rho_0)$，我们有

$$\inf_{x\in\Omega}(\mu * \phi)(x) \ge e^{-\ell_N(\rho_0)}\underline{\phi}(R + \mathrm{diam}(\Omega))/[2\bar{\phi}(0)] = c_0.$$

特别地，$\text{下确界}_{x\in\Omega}(\rho_0 * \phi)(x) \ge c_0$。当 $\eta \le c_0/(H + G^2/c_0)$时，引理5 以及定义 $\rho_1 = (\mathrm{id} - \eta\nabla\delta\ell_N(\rho_0))_{\#}\rho_0$ 共同得到

$$\ell_N(\rho_1) - \ell_N(\rho_0) \le -\frac{\eta}{2}\mathbb{E}_{Y\sim\rho_0}\|\nabla\delta\ell_N(\rho_0)(Y)\|_2^2 \le -\frac{1}{2\eta}W_2^2(\rho_1,\rho_0) \le 0.$$

此外，如果 $\mathrm{supp}(\rho_0) = \mathbb{R}^d$，那么 $\mathrm{supp}(\rho_1) = \mathbb{R}^d$。由 $\ell_N(\rho_1) \le \ell_N(\rho_0)$ 可得，$\inf_{x\in\Omega}(\rho_1 * \phi)(x) \ge c_0$。然后，通过归纳法即可完成证明。

D.2 引理5的证明

我们依次证明引理5 中的两个结论。

第 1 部分。令 $\nu = N^{-1}\sum_{i=1}^N \delta_{X_i}$ 为经验数据分布，并令 $h(x) = -\log x$ 对于 $x > 0$。于是我们可以写成

$$\ell_N(\rho) = -\frac{1}{N}\sum_{i=1}^N \log\left[\rho * \phi\left(X_i\right)\right] = \mathbb{E}_{X\sim\nu}\left[h\left((\rho * \phi)(X)\right)\right] = \mathbb{E}_{X\sim\nu}\left[h\left(\mathbb{E}_{Y\sim\rho}\left[\phi(X - Y)\right]\right)\right]$$

以及

$$\ell_N(\rho^{\mathrm{W},\eta}) = \mathbb{E}_{X\sim\nu}\left[h\left(\mathbb{E}_{Y\sim\rho^{\mathrm{W},\eta}}\left[\phi(X - Y)\right]\right)\right] = \mathbb{E}_{X\sim\nu}\left[h\left(\mathbb{E}_{Y\sim\rho}\left[\phi\left(X - Y + \eta\nabla\delta\ell_N(\rho)(Y)\right)\right]\right)\right].$$

定义任意 $x \in \mathbb{R}^d$ 的 $a(x) = \mathbb{E}_{Y\sim\rho}[\phi(x - Y + \eta\nabla\delta\ell_N(\rho)(Y))]$ 和 $b(x) = \mathbb{E}_{Y\sim\rho}[\phi(x - Y)] = (\rho * \phi)(x)$。于是我们可以写成

$$\ell_N(\rho) = \mathbb{E}_{X\sim\nu}\left[h\left(b(X)\right)\right], \qquad \ell_N(\rho^{\mathrm{W},\eta}) = \mathbb{E}_{X\sim\nu}\left[h\left(a(X)\right)\right].$$

注意，$h'(x) = -x^{-1} < 0$、$h''(x) = x^{-2} > 0$ 和 $h'''(x) = -2x^{-3} < 0$。对于任意 a，$b > 0$，根据泰勒定理

$$h(a) \le h(b) + h'(b)(a - b) + \frac{h''(b)}{2}(a - b)^2 = h(b) - \frac{a - b}{b} + \frac{(a - b)^2}{2b^2}.$$

综合以上两个方程可得

$$\ell_N(\rho^{\mathrm{W},\eta}) - \ell_N(\rho) = \mathbb{E}_{X\sim\nu}\left[h\left(a(X)\right) - h\left(b(X)\right)\right]$$

$$\leq \underbrace{-\mathbb{E}_{X\sim\nu}\left[\frac{a(X)-b(X)}{b(X)}\right]}_{=:\alpha_1} + \underbrace{\frac{1}{2}\mathbb{E}_{X\sim\nu}\left[\frac{[a(X)-b(X)]^2}{b(X)^2}\right]}_{=:\alpha_2}. \tag{D.1}$$

现在分别推导 α_1 和 α_2 的上界。

- 为控制 α_1，令 $G = \sup_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2$，并注意到

$$|\phi(x-y+\eta\nabla\delta\ell_N(\rho)(y)) - \phi(x-y)| \leq G\eta\|\nabla\delta\ell_N(\rho)(y)\|_2$$

对所有 x，$y\in\mathbb{R}^d$ 均成立，因此

$$|a(x)-b(x)| \leq G\eta\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2\right] \tag{D.2}$$

对所有 $x\in\mathbb{R}^d$ 均成立。于是我们有

$$\alpha_1 \overset{\text{(i)}}{\leq} \frac{G^2\eta^2\mathbb{E}^2_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|\right]}{c^2} \overset{\text{(ii)}}{\leq} \frac{G^2\eta^2}{c^2}\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right], \tag{D.3}$$

其中 (i) 由式 (D.2) 以及以下事实得出

$$c = \inf_{x\in\Omega}(\rho*\phi)(x) = \inf_{x\in\Omega} b(x) \tag{D.4}$$

而 (ii) 则由詹森不等式得出。

- 关于 α_2，令 $H = \sup_{x\in\mathbb{R}^d}\|\nabla^2\phi(x)\|_2$，则有

$$\phi(x-y+\eta\nabla\delta\ell_N(\rho)(y)) - \phi(x-y) \geq \langle\nabla\phi(x-y), \eta\nabla\delta\ell_N(\rho)(y)\rangle - \frac{H}{2}\eta^2\|\nabla\delta\ell_N(\rho)(y)\|_2^2,$$

对于任意 x，$y\in\mathbb{R}^d$，因此

$$a(x)-b(x) \geq \mathbb{E}_{Y\sim\rho}\left[\langle\nabla\phi(x-Y), \eta\nabla\delta\ell_N(\rho)(Y)\rangle\right] - \frac{H}{2}\eta^2\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right] \tag{D.5}$$

对于任意 $x\in\mathbb{R}^d$。由于 $b(x)>0$，我们有

$$\begin{aligned}
\alpha_2 &\overset{\text{(i)}}{\leq} -\mathbb{E}_{X\sim\nu}\left[\frac{\mathbb{E}_{Y\sim\rho}\left[\langle\nabla\phi(X-Y), \eta\nabla\delta\ell_N(\rho)(Y)\rangle\right] - \frac{H}{2}\eta^2\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right]}{b(X)}\right] \\
&= -\eta\mathbb{E}_{Y\sim\rho}\left[\left\langle\mathbb{E}_{X\sim\nu}\left(\frac{\nabla\phi(X-Y)}{b(X)}\right), \nabla\delta\ell_N(\rho)(Y)\right\rangle\right] + \frac{H\eta^2}{2}\mathbb{E}_{X\sim\nu}\left[\frac{1}{b(X)}\right]\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right] \\
&\overset{\text{(ii)}}{=} -\eta\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right] + \frac{H\eta^2}{2}\mathbb{E}_{X\sim\nu}\left[\frac{1}{b(X)}\right]\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right] \\
&\overset{\text{(iii)}}{\leq} \left(-\eta + \frac{H\eta^2}{2c}\right)\mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right].
\end{aligned} \tag{D.6}$$

这里 (i) 利用了（D.5）；(ii) 成立，因为

$$\delta\ell_N(\rho)(y) = -\mathbb{E}_{X\sim\nu}\left[\frac{\phi(X-y)}{b(X)}\right], \qquad \nabla\delta\ell_N(\rho)(y) = \mathbb{E}_{X\sim\nu}\left[\frac{\nabla\phi(X-y)}{b(X)}\right], \tag{D.7}$$

因此

$$\mathbb{E}_{Y\sim\rho}\left[\left\langle\mathbb{E}_{X\sim\nu}\left(\frac{\nabla\phi(X-Y)}{b(X)}\right), \nabla\delta\ell_N(\rho)(Y)\right\rangle\right] = \mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2\right];$$

而 (iii) 由 (D.4) 得出。

综合 (D.1)、(D.3) 和 (D.6) 可得

$$\begin{aligned}\ell_N(\rho^{\mathrm{W},\eta}) - \ell_N(\rho) \le \alpha_1 + \alpha_2 &\le \left(-\eta + \frac{H\eta^2}{2c} + \frac{G^2\eta^2}{2c^2} \right) \mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2 \right] \\ &= -\eta\left[1 - \frac{\eta}{2c}\left(H + \frac{G^2}{c} \right) \right] \mathbb{E}_{Y\sim\rho}\left[\|\nabla\delta\ell_N(\rho)(Y)\|_2^2 \right].\end{aligned}$$

最后，我们由 (D.7) 和 (D.4) 得知

$$\left\| \nabla^2\delta\ell_N(\rho)(x) \right\|_2 = \left\| \mathbb{E}_{X\sim\nu}\left[\frac{\nabla^2\phi(X-x)}{b(X)} \right] \right\|_2 \le \frac{\sup_{x\in\mathbb{R}^d} \|\nabla^2\phi(x)\|_2}{\inf_{x\in\mathbb{R}^d} b(x)} = \frac{H}{c}.$$

第 2 部分。 当 $\eta < c/H$ 时，映射 $x \mapsto \eta\nabla\delta\ell_N(\rho)(x)$ 是一个压缩映射。令 $\varphi(x) = x - \eta\nabla\delta\ell_N(\rho)(x)$。根据引理 6，$\varphi : \mathbb{R}^d \to \mathbb{R}^d$ 是双射，且 φ^{-1} 是利普希茨的。$\delta\ell_N(\rho)$ 的二阶可微性意味着 φ 可微，进而 φ^{-1} 也可微。若 $\mathrm{supp}(\rho) = \mathbb{R}^d$，则 $\mathrm{supp}(\rho^{\mathrm{W},\eta}) = \mathrm{supp}(\varphi_{\#}\rho) = \mathbb{R}^d$。

D.3 引理 4 的证明

令 $\nu = N^{-1}\sum_{i=1}^N \delta_{X_i}$ 为经验数据分布。对于任意 $\mu \ll \rho$，我们有

$$\begin{aligned}\ell_N(\mu) &= -\mathbb{E}_{X\sim\nu}\left[\log\left((\mu * \phi)(X) \right) \right] = -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\mu}\left[\phi(X-Y) \right] \right) \right] \\ &= -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\rho}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \cdot \phi(X-Y) \right] \right) \right] \\ &= -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\rho}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \cdot (\rho*\phi)(X) \cdot \frac{\phi(X-Y)}{(\rho*\phi)(X)} \right] \right) \right].\end{aligned}$$

对于任意 $x \in \mathbb{R}^d$，我们可以通过以下方式定义一个新的概率测度 $\rho^{\mathrm{FR},1}_{\cdot,x} \in \mathcal{P}(\mathbb{R}^d)$

$$\frac{\mathrm{d}\rho_x^{\mathrm{FR},1}}{\mathrm{d}\rho}(\cdot) = \frac{\phi(x-\cdot)}{(\rho*\phi)(x)}, \tag{D.8}$$

并且我们可以验证，$\rho^{\mathrm{FR},1}_{\cdot,x}$ 确实是一个概率测度，因为

$$\int_{\mathbb{R}^d} \mathrm{d}\rho_x^{\mathrm{FR},1} = \int_{\mathbb{R}^d} \frac{\phi(x-y)}{(\rho*\phi)(x)} \rho(\mathrm{d}x) = 1.$$

那么，根据 $t \mapsto -\log t$ 的凸性和詹森不等式，我们有

$$\begin{aligned}\ell_N(\mu) &\overset{\text{(i)}}{=} -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\rho_X^{\mathrm{FR},1}}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \cdot (\rho*\phi)(X) \right] \right) \right] \\ &\overset{\text{(ii)}}{\le} -\mathbb{E}_{X\sim\nu}\left[\mathbb{E}_{Y\sim\rho_X^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \cdot (\rho*\phi)(X) \right) \right] \right] \\ &\overset{\text{(iii)}}{=} -\mathbb{E}_{X\sim\nu}\left[\mathbb{E}_{Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \cdot (\rho*\phi)(X) \right) \cdot \frac{\phi(X-Y)}{(\rho*\phi)(X)} \right] \right] \\ &= \underbrace{-\mathbb{E}_{X\sim\nu,Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right) \cdot \frac{\phi(X-Y)}{(\rho*\phi)(X)} \right]}_{=:\beta_1} \underbrace{-\mathbb{E}_{X\sim\nu,Y\sim\rho}\left[\log\left((\rho*\phi)(X) \right) \cdot \frac{\phi(X-Y)}{(\rho*\phi)(X)} \right]}_{=:\beta_2}.\end{aligned} \tag{D.9}$$

这里，(i) 和 (iii) 使用了 (D.8)，而 (ii) 则由詹森不等式以及 $t \mapsto -\log t$ 的凸性得出。接下来，我们分别研究这两项 β_1 和 β_2。对于 β_1，我们有

$$\begin{aligned}\beta_1 &= \mathbb{E}_{Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right) \mathbb{E}_{X\sim\nu}\left[\frac{\phi(X-Y)}{(\rho*\phi)(X)} \right] \right] \\ &= \mathbb{E}_{Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right) \cdot \frac{\mathrm{d}\rho^{\mathrm{FR},1}}{\mathrm{d}\rho}(Y) \right] = \mathbb{E}_{Y\sim\rho^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right) \right]\end{aligned}$$

$$= \mathbb{E}_{Y\sim\rho^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho^{\mathrm{FR},1}}(Y)\right)\right] + \mathbb{E}_{Y\sim\rho^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\rho^{\mathrm{FR},1}}{\mathrm{d}\rho}(Y)\right)\right]$$
$$= -\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\mu\right) + \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right). \tag{D.10}$$

对于 β_2，我们有

$$\beta_2 = \mathbb{E}_{X\sim\nu}\left[\log\left((\rho*\phi)(X)\right)\mathbb{E}_{Y\sim\rho}\left[\frac{\phi(X-Y)}{(\rho*\phi)(X)}\right]\right] = \mathbb{E}_{X\sim\nu}\left[\log\left((\rho*\phi)(X)\right)\right] = -\ell_N(\rho). \tag{D.11}$$

综合式（D.9）、（D.10）和（D.11可得

$$\ell_N(\mu) \le \ell_N(\rho) + \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\mu\right) - \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right), \qquad \forall\,\mu \ll \rho.$$

取 $\mu = \rho^{\mathrm{FR},\,1}$，可得

$$\ell_N\left(\rho^{\mathrm{FR},1}\right) \le \ell_N(\rho) - \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right). \tag{D.12}$$

回顾一下，对于任意 $\gamma\in(0,\,1)$，都有 $\rho^{\mathrm{FR},\,\gamma} = (1-\gamma)\rho + \gamma\rho^{\mathrm{FR},\,1}$。因此

$$\ell_N\left(\rho^{\mathrm{FR},\gamma}\right) \overset{\text{(i)}}{\le} (1-\gamma)\ell_N(\rho) + \gamma\ell_N(\rho^{\mathrm{FR},1}) \overset{\text{(ii)}}{\le} \ell_N(\rho) - \gamma\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right), \tag{D.13}$$

其中，(i) 成立是因为 $\ell_N(\rho)$ 在 ρ上是 ℓ_2-凸的，而 (ii) 则由式（D.12）得出。由 ℓ_2-凸性以及 $\mathsf{KL}(\cdot\,\|\,\rho)$的凸性可得

$$\mathsf{KL}\left(\rho^{\mathrm{FR},\gamma}\,\|\,\rho\right) \le (1-\gamma)\mathsf{KL}\left(\rho\,\|\,\rho\right) + \gamma\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right) = \gamma\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right). \tag{D.14}$$

结合式（D.13）和（D.14）可得

$$\ell_N\left(\rho^{\mathrm{FR},\gamma}\right) \le \ell_N(\rho) - \mathsf{KL}\left(\rho^{\mathrm{FR},\gamma}\,\|\,\rho\right).$$

D.4 定理7的证明

假设对于某个 $n\ge 0$，$\ell_N(\rho_n) \le \ell_N(\rho_0)$ 成立，且 $\mathsf{supp}(\rho_n) = \mathbb{R}^d$。定义

$$c_n = \frac{e^{-\ell_N(\rho_n)}\underline{\phi}(\inf\{r\ge 0:\ \bar{\phi}(r)\le e^{-\ell_N(\rho_n)}/2\} + \mathrm{diam}(\Omega))}{2\bar{\phi}(0)}.$$

于是，$c_n \ge c_0$ 因此

$$0 \le \eta < \frac{c_n}{\sup_{x\in\mathbb{R}^d}\|\nabla^2\phi(x)\|_2 + \sup_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2^2/c_n}.$$

定理 3 以及关于 η 的上界立即给出 $\ell(\widetilde{\rho}_n) \le \ell(\rho_n)$ 以及 $\mathsf{supp}(\widetilde{\rho}_n) = \mathbb{R}^d$。令 $\nu = N^{-1}\sum_{i=1}^N \delta_{X_i}$ 为经验数据分布。根据引理 4，更新规则

$$\frac{\mathrm{d}\rho_{n+1}}{\mathrm{d}\widetilde{\rho}_n}(\cdot) = (1-\gamma) + \gamma\mathbb{E}_{X\sim\nu}\left[\frac{\phi(X-\cdot)}{(\widetilde{\rho}_n*\phi)(X)}\right], \tag{D.15}$$

以及 ϕ 的正性可知，$\ell(\rho_{n+1}) \le \ell(\widetilde{\rho}_n)$ 且 $\mathsf{supp}(\rho_{n+1}) = \mathbb{R}^d$。因此，我们可以用归纳法证明，对于所有 $n\ge 0$，不等式 $\ell(\rho_{n+1}) \le \ell(\widetilde{\rho}_n) \le \ell(\rho_n)$ 成立，且 $\mathsf{supp}(\rho_n) = \mathbb{R}^d$。根据 (B.1)，两个序列 $\{\ell(\rho_n)\}_{n=0}^\infty$ 和 $\{\ell(\widetilde{\rho}_n)\}_{n=0}^\infty$ 均收敛且极限相同。因此，

$$\ell(\rho_n) - \ell(\widetilde{\rho}_n) \to 0. \tag{D.16}$$

现在，假设 $\{\rho_n\}_{n=0}^\infty$ 弱收敛到某个 $\rho_\infty\in\mathcal{P}(\mathbb{R}^d)$。

断言 1。 $\{\delta\ell_N(\rho_n)\}_{n=0}^\infty$在紧集上均匀收敛于 $\delta\ell_N(\rho_\infty)$。

证明。 容易看出

$$\sup_{x\in\mathbb{R}^d}\|\nabla\delta\ell_N(\rho_n)(x)\|_2=\sup_{x\in\mathbb{R}^d}\left\|\mathbb{E}_{X\sim\nu}\left[\frac{\nabla\phi(X-x)}{(\rho_n*\phi)(X)}\right]\right\|_2\le\frac{\sup_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2}{\inf_{x\in\Omega}(\rho_n*\phi)(x)}.$$

假设1 保证上确界$_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2<\infty$。根据引理（2）以及 $\ell(\rho_n)\le\ell(\rho_0)$这一事实，我们有

$$\inf_{x\in\Omega}(\rho_n*\phi)(x)\ge c_0,\qquad\forall n\ge 0. \tag{D.17}$$

因此，$\{\delta\ell_N(\rho_n)\}_{n=0}^{\infty}$是一致等度连续的。因此，只需证明 $\{\delta\ell_N(\rho_n)\}_{n=0}^{\infty}$逐点收敛到 $\delta\ell_N(\rho_\infty)$即可。

由于 ϕ 有界且连续（参见假设1），$(\rho_n*\phi)(x)\to(\rho_\infty*\phi)(x)$ 对每个 $x\in\mathbb{R}^d$均成立。回顾

$$\delta\ell(\rho)(x)=-\mathbb{E}_{X\sim\nu}\left[\frac{\phi(X-x)}{(\rho*\phi)(X)}\right],\qquad\forall\rho\in\mathcal{P}(\mathbb{R}^d). \tag{D.18}$$

基于 ϕ 的有界性以及下界（D.17），我们使用有界收敛定理得到 $\delta\ell_N(\rho_n)(x)\to\delta\ell_N(\rho_\infty)(x)$ 对每个 $x\in\mathbb{R}^d$均成立。这证明了该结论。 □

断言2： $\{\nabla\delta\ell_N(\rho_n)\}_{n=0}^{\infty}$在紧集上一致收敛到 $\nabla\delta\ell_N(\rho_\infty)$。

证明。 证明与断言 1 类似，因此省略。 □

断言3： $\{\delta\ell_N(\widetilde{\rho}_n)\}_{n=0}^{\infty}$在紧集上一致收敛到 $\delta\ell_N(\rho_\infty)$。

证明。 引理 3可推出 $W_2^2(\widetilde{\rho}_n,\rho_n)\le 2\eta[\ell_N(\rho_n)-\ell_N(\widetilde{\rho}_n)]$。由（D.16），$W_2(\widetilde{\rho}_n,\rho_n)\to 0$，因此$W_1(\widetilde{\rho}_n,\rho_n)\to 0$。由于 ϕ 是利普希茨的，我们有

$$\sup_{x\in\mathbb{R}^d}|(\widetilde{\rho}_n*\phi)(x)-(\rho_n*\phi)(x)|\to 0.$$

由上述一致有界性、(D.17) 和 (D.18，我们得到

$$\sup_{x\in\mathbb{R}^d}|\delta\ell_N(\widetilde{\rho}_n)(x)-\delta\ell_N(\rho_n)(x)|\to 0.$$

然后，应用断言 1，证明即告完成。 □

现在回到定理（7）。假设 ρ_∞ 不是最优解。由定理（1）可知，存在 $\varepsilon>0$ 使得 $\delta\ell_N(\rho_\infty)(x)<-1-\varepsilon$，其中某个 $x\in\mathbb{R}^d$成立。与在断言（1）中证明 $\delta\ell_N(\rho_n)$ 逐点收敛到 $\delta\ell_N(\rho_\infty)$ 类似，我们可以利用有界收敛定理证明 $\ell(\rho_n)\to\ell(\rho_\infty)$。因此，$\ell(\rho_\infty)\le\ell(\rho_0)$。引理（2）立即推出

$$\inf_{x\in\Omega}(\rho_\infty*\phi)(x)\ge c_0,\qquad\forall n\ge 0, \tag{D.19}$$

并且 $\lim_{\|x\|_2\to\infty}\delta\ell(\rho_\infty)(x)=0$。因此，函数 $\delta\ell(\rho_\infty)$ 在某个 $x_0\in\mathbb{R}^d$ 处达到其最小值。我们有 $\delta\ell_N(\rho_\infty)(x_0)<-1-\varepsilon$。

为简化记号，令 $f_n(x)=\delta\ell_N(\rho_n(x)$ 且 $f(x)=\delta\ell_N(\rho_\infty)(x)$。根据假设1，$f\in C^d(\mathbb{R}^d)$。引理（7）的第二部分保证存在某个 $y\in(-1-\varepsilon,\ -1-\varepsilon/2)$ 使得

$$S(y)=\left\{x\in\mathbb{R}^d:\ f(x)=y\right\}$$

是紧致的，且 $\inf_{x\in S(y)}\|\nabla f(x)\|_2>0$。记 $\xi=\inf_{x\in S(y)}\|\nabla f(x)\|_2$ 和 $\bar{S}=\{x\in\mathbb{R}^d:\ f(x)\le y\}$。

$\bar{S}$次水平集 S 是紧致的。根据 $f(x_0)<y$ 这一事实以及 f的连续性，S 具有正的勒贝格测度。因此我们有

$$\rho_n(\bar{S})>0,\qquad\forall n\ge 0. \tag{D.20}$$

- 我们首先证明 $\widetilde{\rho}_n(S) \geq \rho_n(S)$ 对充分大的 n成立。根据断言 2，存在 $N > 0$ 使得

$$\sup_{x \in \bar{S}} \|\nabla f_n(x) - \nabla f(x)\|_2 \leq \xi/11, \qquad \forall n > N.$$

根据对 η 的上界假设以及估计（D.19），可得

$$\eta \leq \frac{c_0}{\sup_{x \in \mathbb{R}^d} \|\nabla^2 \phi(x)\|_2} \leq \frac{1}{\sup_{x \in \mathbb{R}^d} \|\nabla^2 f(x)\|_2}.$$

此外，引理（D.23）的第一部分意味着

$$\{x - \eta \nabla f_n(x) : \ x \in \bar{S}\} \subseteq \bar{S}, \qquad \forall n > N.$$

由于 $\widetilde{\rho}_n = (\mathrm{id} - \eta \nabla f_n)_{\#} \rho_n$，我们有

$$\widetilde{\rho}_n(\bar{S}) \geq \rho_n(\bar{S}), \qquad \forall n > N. \tag{D.21}$$

- 接下来我们证明，对于充分大的 $\rho_{n+1}(S) > (1 + \varepsilon/4)\, \widetilde{\rho}_n(S)$ 成立。根据断言 3，存在 $N' > 0$使得

$$\sup_{x \in \bar{S}} |\delta \ell(\widetilde{\rho}_n)(x) - f(x)| \leq \varepsilon/4, \qquad \forall n > N'.$$

因此，

$$\sup_{x \in \bar{S}} \delta \ell(\widetilde{\rho}_n)(x) \leq \sup_{x \in \bar{S}} f(x) + \varepsilon/4 \leq -1 - \varepsilon/4, \qquad \forall n > N'.$$

表达式（D.15）意味着

$$\frac{\mathrm{d}\rho_{n+1}}{\mathrm{d}\widetilde{\rho}_n}(x) = (1 - \gamma) + \gamma \cdot [-\delta \ell(\widetilde{\rho}_n)(x)] \geq 1 + \gamma\varepsilon/4, \qquad \forall n > N', \, x \in \bar{S}.$$

因此，我们有

$$\rho_{n+1}(\bar{S}) \geq (1 + \gamma\varepsilon/4)\widetilde{\rho}_n(\bar{S}), \qquad \forall n > N'. \tag{D.22}$$

综合（D.21）和（D.22）可得

$$\rho_{n+1}(\bar{S}) \geq (1 + \gamma\varepsilon/4)\rho_n(\bar{S}), \qquad \forall n > \max\{N, N'\}.$$

将其与（D.20）结合，立即得到 $\lim_{n \to \infty} \rho_n(\bar{S}) = \infty$，这与 $\rho_n(\mathbb{R}^d) = 1$ 对所有$n \geq 0$均成立相矛盾。因此 ρ_∞ 必定是（1.1）的最优解，即 ρ_∞ 是非参数最大似然估计。

D.5 技术引理

以下是关于恒等映射的 Lipschitz 扰动的一个标准结果。

引理 6. 设 $\mathbb{B}$ 是一个范数为 $\|\cdot\|$的巴拿赫空间。若 $f: \ \mathbb{B} \to \mathbb{B}$ 的 *Lipschitz* 常数为 $c < 1$，则$\varphi: \ \mathbb{B} \to \mathbb{B}$, $x \mapsto x + f(x)$ 是双射，且 φ^{-1} 是$(1 - c)^{-1}$-利普希茨的。

证明。 任取 $y \in \mathbb{B}$。映射 $\psi: \ x \mapsto y - f(x)$ 的 Lipschitz 常数为 $c < 1$ （相对于范数 $\|\cdot\|$）。根据巴拿赫不动点定理，存在唯一的 z 使得 $z = \psi(z)$，这意味着 $y = \varphi(z)$。因此， φ 是双射，且 $\varphi^{-1}: \ \mathbb{B} \to \mathbb{B}$ 定义良好。

对任意 $y_1, \ y_2 \in \mathbb{B}$，定义 $x_i = \varphi^{-1}(y_i)$。于是， $y_i = x_i + f(x_i)$，并且

$$\begin{aligned}
&\|y_2 - y_1\|_2 = \|\varphi(x_2) - \varphi(x_1)\|_2 \geq \|x_2 - x_1\|_2 - \|f(x_2) - f(x_1)\|_2 \geq (1 - c)\|x_2 - x_1\|_2, \\
&\|\varphi^{-1}(y_2) - \varphi^{-1}(y_1)\|_2 = \|x_2 - x_1\|_2 \leq (1 - c)^{-1}\|y_2 - y_1\|_2.
\end{aligned}$$

这证明了 φ^{-1} 的 Lipschitz 光滑性。 □

下面我们将证明，近似梯度下降的一步操作不会扩大函数的某些次水平集。

引理 7. 假设 $f \in C^2(\mathbb{R}^d)$ 有有限最小值 y_0。定义 $S(y) = \{x \in \mathbb{R}^d : f(x) = y\}$对于 $y \in \mathbb{R}$。我们有以下结论。

- 假设 上确界$_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\|_2 \le L < \infty$。任取 $\eta \in [0, 1/L]$ 和 $y > y_0$。定义 $\bar{S} = \{x \in \mathbb{R}^d : f(x) \le y\}$。若 $g : \mathbb{R}^d \to \mathbb{R}^d$ 满足

$$\sup_{x\in\bar{S}} \|g(x) - \nabla f(x)\|_2 \le \frac{1}{11} \inf_{x\in S(y)} \|\nabla f(x)\|_2,$$

then

$$\{x - \eta g(x) : x \in \bar{S}\} \subseteq \bar{S}.$$

- 假设对某个 R，下确界$_{\|x\|_2\ge R} f(x) > y_0$ 成立，此外还有 $f \in C^d(\mathbb{R}^d)$。那么，对于任意 $\varepsilon > 0$，存在 $y \in (y_0, y_0 + \varepsilon)$ 使得 $S(y)$ 是紧致的，并且

$$\inf_{x\in S(y)} \|\nabla f(x)\|_2 > 0.$$

证明。为证明第一部分，任取 $x \in S$，并证明 $x - \eta g(x) \in S$。令 $\delta = \sup_{z\in\bar{S}} \|g(z) - \nabla f(z)\|_2$。由 $\sup_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\|_2 \le L$和 $0 \le \eta \le 1/L$,

$$\begin{aligned}
f(x - \eta g(x)) &\le f(x) + \langle \nabla f(x), -\eta g(x)\rangle + \frac{L}{2}\|\eta g(x)\|_2^2 \\
&\le f(x) + \langle \nabla f(x), -\eta \nabla f(x) - \eta[g(x) - \nabla f(x)]\rangle + \frac{L\eta^2}{2}[\|\nabla f(x)\|_2 + \|g(x) - \nabla f(x)\|_2]^2 \\
&\le f(x) - \eta\|\nabla f(x)\|_2^2 + \eta\delta\|\nabla f(x)\|_2 + \frac{L\eta^2}{2}[\|\nabla f(x)\|_2^2 + 2\delta\|\nabla f(x)\|_2 + \delta^2] \\
&= f(x) + \eta\|\nabla f(x)\|_2^2\left[-\left(1 - \frac{L\eta}{2}\right) + (1 + L\eta)\frac{\delta}{\|\nabla f(x)\|_2} + \frac{L\eta}{2}\left(\frac{\delta}{\|\nabla f(x)\|_2}\right)^2\right] \\
&\le f(x) + \eta\|\nabla f(x)\|_2^2\left[-\frac{1}{2} + \frac{2\delta}{\|\nabla f(x)\|_2} + \frac{1}{2}\left(\frac{\delta}{\|\nabla f(x)\|_2}\right)^2\right]. \qquad \text{(D.23)}
\end{aligned}$$

定义 $\xi =$ 下确界$_{z\in S(y)} \|\nabla f(z)\|_2$。若 $\xi = 0$，则 $\delta = 0$。(D.23) 的界给出

$$f(x - \eta g(x)) \le f(x) - \frac{\eta}{2}\|\nabla f(x)\|_2^2 \le f(x) \le y$$

因此 $x - \eta g(x) \in S$。从现在起，我们假设 $\xi > 0$。

情形 1. 当 $\|\nabla f(x)\|_2 > (\sqrt{5} + 2)\delta$时，有

$$\frac{\delta}{\|\nabla f(x)\|_2} < \frac{1}{\sqrt{5} + 2} = \sqrt{5} - 2.$$

由 (D.23) 可得,

$$f(x - \eta g(x)) - f(x) \le \eta\|\nabla f(x)\|_2^2\left[\frac{1}{2}\left(\frac{\delta}{\|\nabla f(x)\|_2} + 2\right)^2 - \frac{5}{2}\right] \le 0.$$

因此， $f(x - \eta g(x)) \le f(x) \le y$且 $x - \eta g(x) \in S$。

情形 2. 当 $\|\nabla f(x)\|_2 \le (\sqrt{5} + 2)\delta$时，我们使用 $\delta \le \xi/11$ 和 $\sqrt{5} < 5/2$ 得到

$$\|\nabla f(x)\|_2 \le (\sqrt{5} + 2)\delta \le \frac{\sqrt{5} + 2}{11}\xi < \frac{9\xi}{22} < \frac{\xi}{2}.$$

¯因此， $x \notin S(y)$。由于 $x \in S$，我们必须有 $f(x) < y$。

注意，对于任意 $z \in S(y)$，我们有 $\|\nabla f(z)\|_2 \geq \xi > 0$。根据 $\sup_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\|_2 \leq L$，我们有

$$\|z - x\|_2 \geq \frac{\|\nabla f(z) - \nabla f(x)\|_2}{L} \geq \frac{\|\nabla f(z)\|_2 - \|\nabla f(x)\|_2}{L} > \frac{\xi - \xi/2}{L} = \frac{\xi}{2L}.$$

因此，$\inf_{z\in S(y)} \|z - x\|_2 \geq \frac{\xi}{2L}$。¯我们有

$$B \cap S(y) = \varnothing, \qquad \text{where} \qquad B = \left\{x' \in \mathbb{R}^d : \|x' - x\|_2 < \frac{\xi}{2L}\right\}. \tag{D.24}$$

¯我们声称 $B \subseteq S$。若此结论不成立，则对某个 $x' \in B$，$f(x') > y$成立。由于 $f(x) < y$且 f是连续的，存在 $t \in (0, 1)$，使得 $f((1-t)x + tx') = y$，即 $(1-t)x + tx' \in S(y)$。事实 $(1-t)x + tx' \in B$ 导致 $B \cap S(y) \neq \varnothing$，这与 (D.24) 矛盾。

另一方面，我们有

$$\|[x - \eta g(x)] - x\|_2 = \eta\|g(x)\|_2 \leq \eta\|\nabla f(x)\|_2 + \eta\delta \leq \frac{(\sqrt{5}+3)\delta}{L} \leq \frac{(\sqrt{5}+3)\xi}{11L} < \frac{\xi}{2L}.$$

¯因此， $x - \eta g(x) \in B \subseteq S$。这证明了第一部分。

现在来看第二部分。令 $y_1 = \inf_{\|x\|_2\geq R} f(x)$。由于 f 的连续性，像集 $\{f(x) : x \in \mathbb{R}^d\}$ 包含区间 (y_0, y_1)。由于 $f\colon \mathbb{R}^d \to \mathbb{R}$ 是 C^d，Sard 引理断言，临界值集合，即临界点 $\{f(x) : \nabla f(x) = 0\}$ 的像集，其勒贝格测度为 0。因此，对于任意 $\varepsilon > 0$，存在一个正则值 $y \in (y_0, \min\{y_0 + \varepsilon, y_1\})$，即只要 $f(x) = y$，就有 $\nabla f(x) \neq 0$。由于 $y < y_1$，$S(y) \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \geq R\}$ 必须是紧致的。∇f 的连续性意味着 $\inf_{x\in S(y)} \|\nabla f(x)\|_2 > 0$。

□

E 粒子梯度流的适定性

E.1 Fisher-Rao梯度流（定理 5 的证明）

本节将证明，对于常微分方程组（3.15）

$$\dot{\omega}_t^{(j)} = -\omega_t^{(j)}\left[1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(l)}\phi\left(X_i - \mu^{(l)}\right)}\right], \quad \forall t \geq 0,\ j \in [m]$$

以初值 $\omega_0 \in \Delta^{m-1}$出发，解存在且唯一，并且 $(\rho_t)_{t\geq 0}$ 其中 $\rho_t := \sum_{l=1}^{m}\omega_t^{(l)}\delta_{\mu^{(l)}}$ 是式（3.11）意义下的Fisher-Rao梯度流。

首先，我们将使用皮卡-林德洛夫定理证明解的存在唯一性。定义函数 $f : \mathbb{R}^m \to \mathbb{R}^m$ 如下

$$f(y) = [f_j(y)]_{1\leq j\leq m}, \qquad f_j(y) = -y_j\left[1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^{m}y_l\phi\left(X_i - \mu^{(l)}\right)}\right].$$

然后我们可以将常微分方程组改写为 $\dot{\omega}_t = f(\omega_t)$。对于任意满足 $\|y - \omega_0\|_2 \leq \varepsilon$ 的 $y \in \mathbb{R}^m$，其中某个 $\delta > 0$ 将在后文指定，我们知道

$$\begin{aligned}
\sum_{l=1}^{m} y_l\phi\left(X_i - \mu^{(l)}\right) &\geq \sum_{l=1}^{m}\omega_0^{(l)}\phi\left(X_i - \mu^{(l)}\right) - \sum_{l=1}^{m}\left(\omega_0^{(l)} - y_l\right)\phi\left(X_i - \mu^{(l)}\right) \\
&\overset{\text{(i)}}{\geq} \min_{i\in[N],l\in[m]}\phi\left(X_i - \mu^{(l)}\right) - \|y - \omega_0\|_2\sum_{l=1}^{m}\phi^2\left(X_i - \mu^{(l)}\right) \\
&\overset{\text{(ii)}}{\geq} \min_{i\in[N],l\in[m]}\phi\left(X_i - \mu^{(l)}\right) - \frac{\varepsilon m}{(2\pi)^d}
\end{aligned}$$

对于任意 $i \in [N]$，其中 (i) 由 $\omega_0^{(l)} \in \Delta^{m-1}$ 和柯西-施瓦茨不等式得出，而 (ii) 成立是因为$\|\phi\|_\infty \leq (2\pi)^{-d/2}$。因此，取

$$\varepsilon := \frac{(2\pi)^d}{2m} \min_{i\in[N], l\in[m]} \phi\left(X_i - \mu^{(l)}\right),$$

我们知道，对于任意 $\|y - \omega_0\|_2 \leq \varepsilon$，都有

$$\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right) \geq \frac{1}{2} \min_{i\in[N], l\in[m]} \phi\left(X_i - \mu^{(l)}\right) \triangleq \delta. \tag{E.1}$$

因此，我们可以验证

$$\begin{aligned}\left|\left[\nabla f_j(y)\right]_j\right| &= \left|-\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)}\right] - \left[\frac{1}{N}\sum_{i=1}^N \frac{y_j \phi^2\left(X_i - \mu^{(j)}\right)}{\left[\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)\right]^2}\right]\right| \\ &\leq 1 + \frac{\|\phi\|_\infty}{\delta} + \frac{(1+\varepsilon)\|\phi\|_\infty^2}{\delta^2} = 1 + \frac{1}{(2\pi)^{d/2}\delta} + \frac{1+\varepsilon}{(2\pi)^d \delta^2},\end{aligned}$$

并且对于 $l \neq j$

$$\left|\left[\nabla f_j(y)\right]_l\right| = \left|-y_j\left[\frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(j)}\right)\phi\left(X_i - \mu^{(l)}\right)}{\left[\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)\right]^2}\right]\right| \leq \frac{(1+\varepsilon)\|\phi\|_\infty^2}{\delta^2} = \frac{1+\varepsilon}{(2\pi)^d \delta^2}.$$

因此，我们知道，对于任意

$$\max_{y:\|y-\omega_0\|_2\leq\varepsilon} \|\nabla f_j(y)\|_2 \leq \sqrt{m}\left(1 + \frac{1}{(2\pi)^{d/2}\delta} + \frac{1+\varepsilon}{(2\pi)^d\delta^2}\right) \triangleq C_{\mathsf{lip}}, \tag{E.2}$$

因此， $f(y)$ 在 $\{y$ 上是 C_{Lip}-利普希茨连续的: $\|y - \omega_0\|_2 \leq \varepsilon\}$ 其中 $C_{\mathsf{Lip}} := \sqrt{m}C_{\mathsf{lip}}$。此外，很容易证明

$$\max_{y:\|y-\omega_0\|_2\leq\varepsilon} \|f_j(y)\|_2 \leq -y_j + \frac{1}{N}\sum_{i=1}^N \frac{y_j\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)} \leq 1 + \varepsilon \triangleq M.$$

根据皮卡-林德洛夫定理，存在 $t_0 > 0$，使得该常微分方程在时间区间 $[0, t_0]$ 上有唯一解。我们首先验证，对于任意 $j \in [m]$ 和 $t \in [0, t_0]$，都有 $\omega_t^{(j)} > 0$。若此结论不成立，则定义

$$t_\star := \min\left\{t \in [0, t_0] : \exists j \in [m] \text{ s.t. } \omega_t^{(j)} \leq 0\right\}$$

并假设 $\omega^{(j_\star)}{}_{t_\star} \leq 0$ 对于 $j_\star \in [m]$。那么我们知道，对于任意 $j \in [m]$ 和 $0 \leq t \leq t_\star$，都有 $\omega_t^{(j)} > 0$，因此对于所有 $t \in [0, t_\star]$，都有$\dot{\omega}_t^{(j)} \geq -\omega_t^{(j)}$。接着，我们可以使用格朗沃尔引理，得到对于所有 $t \in [0, t_\star]$，都有 $\omega_t^{(j)} \geq \omega_0^{(j)} e^{-t}$，因此 $\omega^{(j_\star)}{}_{t_\star} > 0$，这就产生了矛盾。此外，我们还可以验证

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{j=1}^m \omega_t^{(j)} = \sum_{j=1}^m \dot{\omega}_t^{(j)} = -\sum_{j=1}^m \omega_t^{(j)}\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^m \omega_t^{(l)}\phi\left(X_i - \mu^{(l)}\right)}\right] = 0$$

对于所有 $t \in [0, t_0]$。因此 $\omega_t \in \Delta^{m-1}$ 对于任意 $0 \leq t \leq t_0$。重复上述相同步骤（注意，上述证明仅依赖于 $\omega_0 \in \Delta^{m-1}$，而 t_0 仅依赖于普适常数 C_{Lip}和 M，并不依赖于 ω_0），我们可以证明该常微分方程在 $[t_0, 2t_0]$、$[2t_0, 3t_0]$等区间上存在唯一解。这表明常微分方程组（3.15）的解存在且唯一。

接下来，我们证明$(\rho_t)_{t\geq 0}$ 定义为 $\rho_t := \sum_{l=1}^m \omega_t^{(l)}\delta_{\mu^{(l)}}$ 是（3.11）的解。注意， ρ_t 是一个概率测度，因为我们已经证明，对于任意 $\omega_t \in \Delta^{m-1}$， $t \geq 0$。对于任意测试函数 $\varphi(x) \in C_c^\infty$，有

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} \varphi(x)\,\rho_t(\mathrm{d}x) = \frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{j=1}^m \omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right)\right] = \sum_{j=1}^m \dot{\omega}_t^{(j)}\varphi\left(\mu_t^{(j)}\right)$$

$$= -\sum_{j=1}^{m}\left[1+\delta\ell_N\left(\rho_t\right)\left(\mu_t^{(j)}\right)\right]\omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right)$$
$$= -\int_{\mathbb{R}^d}\left[1+\delta\ell_N\left(\rho_t\right)(x)\right]\varphi(x)\,\rho_t(\mathrm{d}x).$$

这就证明了

$$\partial_t\rho_t = -\left[\delta\ell\left(\rho_t\right)+1\right]\rho_t$$

以分布的意义成立。

E.2 Wasserstein梯度流（定理 6的证明）

在本节中，我们将证明常微分方程组（3.19）

$$\dot{\mu}_t^{(j)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\left(X_i-\mu_t^{(j)}\right),\quad \forall t\geq 0,\, j\in[m]$$

具有唯一解，并且 $(\rho_t)_{t\geq 0}$ 其中 $\rho_t := m^{-1}\sum_{l=1}^{m}\delta_{\mu_t^{(l)}}$ 是根据（3.18）意义下的 Wasserstein梯度流。

我们将使用皮卡-林德洛夫定理来证明解的存在唯一性。定义一个充分大的常数

$$R := \max_{1\leq i\leq N}\|X_i\|_2.$$

对于每个 $j\in[m]$，定义函数 $f^{(j)}:\ \mathbb{R}^{md}\to\mathbb{R}^d$ 为

$$f^{(j)}(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}\left(X_i-z_j\right),$$

其中 $z=[z_j]_{1\leq j\leq m}\in\mathbb{R}^{md}$和 $z_1,\ldots,z_m\in\mathbb{R}^d$，并令 $f(z)=[f^{(j)}(z)]_{1\leq j\leq m}$。于是可将常微分方程组写为 $\dot{\mu}_t=f(\mu_t)$，其中 $\mu_t=[\mu_t^{(j)}]_{1\leq j\leq m}$。记 $f^{(j)}(z)=[f_k^{(j)}(z)]_{1\leq k\leq d}$。对于任意满足 $\max_{j\in[m]}\|z_j\|_2\leq 2R$ 的 $z\in\mathbb{R}^{md}$，我们有

$$\min_{1\leq i\leq N}\frac{1}{m}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)\geq\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{9}{2}R^2\right)\triangleq\delta. \tag{E.3}$$

接下来我们可以计算 $l\neq j$

$$\begin{aligned}\left\|\nabla_{z_l}f_k^{(j)}(z)\right\|_2 &= \left\|\nabla_{z_l}\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}e_k^\top\left(X_i-z_j\right)\right\|_2\\ &= \left\|-\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)m^{-1}\phi\left(X_i-z_l\right)}{\left[m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)\right]^2}e_k^\top\left(X_i-z_j\right)\left(X_i-z_l\right)\right\|_2\\ &\overset{\text{(i)}}{\leq}\frac{m^{-1}\|\phi\|_\infty^2}{\delta^2}\frac{1}{N}\sum_{i=1}^{N}\left|e_k^\top\left(X_i-z_j\right)\right|\|X_i-z_l\|_2\overset{\text{(ii)}}{\leq}\frac{9m^{-1}}{\delta^2(2\pi)^d}R^2,\end{aligned}$$

其中 (i) 利用了 (E.3)，(ii) 则由 最大值$_{i\in[N]}\|X_i\|_2\leq R$ 和 最大值$_{j\in[m]}\|z_j\|_2\leq 2R$ 得出。类似地，我们有

$$\begin{aligned}\left\|\nabla_{z_j}f_k^{(j)}(z)\right\|_2 &= \left\|\nabla_{z_j}\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}e_k^\top\left(X_i-z_j\right)\right\|_2\\ &\leq\left\|-\frac{1}{N}\sum_{i=1}^{N}\frac{\phi^2\left(X_i-z_j\right)m^{-1}}{\left[m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)\right]^2}e_k^\top\left(X_i-z_j\right)\left(X_i-z_j\right)\right\|_2\end{aligned}$$

$$
\begin{aligned}
&+\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}\left[e_k^{\top}\left(X_i-z_j\right)\left(X_i-z_j\right)-e_k\right]\right\|_2 \\
&\leq 9\left(\frac{m^{-1}}{\delta^2\left(2\pi\right)^d}+\frac{1}{\delta\left(2\pi\right)^{d/2}}\right)R^2+\frac{\|\phi\|_\infty}{\delta}.
\end{aligned}
$$

因此，我们有

$$
\left\|\nabla_z f_k^{(j)}\left(z\right)\right\|_2=\sqrt{\sum_{l=1}^{m}\left\|\nabla_{z_l} f_k^{(j)}\left(z\right)\right\|_2^2}\leq\sqrt{\left(m+1\right)\left(\frac{9m^{-1}}{\delta^2\left(2\pi\right)^d}R^2\right)^2+2\frac{1}{\delta^2\left(2\pi\right)^d}}\triangleq C_{\mathsf{lip}}. \tag{E.4}
$$

因此， $f_{(j)}(z)$ 关于 $\{z^-$是 $\sqrt{d}C_{\mathsf{Lip}}$-利普希茨连续的：$\max_{j\in[m]}\|z_j\|_2\leq 2R\}$，因此 $f(z)$ 是 C_{Lip}-利普希茨连续的，其中 $C_{\mathsf{Lip}}:=C_{\mathsf{lip}}\sqrt{md}$。此外，很容易证明，对于任意满足 $\max_{j\in[m]}\|z_j\|_2\leq 2R$ 的 $z\in\mathbb{R}^{md}$，

$$
\left\|f^{(j)}\left(z\right)\right\|_2\leq\frac{\|\phi\|_\infty}{\delta}3R\triangleq M
$$

对所有 $1\leq j\leq m$ 均成立。

回顾一下， $\mu_0^{(1)},\ldots,\mu_0^{(m)}$ 是从 Uniform($\{X_i\}_{1\leq i\leq N}$) 中独立同分布抽样得到的，因此 $\max_{j\in[m]}\|\mu_0^{(j)}\|_2\leq R$。于是我们有

$$
\left\{z:\|z-\mu_0\|_2\leq R\right\}\subseteq\left\{z:\max_{j\in[m]}\|z_j\|_2\leq 2R\right\},
$$

其中 $\mu_0=[\mu_0^{(j)}]_{1\leq j\leq m}$。因此， $f(z)$ 关于 $\{z$是 $\sqrt{md}C_{\mathsf{Lip}}$-利普希茨连续的： $\|z-\mu_0\|_2\leq R\}$。接着我们可以利用皮卡-林德洛夫定理证明，存在 $t_0>0$ ，使得该常微分方程在时间区间 $[0,t_0]$ 上有唯一解。对于任意 $t\in[0,t_0]$ 和 $j\in[m]$，我们可以计算

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|\mu_t^{(j)}\|_2^2=2\langle\mu_t^{(j)},\dot{\mu}_t^{(j)}\rangle&=\frac{2}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\mu_t^{(j)\top}\left(X_i-\mu_t^{(j)}\right)\\
&\leq\frac{2}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\left(\|X_i\|_2\|\mu_t^{(j)}\|_2-\|\mu_t^{(j)}\|_2^2\right)\\
&\leq\frac{2}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\left(R-\|\mu_t^{(j)}\|_2\right)\|\mu_t^{(j)}\|_2,
\end{aligned}
$$

其中在倒数第二步使用了柯西-施瓦茨不等式。这表明 $\mathrm{d}\ \mathrm{d}t\|\mu_t^{(j)}\|_2^2<0$ 只要 $\|\mu_{t_0}^{(j)}\|_2\gtrsim R$，因此 $\max_{j\in[m]}\|\mu_{t_0}^{(j)}\|_2\leq R$。然后，我们可以重复上述相同的分析（注意，上述证明只要求 $\max_{j\in[m]}\|\mu_0^{(j)}\|_2\leq R$，且 t_0 只依赖于普适常数C_{Lip} 以及 M），从而证明该常微分方程在 $[t_0,2t_0]$， $[2t_0,3t_0]$等区间上都有唯一解。这表明常微分方程组（3.15）的解存在且唯一。

最后我们验证 $(\rho_t)_{t\geq 0}$ 定义为 $\rho_t:=\sum_{l=1}^{m}m^{-1}\delta_{\mu_t^{(l)}}$ 满足（3.11）。对于任意测试函数 $\varphi(x)\in C_c^\infty$，有

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi\left(x\right)\rho_t\left(\mathrm{d}x\right)=\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{m}\sum_{j=1}^{m}\varphi\left(\mu_t^{(j)}\right)\right]&=\frac{1}{m}\sum_{j=1}^{m}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right),\dot{\mu}_t^{(j)}\right\rangle\\
&=\frac{1}{m}\sum_{j=1}^{m}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right),-\nabla\delta\ell_N\left(\rho_t\right)\left(\mu_t^{(j)}\right)\right\rangle\\
&=-\int_{\mathbb{R}^d}\left\langle\nabla\varphi\left(x\right),\nabla\delta\ell_N\left(\rho_t\right)\right\rangle\rho_t\left(\mathrm{d}x\right).
\end{aligned}
$$

这证明了

$$\partial_t \rho_t = \operatorname{div}\left(\rho_t \nabla \delta \ell\left(\rho_t\right)\right)$$

以分布的意义成立。

E.3 Wasserstein-Fisher-Rao 梯度流（定理 3 的证明）

本节将证明常微分方程组（3.8）

$$\begin{aligned}
\dot{\mu}_t^{(j)} &= \frac{1}{N}\sum_{i=1}^{N} \frac{\phi\left(X_i - \mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i - \mu_t^{(l)}\right)}\left(X_i - \mu_t^{(j)}\right), \\
\dot{\omega}_t^{(j)} &= \left[\frac{1}{N}\sum_{i=1}^{N} \frac{\phi\left(X_i - \mu_t^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(X_i - \mu_t^{(l)}\right)} - 1\right]\omega_t^{(j)},
\end{aligned}$$

存在唯一解，并且 $(\rho_t)_{t\geq 0}$ 其中 $\rho_t := \sum_{l=1}^{m}\omega_t^{(l)}\delta_{\mu_t^{(l)}}$ 是一个满足（3.6）意义下的 Wasserstein-Fisher-Rao 梯度流。我们将综合前两节所用的证明技巧。

我们将再次使用皮卡-林德洛夫定理证明解的存在唯一性。对于每个 $j \in [m]$，定义两个函数 $f^{(j)}: \Delta^{m-1}\times\mathbb{R}^{md}\to\mathbb{R}^d$ 和 $g^{(j)}: \Delta^{m-1}\times\mathbb{R}^{md}\to\mathbb{R}$，如下所示

$$\begin{aligned}
f^{(j)}(y,z) &= \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i - z_j\right)}{\sum_{l=1}^{m} y_l\phi\left(X_i - z_l\right)}\left(X_i - z_j\right), \\
g^{(j)}(y,z) &= -\left[1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i - z_j\right)}{\sum_{l=1}^{m} y_l\phi\left(X_i - z_l\right)}\right] y_j,
\end{aligned}$$

其中 $y = [y_j]_{1\leq j\leq m} \in \Delta^{m-1}$, $z = [z_j]_{1\leq j\leq m} \in \mathbb{R}^{md}$ 满足 $z_1, \ldots, z_m \in \mathbb{R}^d$。令

$$f(y,z) := \begin{bmatrix} f^{(1)}(y,z) \\ \vdots \\ f^{(m)}(y,z) \end{bmatrix}, \quad g(y,z) := \begin{bmatrix} g^{(1)}(y,z) \\ \vdots \\ g^{(m)}(y,z) \end{bmatrix}, \quad \text{and} \quad h(y,z) = \begin{bmatrix} f(y,z) \\ g(y,z) \end{bmatrix}.$$

于是，我们可以将常微分方程组写为

$$\begin{bmatrix} \dot{\mu}_t \\ \dot{\omega}_t \end{bmatrix} = h\left(\begin{bmatrix} \mu_t \\ \omega_t \end{bmatrix}\right),$$

其中 $\mu_t = [\mu_t^{(j)}]_{1\leq j\leq m}$ 和 $\omega_t = [\omega_t^{(j)}]_{1\leq j\leq m}$。记 $f^{(j)}(z) = [f_k^{(j)}(z)]_{1\leq k\leq d}$。

对于任意满足 $\max_{j\in[m]}\|z_j\|_2 \leq 2R$ 的 $z\in\mathbb{R}^{md}$，以及满足 $\|y-\omega_0\|_2 \leq \varepsilon$ 的 $y\in\mathbb{R}^m$，其中

$$R := \max_{1\leq i\leq N}\|X_i\|_2, \qquad \varepsilon := \frac{(2\pi)^{d/2}}{2m}\exp\left(-\frac{9}{2}R^2\right),$$

对于任意 $i\in[N]$

$$\begin{aligned}
\sum_{l=1}^{m} y_l\phi\left(X_i - z_l\right) &\geq \sum_{l=1}^{m}\omega_0^{(l)}\phi\left(X_i - z_l\right) - \sum_{l=1}^{m}\left(\omega_0^{(l)} - y_l\right)\phi\left(X_i - z_l\right) \\
&\overset{\text{(i)}}{\geq} \min_{i\in[N],l\in[m]}\phi\left(X_i - z_l\right) - \|y-\omega_0\|_2\sum_{l=1}^{m}\phi^2\left(X_i - \mu^{(l)}\right) \\
&\overset{\text{(ii)}}{\geq} \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{9}{2}R^2\right) - \frac{\varepsilon m}{(2\pi)^d} = \frac{1}{2(2\pi)^{d/2}}\exp\left(-\frac{9}{2}R^2\right) \triangleq \delta.
\end{aligned} \tag{E.5}$$

这里，(i) 推出自 $\omega_0 \in \Delta^{m-1}$ 和柯西-施瓦茨不等式，而 (ii) 和 (iii) 成立是因为 $\|X_i - z_l\|_2 \leq 3R$对于任意 $i \in [N]$和 $l \in [m]$。对于任意 $j \in [m]$，记 $f^{(j)}(z) = [f_k^{(j)}(z)]_{1 \leq k \leq d}$。与 (E.4) 的证明类似，我们可以利用 (E.5) 证明

$$\left\|\nabla_z f_k^{(j)}(y,z)\right\|_2 \leq \sqrt{(m+1)\left(\frac{9m^{-1}}{\delta^2(2\pi)^d}R^2\right)^2 + 2\frac{1}{\delta^2(2\pi)^d}}.$$

我们还有

$$\begin{aligned}\nabla_y f_k^{(j)}(y,z) &= \nabla_y \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)} e_k^\top (X_i - z_j) \\ &= -\frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2} e_k^\top (X_i - z_j) \begin{bmatrix} \phi(X_i - z_1) \\ \vdots \\ \phi(X_i - z_m) \end{bmatrix},\end{aligned}$$

因此

$$\left\|\nabla_y f_k^{(j)}(y,z)\right\|_2 \leq \frac{\sqrt{m}\|\phi\|_\infty^2}{\delta^2} \max_{i\in[N], j\in[m]} \|X_i - z_j\|_2 \leq \frac{3\sqrt{m}R}{\delta^2(2\pi)^d}.$$

此外，我们可以计算

$$\begin{aligned}\nabla_y g^{(j)}(y,z) &= -\nabla_y \left[\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right) y_j\right] \\ &= -\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right] e_j + \frac{1}{N}\sum_{i=1}^N \frac{y_j \phi(X_i - z_j)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2} \begin{bmatrix} \phi(X_i - z_1) \\ \vdots \\ \phi(X_i - z_m) \end{bmatrix}\end{aligned}$$

且对于每个 $l \in [m]$

$$\begin{aligned}\nabla_{z_l} g^{(j)}(y,z) &= -\nabla_{z_l} \left[\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right) y_j\right] \\ &= y_j \frac{1}{N}\sum_{i=1}^N \left[\frac{\nabla_{z_l}\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)} - \frac{\phi(X_i - z_j) y_l \nabla_{z_l}\phi(X_i - z_l)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2}\right] \\ &= \frac{1}{N}\sum_{i=1}^N \left[\mathbb{1}\{l = j\} \frac{y_j \phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}(X_i - z_j) - \frac{y_j \phi(X_i - z_j) y_l \phi(X_i - z_l)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2}(X_i - z_l)\right].\end{aligned}$$

因此，我们有

$$\begin{aligned}\left\|\nabla_y g^{(j)}(y,z)\right\|_2 &\leq \left|1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right| + \frac{1}{N}\sum_{i=1}^N \left|\frac{y_j \phi(X_i - z_j)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2}\right| \sqrt{m}\|\phi\|_\infty \\ &\leq 1 + \frac{\|\phi\|_\infty}{\delta} + \frac{(1+\varepsilon)\sqrt{m}}{\delta^2}\|\phi\|_\infty^2 = 1 + \frac{1}{(2\pi)^{d/2}\delta} + \frac{(1+\varepsilon)\sqrt{m}}{(2\pi)^d \delta^2}\end{aligned}$$

and

$$\begin{aligned}\left\|\nabla_{z_l} g^{(j)}(y,z)\right\|_2 &\leq \frac{1}{N}\sum_{i=1}^N \frac{(1+\varepsilon)\|\phi\|_\infty}{\delta}\|X_i - z_j\|_2 + \frac{1}{N}\sum_{i=1}^N \frac{(1+\varepsilon)^2\|\phi\|_\infty^2}{\delta^2}\|X_i - z_l\|_2 \\ &\leq \frac{3(1+\varepsilon)R}{(2\pi)^{d/2}\delta} + \frac{3(1+\varepsilon)^2 R\|\phi\|_\infty^2}{(2\pi)^d \delta^2}.\end{aligned}$$

因此，对于任意满足 $\max_{j\in[m]}\|z_j\|_2 \le 2R$ 且$\|y-\omega_0\|_2\le\varepsilon$的 $k\in[d]$、$j\in[m]$、$z\in\mathbb{R}^{md}$，以及 $y\in\mathbb{R}^m$，我们有

$$\begin{aligned}\left\|\nabla f_k^{(j)}(y,z)\right\|_2 &= \sqrt{\left\|\nabla_z f_k^{(j)}(y,z)\right\|_2^2+\left\|\nabla_y f_k^{(j)}(y,z)\right\|_2^2}\\ &\le \sqrt{(m+1)\left(\frac{9m^{-1}}{\delta^2(2\pi)^d}R^2\right)^2+\frac{2}{\delta^2(2\pi)^d}+\left(\frac{3\sqrt{m}R}{\delta^2(2\pi)^d}\right)^2}\triangleq C_{\mathsf{lip},f},\end{aligned}$$

这表明 $f^{(j)}(y,z)$ 是 $\sqrt{d}C_{\mathsf{lip},f}$-利普希茨连续的，并且

$$\begin{aligned}\left\|\nabla g^{(j)}(y,z)\right\|_2 &= \sqrt{\left\|\nabla_y g^{(j)}(y,z)\right\|_2^2+\sum_{l=1}^m\left\|\nabla_{z_l} g^{(j)}(y,z)\right\|_2^2}\\ &\le \sqrt{\left(1+\frac{1}{(2\pi)^{d/2}\delta}+\frac{(1+\varepsilon)\sqrt{m}}{(2\pi)^d\delta^2}\right)^2+m\left(\frac{3(1+\varepsilon)R}{(2\pi)^{d/2}\delta}+\frac{3(1+\varepsilon)^2R\|\phi\|_\infty^2}{(2\pi)^d\delta^2}\right)^2}\triangleq C_{\mathsf{lip},g}.\end{aligned}$$

这表明 $g^{(j)}(y,z)$ 是 $C_{\mathsf{lip},g}$-利普希茨连续的。因此，我们可以得出结论，$h(y,z)$ 关于 $\{(y,z)$ 是 C_{Lip}-连续的：$\|y-\omega_0\|_2\le\varepsilon,\max_{j\in[m]}\|z_j\|_2\le 2R\}$，其中 $C_{\mathsf{Lip}}=\sqrt{mC_{\mathsf{lip},g}^2+mdC_{\mathsf{lip},f}^2}$。此外，容易验证，对于任意满足 $\max_{j\in[m]}\|z_j\|_2\le 2R$ 的 $z\in\mathbb{R}^{md}$，以及任意满足$\|y-\omega_0\|_2\le\varepsilon$的 $y\in\mathbb{R}^m$，

$$\max_{j\in[m]}\left\|f^{(j)}(y,z)\right\|_2\le\frac{3R}{\delta(2\pi)^{d/2}},\quad \max_{j\in[m]}\left|g^{(j)}(y,z)\right|\le\left(1+\frac{1}{\delta(2\pi)^{d/2}}\right)(1+\varepsilon)$$

因此

$$\|h(y,z)\|_2\le\sqrt{m\left(\frac{3R}{\delta(2\pi)^{d/2}}\right)^2+m\left[\left(1+\frac{1}{\delta(2\pi)^{d/2}}\right)(1+\varepsilon)\right]^2}\triangleq M.$$

回忆一下，$\mu_0^{(1)},\ldots,\mu_0^{(m)}$ 是从均匀分布 Uniform($\{X_i\}_{1\le i\le N}$) 中独立同分布抽样得到的，因此

$$(\omega_0,\mu_0)\in(y,z):\left\{(y,z):\|y-\omega_0\|_2\le\varepsilon,\max_{j\in[m]}\|z_j\|_2\le 2R\right\}$$

其中 $\mu_0=[\mu_0^{(j)}]_{1\le j\le m}$。我们现在可以应用皮卡-林德洛夫定理，证明存在一个仅依赖于 C_{Lip}和 M的 $t_0>0$，使得该常微分方程在时间区间 $[0,t_0]$上有唯一解。我们可以使用附录E.1中定理5证明里的相同论证，证明对于所有 $t\in[0,t_0]$，都有 $\omega_t\in\Delta^{m-1}$；也可以使用附录 E.2 中定理 6 证明里的相同论证，证明对于所有 $t\in[0,t_0]$，都有 $\max_{j\in[m]}\|\mu_t^{(j)}\|_2\le R$。随后，我们可以重复上述相同的分析（注意，上述证明只需要 $\omega_0\in\Delta^{m-1}$ 和 $\max_{j\in[m]}\|\mu_0^{(j)}\|_2\le R$，而 t_0 仅依赖于普适常数 C_{Lip} 和 M），从而证明该常微分方程在 $[t_0,2t_0]$、$[2t_0,3t_0]$ 等区间上有唯一解。这证明了常微分方程组 (3.8) 的解的存在唯一性。

最后，我们验证 $(\rho_t)_{t\ge0}$ 定义为 $\rho_t:=\sum_{l=1}^m\omega_t^{(l)}\delta_{\mu_t^{(l)}}$ 是（3.11）的解。注意，ρ_t 是一个概率测度，因为我们已经证明，对于任意 $\omega_t\in\Delta^{m-1}$，$t\ge0$。对于任意测试函数 $\varphi(x)\in C_c^\infty$，有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(x)\rho_t(\mathrm{d}x)&=\frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{j=1}^m\omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right)\right]=\sum_{j=1}^m\left[\dot\omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right)+\omega_t^{(j)}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right),\dot\mu_t^{(j)}\right\rangle\right]\\&=-\sum_{j=1}^m\left[1+\delta\ell_N(\rho_t)\left(\mu_t^{(j)}\right)\right]\omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right)+\sum_{j=1}^m\omega_t^{(j)}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right),-\nabla\delta\ell_N(\rho_t)\left(\mu_t^{(j)}\right)\right\rangle\end{aligned}$$

$$= -\int_{\mathbb{R}^d} \left[1 + \delta \ell_N\left(\rho_t\right)(x)\right] \varphi(x) \rho_t(\mathrm{d}x) - \int_{\mathbb{R}^d} \left\langle \nabla \varphi(x), \nabla \delta \ell_N\left(\rho_t\right)\right\rangle \rho_t(\mathrm{d}x).$$

这证明了

$$\partial_t \rho_t = \operatorname{div}\left(\rho_t \nabla \delta \ell\left(\rho_t\right)\right) - \left[\delta \ell\left(\rho_t\right) + 1\right] \rho_t$$

在分布的意义下成立。

F Wasserstein梯度流的性质

本节将介绍一些关于用于学习高斯混合分布的 Wasserstein梯度流的初步结果。我们还将讨论这些结果的意义，以及获得更一般结果所面临的技术难点。

我们首先建立 Wasserstein梯度流与欧几里得空间中的经典梯度流之间的联系。假设我们使用 $\{X_i\}_{1\le i\le N}$ 拟合一个由 m个分量组成的高斯混合模型

$$\frac{1}{m}\sum_{j=1}^{m} \mathcal{N}\left(\mu^{(j)}, I_d\right),$$

其中， $\{\mu^{(j)}\}_{1\le j\le m}$ 是 m 个高斯分量的位置。负似然函数为

$$\ell_{N,m}\left(\mu^{(1)}, \ldots, \mu^{(m)}\right) := -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{1}{m}\sum_{j=1}^{m} \phi\left(X_i - \mu^{(j)}\right)\right]. \tag{F.1}$$

用于最小化（F.1）的梯度流，记为 $(\mu_t)_{t\ge 0}$ 其中 $\mu_t = [\mu_t^{(j)}]_{1\le j\le m}$，由以下常微分方程组给出

$$\dot{\mu}_t^{(j)} = -\nabla_{\mu^{(j)}} \ell_{N,m}\left(\mu_t^{(1)}, \ldots, \mu_t^{(m)}\right) \tag{F.2}$$

以 $\mu_0^{(1)}, \ldots, \mu_0^{(m)} \overset{i.i.d.}{\sim}$均匀分布$(\{X_i\}_{1\le i\le N})$为初始化。以下定理表明，梯度流（F.2）刻画了由离散分布 $\frac{1}{m}\sum_{l=1}^{m} \delta_{\mu_0^{(l)}}$ 初始化的 Wasserstein梯度流（3.18）中粒子位置的演化。证明见附录F.1。

定理 8。 考虑欧几里得梯度流$(\mu_t)_{t\ge 0}$，其中*F.2*。那么，流$(\rho_t)_{t\ge 0}$ 定义为 s

$$\rho_t := \frac{1}{m}\sum_{l=1}^{m} \delta_{\mu_t^{(l)}} \tag{F.3}$$

即为 *Wasserstein*梯度流，也就是说，（*F.3*）是偏微分方程（*3.18*）的分布解。

对于最小化（F.1）的梯度下降算法与粒子 Wasserstein梯度下降之间，也可以建立类似的联系（参见算法3），此处为简洁起见不再赘述。

接下来，我们关注 Wasserstein 梯度流的无限样本极限，并分析其收敛性质。总体损失函数为

$$\ell_\infty(\rho) = -\mathbb{E}_{X\sim\rho^\star * \phi}\left\{\log\left[\rho * \phi(X)\right]\right\} = \mathrm{KL}\left(\rho^\star * \phi \,\|\, \rho * \phi\right) + \text{const.} \tag{F.4}$$

在附录中C.1 我们已计算得出

$$\delta \ell_\infty(\rho) = -\int_{\mathbb{R}^d} \frac{\rho^\star * \phi(y)}{\rho * \phi(y)} \phi(x - y)\, \mathrm{d}y. \tag{F.5}$$

我们知道 W asserstein梯度流 $(\rho_t)_{t\ge 0}$ 关于 $\ell_\infty(\rho)$ 由以下偏微分方程描述 :

$$\partial_t \rho_t = \operatorname{div}\left(\rho_t \nabla \delta \ell_\infty\left(\rho_t\right)\right) \tag{F.6}$$

其中 $\rho_0 = \rho^\star * \mathcal{N}(0, I_d)$ 是无限样本情况下的数据分布。这个Wasserstein梯度流有如下粒子解释：假设在时间 $t = 0$，我们将粒子 $x_0 \sim \rho_0$初始化于向量场 $(v_t)_{t\geq 0}$ 中，其中 $v_t = -\nabla \delta \ell_\infty(\rho_t)$，也就是说

$$\dot{x}_t = v_t\left(x_t\right),$$

即 $x_t \sim \rho_t$，也就是说 $(x_t)_{t\geq 0}$ 的边缘分布按照 Wasserstein梯度流演化
。下面的定理表明，当真实的混合分布 $\rho^\star$ 是一个单点集时（不失一般性地假设 $\rho^\star = \delta_0$），Wasserstein梯度流收敛于 $\rho^\star$。证明见附录F.2。

定理 9。 考虑 (F.6) 中的 *Wasserstein*梯度流，其中 $\rho^\star = \delta_0$。对于任意 $\varepsilon < 1$，有$\int_{\mathbb{R}^d} \|x\|_2^2 \rho_t(\mathrm{d}x) = O(\varepsilon)$ 只要

$$t \geq \exp\left(2d\right) \varepsilon^{-1-\max\left\{8, \sqrt{8d}\right\}}.$$

尽管定理 9 仅关注 $\rho^\star$ 为单点集的情形，但该收敛结果已经为我们理解更一般情形下 Wasserstein梯度流的行为提供了一些直观认识。考虑一个具有 K 个分量的充分分离的高斯混合模型。假设混合分布为 $\rho^\star = \sum_{j=1}^K \omega_j^\star \delta_{\mu_j^\star}$，且每个高斯分量的位置 $\{\mu_j^\star\}_{1\leq j\leq K}$彼此充分分离。由于推前映射 $v_t = -\nabla \delta \ell_\infty(\rho_t)$ 是局部化的（见 F.5），因此存在某个 $T > 0$，使得从 $\rho^\star * \mathcal{N}(0, I_d)$ 初始化的 Wasserstein梯度流（F.6）在时间 T之前可以近似为

$$\rho_t \approx \sum_{j=1}^K \omega_j^\star \rho_t^{(j)} \qquad \forall\, t \in [0, T],$$

其中，对于每个 $j \in [K]$, $\rho_t^{(j)}$，Wasserstein梯度流为 $\partial_t \rho_t^{(j)} = \mathsf{div}(\rho_t^{(j)} \nabla \delta \ell_\infty(\rho_t^{(j)}))$，其初始化为 $\rho_0^{(j)} = \mathcal{N}(\mu_j^\star, I_d)$。这表明，Wasserstein梯度流近似收敛于 $\rho^\star$，因为根据定理 9，每个 $\rho_t^{(j)}$ 都会收敛于 $\delta_{\mu_j^\star}$。然而，这一观察也表明，Wasserstein梯度流对权重不匹配并不稳健。考虑使用 $\rho_0 = \widetilde{\rho} * \mathcal{N}(0, 1)$ 初始化 Wasserstein梯度流 (F.6)，其中 $\widetilde{\rho} = \sum_{j=1}^K \widetilde{\omega}_j \delta_{\mu_j^\star}$是一个支撑集 $\{\mu_j^\star\}_{1\leq j\leq K}$正确但权重 $\{\widetilde{\omega}_j\}_{1\leq j\leq K} \neq \{\omega_j^\star\}_{1\leq j\leq K}$错误的混合分布。于是我们还有

$$\rho_t \approx \sum_{j=1}^K \widetilde{\omega}_j \rho_t^{(j)} \qquad \forall\, t \in [0, T],$$

这表明，当 $0 \leq t \leq T$ 时，ρ_t 近似收敛到 $\widetilde{\rho}$，而非 $\rho^\star$。请注意，只要分离度 $\min_{i\neq j} \|\mu_i^\star - \mu_j^\star\|_2 \to \infty$ 满足条件，这类近似有效的时间长度 T 就可以任意长。上述讨论表明，使用正确的初始权重对于Wasserstein梯度流收敛到真实混合分布至关重要。

我们还希望将定理 9中的收敛速率与 Bures-Wasserstein 梯度流所提供的基准进行比较。Bures-Wasserstein 梯度流定义在 $\mathbb{R}^d$ 上的非退化高斯分布空间中，记作 $\mathsf{BW}(\mathbb{R}^d) = \mathbb{R}^d \times \mathbb{S}_{++}^d$（其中，我们将非退化高斯分布 $\nu = \mathcal{N}(\mu, \Sigma)$ 与 $(\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{S}_{++}^d$对应起来），并配备 Wasserstein 距离（3.3），其闭式表达如下

$$d_{\mathsf{W}}^2\left(\nu_1, \nu_2\right) = \|\mu_1 - \mu_2\|_2^2 + \mathsf{tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}\right)^{1/2}\right]$$

当 $\nu_1 = \mathcal{N}(\mu_1, \Sigma_1)$ 和 $\nu_2 = \mathcal{N}(\mu_2, \Sigma_2)$ 均为非退化高斯分布时。Bures-Wasserstein 梯度流 $(\nu_t)_{t\geq 0}$ 可视为受限于 $\mathsf{BW}(\mathbb{R}^d)$ 的 Wasserstein梯度流 $(\rho_t)_{t\geq 0}$。有兴趣的读者可参阅 Altschuler等人（2021）；Lambert等人（2022），了解更详细的讨论。从定理 9 的证明中可以看出，Wasserstein梯度流的推前映射 $v_t(x)$ 以指数速度衰减至 $\|x\|_2 \to \infty$，这将使 Wasserstein梯度流 $(\rho_t)_{t\geq 0}$ 变得越来越重尾。然而，Bures-Wasserstein 梯度流的推前映射始终

是线性的，且 Bures-Wasserstein 梯度流 $(\nu_t)_{t\geq 0}$ 始终是高斯分布。例如，在附录F.3中，我们可以在 $t=0$ 处显式计算这两种梯度流的推前映射：

$$v_0(x) = -\frac{1}{3}\left(\frac{4}{3}\right)^{d/2}\exp\left(-\frac{\|x\|_2^2}{6}\right)x \qquad \text{（Wasserstein），}$$

$$v_0(x) = \frac{x}{4} \qquad \text{（Bures-Wasserstein）。}$$

因此，很自然可以预期，Bures-Wasserstein 梯度流 $(\nu_t)_{t\geq 0}$ 从 $\nu_0=\mathcal{N}(0, I_d)$ 初始化后，比 Wasserstein梯度流 $(\rho_t)_{t\geq 0}$ 从 $\rho_0=\delta_0$ 初始化后收敛得更快。在附录 F.3 中，我们证明 Bures-Wasserstein 梯度流 $(\nu_t=\mathcal{N}(\mu_t, \Sigma_t))_{t\geq 0}$ 由以下常微分方程刻画：

$$\begin{aligned}\dot{\mu}_t &= 0\\ \dot{\Sigma}_t &= -2\left(\Sigma_t+I_d\right)^{-1}\Sigma_t^2\left(\Sigma_t+I_d\right)^{-1}.\end{aligned}$$

我们还证明，Σ_t 介于

$$\frac{1}{1+2t}I \preceq \Sigma_t \preceq \frac{2}{2+t}I,$$

并且因此 $\int_{\mathbb{R}^d}\|x\|_2^2\nu_t(\mathrm{d}x) = O(d/t)$。由于 Bures-Wasserstein 梯度流并非以指数速度收敛（我们可以看到，其收敛速率关于 t 是多项式的），我们猜想 Wasserstein梯度流同样不具有指数收敛性质。

最后，我们在图 6 中通过数值实验展示，损失函数 $\ell_\infty(\rho)$ （参见（F.4））即使在 $\rho^\star=\delta_0$时也不是测地凸的（Ambrosio 等人，2008）。我们还可以验证，Polyak-Łojasiewicz（PL）不等式

$$\forall\rho: \quad \|\nabla_{\mathsf{W}}\ell_\infty(\rho)\|_\rho^2 \geq C_{\mathsf{PL}}\left[\ell_\infty(\rho)-\ell_\infty(\rho^\star)\right] \quad \text{for some } C_{\mathsf{PL}}>0$$

一般情况下并不成立：考虑 $\rho^\star=\frac{1}{2}\delta_{-1}+\frac{1}{2}\delta_1$ 和 $\rho=\delta_0$，则不难验证$\nabla_{\mathsf{W}}\ell_\infty(\rho)=0$，但 $\ell_\infty(\rho)>\ell_\infty(\rho^\star)$。因此，当损失函数是测地凸时，我们无法使用标准证明技巧（例如 Ambrosio 等人(2008)，或在满足 PL 不等式时使用 Chewi 等人 (2020年)，来证明 Wasserstein梯度流（F.6）的指数收敛。

F.1 定理8的证明

不难计算出 $\ell_{N,m}$ 的梯度。对于任意 $j\in[m]$，有

$$\begin{aligned}\nabla_{\mu^{(j)}}\ell_{N,m}\left(\mu^{(1)},\ldots,\mu^{(m)}\right) &= -\frac{1}{N}\sum_{i=1}^N\frac{1}{\sum_{l=1}^m\phi\left(X_i-\mu^{(l)}\right)}\nabla_{\mu^{(j)}}\phi\left(X_i-\mu^{(j)}\right)\\ &= -\frac{1}{N}\sum_{i=1}^N\frac{\phi\left(X_i-\mu^{(l)}\right)}{\sum_{l=1}^m\phi\left(X_i-\mu^{(l)}\right)}\left(X_i-\mu^{(l)}\right).\end{aligned}$$

因此，欧几里得梯度流（F.2）为

$$\dot{\mu}_t^{(j)} = -\nabla_{\mu^{(j)}}\ell_{N,m}\left(\mu_t^{(1)},\ldots,\mu_t^{(m)}\right) = \frac{1}{N}\sum_{i=1}^N\frac{\phi\left(X_i-\mu_t^{(l)}\right)}{\sum_{l=1}^m\phi\left(X_i-\mu_t^{(l)}\right)}\left(X_i-\mu_t^{(l)}\right).$$

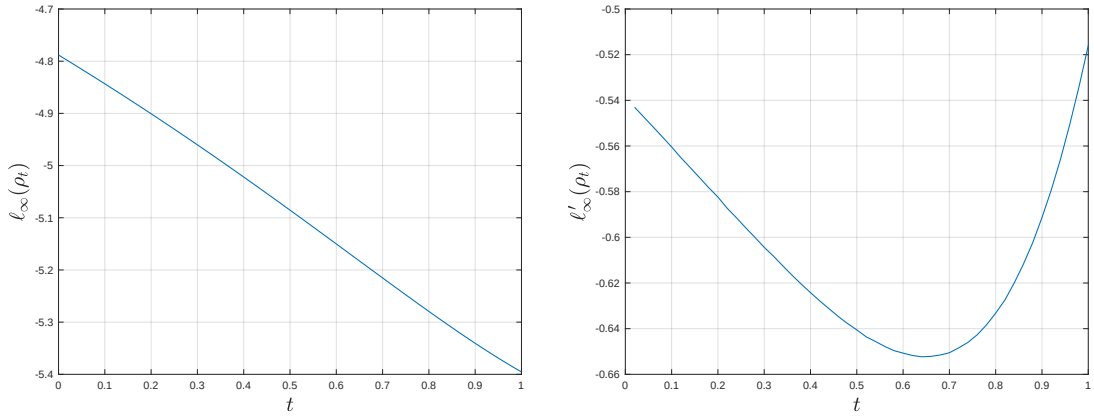
然后我们可以应用定理 6 完成证明。

F.2 定理9 的证明

步骤 1：刻画推前映射。首先，很容易验证，对于所有 $t\geq 0$，$(\rho_t)_{t\geq 0}$ 都是球对称的，也就是说，$\rho_t(\mathrm{d}x)$ 仅取决于 $\|x\|_2$。时刻 t 的推前映射 $v_t(x):\mathbb{R}^d\to\mathbb{R}^d$ 为

$$v_t(x) = -\nabla\delta\ell_\infty(\rho_t)(x) = \nabla_x\int_{\mathbb{R}^d}\frac{\rho^\star*\phi(y)}{\rho_t*\phi(y)}\phi(x-y)\,\mathrm{d}y$$

(a) 常速测地线 $(\rho_t)_{0\le t\le 1}$ 连接 $\rho_0=\mathcal{N}(0,\ 3)$ 与 $\rho_1=\mathcal{N}(0,\ 1)$



(b) 常速测地线 $(\rho_t)_{0\le t\le 1}$ 连接 $\rho_0=\mathcal{N}(0,\ 1)$ 与 $\rho_1=\delta_0$

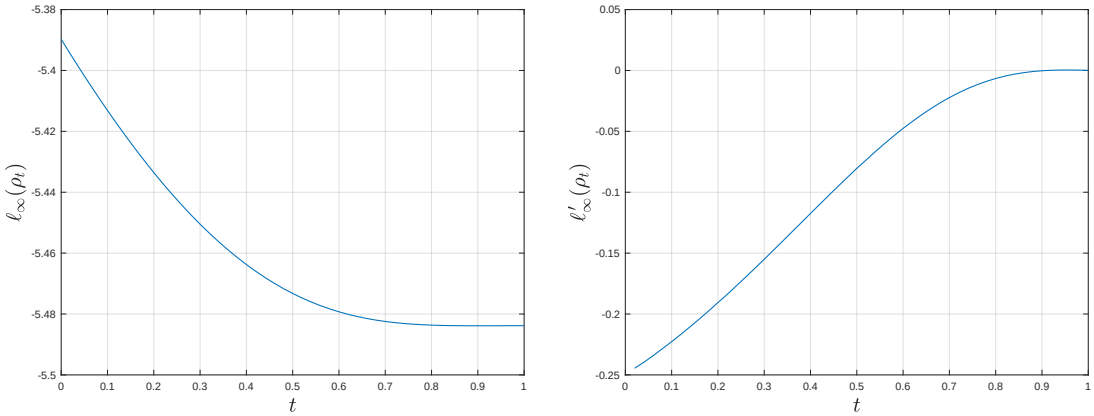


图 6：损失函数 $\ell_\infty(\rho_t)$ 或其导数 $\ell'_\infty(\rho_t)$ 与 t 的关系。在图 (a) 中，$(\rho_t)_{0\le t\le 1}$ 是连接 $\rho_0=\mathcal{N}(0,\ 3)$ 与 $\rho_1=\mathcal{N}(0,\ 1)$ 的常速测地线。在图 (b) 中，$(\rho_t)_{0\le t\le 1}$ 是连接 $\rho_0=\mathcal{N}(0,\ 1)$ 与 $\rho_1=\delta_0$ 的常速测地线。这表明 $\ell_\infty(\rho)$ 并非全局测地凸，但可能在 $\rho^\star=\delta_0$ 附近局部测地凸。

$$
\begin{aligned}
&= -\int_{\mathbb{R}^d} \frac{\rho^\star * \phi(y)}{\rho_t * \phi(y)} (x-y)\,\phi(x-y)\,\mathrm{d}y \\
&= \int_{\mathbb{R}^d} \nabla_y \frac{\rho^\star * \phi(y)}{\rho_t * \phi(y)} \phi(y-x)\,\mathrm{d}y \\
&= \int_{\mathbb{R}^d} h_t(y)\,\phi(y-x)\,\mathrm{d}y,
\end{aligned} \tag{F.7}
$$

其中倒数第二行由斯坦引理或高斯分部积分得到，最后一行中的 $h_t(y)$ 定义为

$$
\begin{aligned}
h_t(y) &\coloneqq \nabla \frac{\phi(y)}{\rho_t * \phi(y)} = \nabla_y \frac{1}{\int \exp\left(-\frac{1}{2}\|z\|_2^2 + y^\top z\right) \rho_t(\mathrm{d}z)} \\
&= -\frac{\int \exp\left(-\frac{1}{2}\|z\|_2^2 + y^\top z\right) z \rho_t(\mathrm{d}z)}{\left[\int \exp\left(-\frac{1}{2}\|z\|_2^2 + y^\top z\right) \rho_t(\mathrm{d}z)\right]^2} = -\phi(y) \frac{\int \phi(y-z)\, z \rho_t(\mathrm{d}z)}{\left[\int \phi(y-z)\,\rho_t(\mathrm{d}z)\right]^2} \\
&= -\phi(y) \frac{\int \phi(y-z)\, z \rho_t(\mathrm{d}z)}{[\rho_t * \phi(y)]^2} = -\frac{\phi(y)}{\rho_t * \phi(y)} \cdot \frac{\int \phi(y-z)\, z \rho_t(\mathrm{d}z)}{\rho_t * \phi(y)}.
\end{aligned} \tag{F.8}
$$

对于任意 $y \in \mathbb{R}^d$，我们可以计算

$$\begin{aligned}
\int \phi(y-z)\, z \rho_t(\mathrm{d}z) &= \int_{y^\top z>0} \phi(y-z)\, z \rho_t(\mathrm{d}z) + \int_{y^\top z<0} \phi(y-z)\, z \rho_t(\mathrm{d}z) + \int_{y^\top z=0} \phi(y-z)\, z \rho_t(\mathrm{d}z) \\
&\overset{\text{(i)}}{=} \int_{y^\top z>0} [\phi(y-z) - \phi(y+z)]\, z \rho_t(\mathrm{d}z) \\
&= \phi(y) \int_{y^\top z>0} \left[\exp\left(y^\top z\right) - \exp\left(-y^\top z\right)\right] z \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) \\
&\overset{\text{(ii)}}{=} \phi(y) \int_{y^\top z>0} \left[\exp\left(y^\top z\right) - \exp\left(-y^\top z\right)\right] \frac{y^\top z}{\|y\|_2^2} y \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) \\
&= \underbrace{\int_{y^\top z>0} \left[\exp\left(y^\top z\right) - \exp\left(-y^\top z\right)\right] \frac{y^\top z}{\|y\|_2^2} \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z)}_{=:a_t} \cdot \phi(y)\, y. \qquad \text{(F.9)}
\end{aligned}$$

这里，(i) 和 (ii) 均由 ρ_t 的球对称性推出。容易验证，由于 ρ_t 的球对称性，最后一行中的积分与 y 无关，因此 a_t 是一个与 y 无关的普适常数。注意，当 $y^\top z > 0$ 时，我们有 $\exp(y^\top z) - \exp(-y^\top z) \geq 2y^\top z$，因此

$$a_t \geq 2 \int_{y^\top z>0} \frac{\left(y^\top z\right)^2}{\|y\|_2^2} \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) = \int_{\mathbb{R}^d} \frac{\left(y^\top z\right)^2}{\|y\|_2^2} \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z).$$

由于 a_t 与 y 无关，我们取 $y = e_i$ 作为 $i \in [d]$，从而得到

$$a_t \geq \int_{\mathbb{R}^d} z_i^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z), \qquad \forall i \in [d].$$

对 d 取平均后，我们得到

$$a_t \geq \frac{1}{d} \int_{\mathbb{R}^d} \|z\|_2^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) = \frac{m_t}{d}, \qquad \text{(F.10)}$$

其中我们定义

$$m_t \coloneqq \int \|z\|_2^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z).$$

将 (F.8)、(F.9) 和 (F.10) 合并可得

$$h_t(y) = -a_t y \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2, \qquad \text{where} \qquad a_t \geq \frac{m_t}{d}.$$

然后，我们利用（F.7）来刻画推前映射：

$$\begin{aligned}
v_t(x) &= \int_{y^\top x>0} h(y)\, \phi(y-x)\, \mathrm{d}y + \int_{y^\top x<0} h(y)\, \phi(y-x)\, \mathrm{d}y + \int_{y^\top x=0} h(y)\, \phi(y-x)\, \mathrm{d}y \\
&\overset{\text{(i)}}{=} \int_{y^\top x>0} h(y) \left[\phi(y-x) - \phi(-y-x)\right] \mathrm{d}y \\
&= -\phi(x)\, a_t \int_{y^\top x>0} y \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \left[\exp\left(y^\top x\right) - \exp\left(-y^\top x\right)\right] \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y \\
&\overset{\text{(ii)}}{=} -\phi(x)\, a_t \int_{y^\top x>0} \frac{x^\top y}{\|x\|_2^2} x \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \left[\exp\left(y^\top x\right) - \exp\left(-y^\top x\right)\right] \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y \\
&= -a_t \underbrace{\int_{y^\top x>0} \frac{x^\top y}{\|x\|_2^2} \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \left[\exp\left(y^\top x\right) - \exp\left(-y^\top x\right)\right] \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y}_{=:b_t} \cdot \phi(x)\, x.
\end{aligned}$$

与（F.9）类似，此处（i）和（ii）均可由 ρ_t的球对称性推出；最后一行中的积分不依赖于 x ，这是由于 ρ_t具有球对称性。因此， b_t 是一个与 x无关的普适常数。注意，当 $y^\top x > 0$时，我们有$\exp(y^\top x) - \exp(-y^\top x) \geq 2y^\top x$，因此

$$\begin{aligned} b_t &\geq 2\int_{y^\top x>0} \frac{x^\top y}{\|x\|_2^2}\left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 y^\top x \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y \\ &= \int_{\mathbb{R}^d} \frac{\left(x^\top y\right)^2}{\|x\|_2^2}\left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y. \end{aligned}$$

注意到 $\|\rho_t * \phi\|_\infty \leq \|\phi\|_\infty \leq (2\pi)^{-d/2}$，因此

$$b_t \geq \int \frac{\left(x^\top y\right)^2}{\|x\|_2^2} \exp\left(-\frac{3}{2}\|y\|_2^2\right) \mathrm{d}y = \frac{1}{3}\left(\frac{2\pi}{3}\right)^{d/2}$$

因此我们有

$$v_t(x) = -a_t b_t \phi(x)\, x = -c_t \phi(x)\, x \tag{F.11}$$

其中

$$c_t \coloneqq a_t b_t \geq \frac{1}{3d}\left(\frac{2\pi}{3}\right)^{d/2} m_t. \tag{F.12}$$

步骤 2：证明 Wasserstein 梯度流的收敛性。回顾 Wasserstein 梯度流的粒子解释如下：令 $x_0 \sim \rho_0 = \mathcal{N}(0, I_d)$ 和 $\dot{x}_t = v_t(x_t)$，则对任意 $t \geq 0$，我们有$x_t \sim \rho_t$。由此可以计算

$$\begin{aligned} \partial_t \mathbb{E}\left[\|x_t\|_2^2\right] &= 2\mathbb{E}\left[\langle x_t, \dot{x}_t\rangle\right] = 2\mathbb{E}\left[\langle x_t, v_t(x_t)\rangle\right] \overset{\text{(i)}}{=} -2c_t \mathbb{E}\left[\|x_t\|_2^2 \phi(x_t)\right] \\ &\overset{\text{(ii)}}{\leq} -\frac{2}{3d}\left(\frac{2\pi}{3}\right)^{d/2} m_t \mathbb{E}\left[\|x_t\|_2^2 \phi(x_t)\right] \\ &\overset{\text{(iii)}}{=} -\frac{2}{3d}\left(\frac{4\pi^2}{3}\right)^{d/2} \mathbb{E}^2\left[\|x_t\|_2^2 \phi(x_t)\right], \end{aligned} \tag{F.13}$$

其中 (i) 根据（F.11）得出，(ii) 利用了（F.12），而 (iii) 成立是因为

$$m_t = \int \|z\|_2^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) = (2\pi)^{d/2}\, \mathbb{E}\left[\|x_t\|_2^2 \phi(x_t)\right].$$

对任意 $\tau > 0$，由柯西-施瓦茨不等式可得

$$\mathbb{E}\left[\|x_0\|_2^2 \mathbb{1}\left\{\|x_0\|_2^2 > d + \tau\right\}\right] \overset{\text{(i)}}{\leq} \left[\mathbb{E}\|x_0\|_2^4\right]^{1/2} \left[\mathbb{P}\left(\|x_0\|_2^2 > d + \tau\right)\right]^{1/2}.$$

注意到 $\|x_0\|_2^2 \sim \chi^2(d)$，因此 $\mathbb{E}\|x_0\|_2^4 = \text{方差}(\|x_0\|_2^2) + (\mathbb{E}\|x_0\|_2^2)^2 = 2d + d^2$。此外，根据 χ^2 随机变量的尾概率界（例如 Wainwright (2019年，公式 (2.18)），我们有

$$\mathbb{P}\left(\|x_0\|_2^2 > d + \tau\right) \leq \exp\left(-\min\left\{\frac{\tau^2}{8d}, \frac{\tau}{8}\right\}\right).$$

因此，我们有

$$\begin{aligned} \mathbb{E}\left[\|x_0\|_2^2 \mathbb{1}\left\{\|x_0\|_2^2 > d + \tau\right\}\right] &\leq \sqrt{2d + d^2} \exp\left(-\min\left\{\frac{\tau^2}{8d}, \frac{\tau}{8}\right\}\right) \\ &\leq (d+1)\exp\left(-\min\left\{\frac{\tau^2}{8d}, \frac{\tau}{8}\right\}\right) \leq \varepsilon \end{aligned}$$

只要我们选择

$$\tau \triangleq \max\left\{8\log\frac{d+1}{\varepsilon}, \sqrt{8d\log\frac{d+1}{\varepsilon}}\right\}.$$

由于推前映射 $v_t(x)$始终指向零（参见 (F.11) 和 (F.12)），我们知道 $\|x_t\|_2$ 是关于 t的非增函数。因此我们有

$$\begin{aligned}
\mathbb{E}\left[\|x_t\|_2^2\,\phi(x_t)\right] &\geq \mathbb{E}\left[\|x_t\|_2^2\,\phi(x_t)\,\mathbb{1}\left\{\|x_0\|_2^2 \leq d+\tau\right\}\right] \\
&\overset{\text{(i)}}{\geq} (2\pi)^{-d/2}\exp\left(-\frac{d+\tau}{2}\right)\mathbb{E}\left[\|x_t\|_2^2\,\mathbb{1}\left\{\|x_0\|_2^2 \leq d+\tau\right\}\right] \\
&\geq (2\pi)^{-d/2}\exp\left(-\frac{d+\tau}{2}\right)\left(\mathbb{E}\left[\|x_t\|_2^2\right] - \mathbb{E}\left[\|x_t\|_2^2\,\mathbb{1}\left\{\|x_0\|_2^2 > d+\tau\right\}\right]\right) \\
&\overset{\text{(ii)}}{\geq} (2\pi)^{-d/2}\exp\left(-\frac{d+\tau}{2}\right)\left(\mathbb{E}\left[\|x_t\|_2^2\right] - \mathbb{E}\left[\|x_0\|_2^2\,\mathbb{1}\left\{\|x_0\|_2^2 > d+\tau\right\}\right]\right) \\
&\geq (2\pi)^{-d/2}\exp\left(-\frac{d+\tau}{2}\right)\left(\mathbb{E}\left[\|x_t\|_2^2\right] - \varepsilon\right), \qquad \text{(F.14)}
\end{aligned}$$

其中，(i) 和 (ii) 均由 $\|x_t\|_2$ 是非增函数这一事实得出。综合 (F.13) 和 (F.14) 可得

$$\begin{aligned}
\partial_t\mathbb{E}\left[\|x_t\|_2^2\right] &\leq -\frac{2}{3d}\left(\frac{4\pi^2}{3}\right)^{d/2}(2\pi)^{-d}\exp\left[-(d+\tau)\right]\left(\mathbb{E}\left[\|x_t\|_2^2\right]-\varepsilon\right)^2 \\
&= -\frac{2}{3d}\left(\frac{1}{3}\right)^{d/2}\exp\left[-(d+\tau)\right]\left(\mathbb{E}\left[\|x_t\|_2^2\right]-\varepsilon\right)^2.
\end{aligned}$$

令 $f(t) = \mathbb{E}[\|x_t\|_2^2]$，我们知道 $f(0) = d$且

$$\frac{\mathrm{d}f}{\mathrm{d}t} \leq -\frac{2}{3d}\left(\frac{1}{3}\right)^{d/2}\exp\left[-(d+\tau)\right](f-\varepsilon)^2.$$

解这个常微分不等式可得

$$\frac{1}{f(t)-\varepsilon} - \frac{1}{f(0)-\varepsilon} \geq \frac{2}{3d}\left(\frac{1}{3}\right)^{d/2}\exp\left[-(d+\tau)\right]t,$$

这等价于

$$\mathbb{E}\left[\|x_t\|_2^2\right] \leq \varepsilon + \left\{\frac{2}{3d}\left(\frac{1}{3}\right)^{d/2}\exp(-d-\tau)\,t + \frac{1}{d-\varepsilon}\right\}^{-1}.$$

那么我们立即知道，只要满足

$$t \geq \exp(2d)\,\varepsilon^{-1-\max\left\{8,\sqrt{8d}\right\}}.$$

F.3 Bures-Wasserstein 梯度流的计算

定义 $\ell(\mu,\Sigma) = \ell_\infty(\rho)$，其中我们将 $\rho = \mathcal{N}(\mu,\Sigma)$ 参数化。然后我们可以计算

$$\begin{aligned}
\ell(\mu,\Sigma) &= -\int \log\left[(2\pi)^{-d/2}\left[\det(\Sigma+I_d)\right]^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^\top(\Sigma+I_d)^{-1}(x-\mu)\right)\right]\phi(x)\,\mathrm{d}x + \text{constant} \\
&= \frac{1}{2}\log\det(\Sigma+I_d) + \int\frac{1}{2}(x-\mu)^\top(\Sigma+I_d)^{-1}(x-\mu)\,\phi(x)\,\mathrm{d}x + \text{constant}
\end{aligned}$$

$$= \frac{1}{2}\log\det\left(\Sigma + I_d\right) + \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(0,I)}\left[(x-\mu)^\top\left(\Sigma+I_d\right)^{-1}(x-\mu)\right] + \text{constant}$$
$$= \frac{1}{2}\log\det\left(\Sigma + I_d\right) + \frac{1}{2}\text{tr}\left[\left(\Sigma+I_d\right)^{-1}\right] + \frac{1}{2}\mu^\top\left(\Sigma+I_d\right)^{-1}\mu + \text{constant}.$$

然后，我们可以如下计算 $\ell(\mu,\Sigma)$ 的欧几里得梯度：

$$\begin{aligned}
\nabla_\mu \ell(\mu,\Sigma) &= \left(\Sigma+I_d\right)^{-1}\mu,\\
\nabla_\Sigma \ell(\mu,\Sigma) &= \frac{1}{2}\left(\Sigma+I_d\right)^{-1} - \frac{1}{2}\left(\Sigma+I_d\right)^{-2} - \frac{1}{2}\left(\Sigma+I_d\right)^{-1}\mu\mu^\top\left(\Sigma+I_d\right)^{-1}\\
&= \frac{1}{2}\left(\Sigma+I_d\right)^{-1}\left(\Sigma+I_d-I_d-\mu\mu^\top\right)\left(\Sigma+I_d\right)^{-1}\\
&= \frac{1}{2}\left(\Sigma+I_d\right)^{-1}\left(\Sigma-\mu\mu^\top\right)\left(\Sigma+I_d\right)^{-1}.
\end{aligned}$$

根据 Lambert 等人 (2022年，附录 B.3)，当从 $(\mu_0,\Sigma_0) = (0, I_d)$初始化时，Bures-Wasserstein 梯度流可用以下常微分方程描述：

$$\begin{aligned}
\dot{\mu}_t &= -\left(\Sigma_t+I_d\right)^{-1}\mu_t\\
\dot{\Sigma}_t &= -\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\left[\Sigma_t-\mu\mu^\top\right]\left(\Sigma_t+I_d\right)^{-1} - \left(\Sigma_t+I_d\right)^{-1}\left[\Sigma_t-\mu\mu^\top\right]\left(\Sigma_t+I_d\right)^{-1}\Sigma_t
\end{aligned}$$

初始条件为 $\mu_0 = 0$ 和 $\Sigma_0 = I_d$。容易验证，对于所有 $t \geq 0$，$\mu_t = 0$ 均成立，而 Σ_t 的动态由下式支配：

$$\begin{aligned}
\dot{\Sigma}_t &= -\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1} - \left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\\
&= -\Sigma_t\left(\Sigma_t+I_d\right)^{-1} + 2\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1} - \left(\Sigma_t+I_d\right)^{-1}\Sigma_t\\
&= -2I_d + 2\left(\Sigma_t+I_d\right)^{-1} + 2\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\\
&= -2\left(\Sigma_t+I_d\right)^{-1}\Sigma_t^2\left(\Sigma_t+I_d\right)^{-1}
\end{aligned}$$

初始条件为 $\Sigma_0 = I_d$。我们可以验证，Σ_t 的非对角元素始终为零，且其对角元素彼此相同，并遵循以下常微分方程演化：

$$\dot{\sigma}_t = -2\frac{\sigma_t^2}{(\sigma_t+1)^2}$$

初始条件为 $\sigma_0 = 1$。容易验证，σ_t 单调递减且始终非负，也就是说，$0 \leq \sigma_t \leq 1$ 始终成立。因此我们有：

$$-2\sigma_t^2 \leq \dot{\sigma}_t \leq -\frac{1}{2}\sigma_t^2.$$

由此可得

$$\frac{1}{1+2t} \leq \sigma_t \leq \frac{2}{2+t},$$

因此

$$\frac{1}{1+2t}I \preceq \Sigma_t \preceq \frac{2}{2+t}I,$$

这表明，ρ_t 以 $O(d/t)$的速度收敛到 $\rho^\star$。

当 $t = 0$ 时，我们可以显式计算 Wasserstein梯度流的推前映射，这直观地解释了为什么 Wasserstein梯度流不会指数快速收敛。我们首先计算

$$\nabla\frac{\rho^\star * \phi(y)}{\rho_0 * \phi(y)} == \nabla\frac{(\det I)^{-d/2}\exp\left(-\frac{1}{2}\|y\|_2^2\right)}{(\det 2I)^{-d/2}\exp\left(-\frac{1}{4}\|y\|_2^2\right)} = 2^{d/2}\nabla\exp\left(-\frac{1}{4}\|y\|_2^2\right) = -2^{d/2-1}y\exp\left(-\frac{1}{4}\|y\|_2^2\right),$$

那么， $t=0$ 处的推前映射为 $x \mapsto v_0(x)$，其中

$$
\begin{aligned}
v_0(x) &= \int \left(\nabla_y \frac{\rho^\star * \phi(y)}{\rho_0 * \phi(y)}\right) \phi(y-x)\,\mathrm{d}y = -2^{d/2-1} \int y \exp\left(-\frac{1}{4}\|y\|_2^2\right) \cdot \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|x-y\|_2^2\right) \mathrm{d}y \\
&= -\frac{2^{d/2-1}}{(2\pi)^{d/2}} \int y \exp\left(-\frac{1}{2}\|x\|_2^2 + x^\top y - \frac{3}{4}\|y\|_2^2\right) \mathrm{d}y = -\frac{2^{d/2-1}}{(2\pi)^{d/2}} \int y \exp\left(-\frac{1}{6}\|x\|_2^2 - \frac{3}{4}\left\|y - \frac{2}{3}x\right\|_2^2\right) \mathrm{d}y \\
&= -2^{d/2-1}\left(\frac{2}{3}\right)^{d/2} \exp\left(-\frac{1}{6}\|x\|_2^2\right) \int \frac{1}{(2\pi)^{d/2}(2/3)^{d/2}} y \exp\left(-\frac{3}{4}\left\|y - \frac{2}{3}x\right\|_2^2\right) \mathrm{d}y \\
&= -\frac{1}{3}\left(\frac{4}{3}\right)^{d/2} \exp\left(-\frac{1}{6}\|x\|_2^2\right) x.
\end{aligned}
$$

另一方面，根据 Lambert 等人 (2022年，附录 B.3)，时刻$t=0$ 的 Bures-Wasserstein 梯度为

$$
\nabla_{\mathsf{BW}} \ell_\infty(\rho_0) = \begin{bmatrix} \nabla_\mu \ell(\mu_0, \Sigma_0) \\ 2\nabla_\Sigma \ell(\mu_0, \Sigma_0) \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{4} I_d \end{bmatrix},
$$

因此， $t=0$ 处的推前映射为 $x \mapsto x/4$。

参考文献

Allen-Zhu, Z., Li, Y., and Song, Z.（2019年）。通过过参数化建立深度学习的收敛理论。*ICML*。

Altschuler, J., Chewi, S., Gerber, P. R., and Stromme, A.（2021）。Bures-Wasserstein 流形上的平均：梯度下降的无维度依赖收敛。神经信息处理系统进展，34:22132–22145。

Ambrosio, L., Gigli, N., and Savaré, G.（2008）。梯度流：度量空间与概率测度空间中的梯度流。Springer Science & Business Media。

Bauer, M.、Bruveris, M. 和 Michor, P. W.（2016）。光滑密度空间上 Fisher–Rao 度量的唯一性。伦敦数学会会刊，48(3)：499–506。

Bubeck, S.（2015）。凸优化：算法与复杂度。Now Publishers Inc.

Chewi, S.（2022年）。对数凹采样。书稿。

Chewi, S.、Maunu, T.、Rigollet, P. 和 Stromme, A. J.（2020年）。Bures–Wasserstein 重心的梯度下降算法。载于学习理论会议，第 1276–1304 页。PMLR。

Chizat, L. 和 Bach, F.（2018年）。利用最优传输研究过参数化模型梯度下降的全局收敛。*arXiv* 预印本 *arXiv:1805.09545*。

Chizat, L.、Peyré, G.、Schmitzer, B. 和 Vialard, F.-X.（2018年）。最优传输与 Fisher–Rao 度量之间的插值距离。计算数学基础，18(1)：1–44。

Dicker, L. H. 和 Zhao, S. D.（2016）。通过非参数经验贝叶斯和最大似然推断进行高维分类。*Biometrika*, 103(1):21–34。

Gallouët, T. O. 和 Monsaingeon, L.（2017）。用于 Kantorovich–Fisher–Rao 梯度流的 JKO 分裂格式。*SIAM* 数学分析杂志，49(2):1100–1130。

Groeneboom, P.、Jongbloed, G. 和 Wellner, J. A.（2008）。用于计算混合模型中非参数函数估计的支撑缩减算法。《斯堪的纳维亚统计学杂志》，35(3):385–399。

Jewell, N. P. (1982年). 指数分布混合. 《统计年刊》，第479–4页 84.

Jiang, W. 和 Zhang, C.-H. (2009年). 正态均值的广义最大似然经验贝叶斯估计. 《统计年刊》, 37(4):1647–1684. Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J. 和 Jordan, M. I. (2016). 高斯混合模型似然中的局部极大值：结构性结果及其算法后果. 神经信息处理系统进展，29:4116–4124. Jordan, R., Kinderlehrer, D. 和 Otto, F. (1998). 福克–普朗克方程的变分形式.*SIAM* 数学分析杂志，29(1):1–17. Kiefer, J. 和 Wolfowitz, J. (1956年). 存在无限多个附带参数时最大似然估计量的相合性. 《数学统计年刊》, 第887–906页. Koenker, R. 和 Mizera, I. (2014年). 凸优化、形状约束、复合决策和经验贝叶斯规则. 《美国统计学会杂志》，109(506):674–685. Kondratyev, S., Monsaingeon, L. 和 Vorotnikov, D. (2016). 有限拉东测度空间上的一种新最优传输距离. 《微分方程进展》，21(11/12):1117–1164. Lambert, M., Chewi, S., Bach, F., Bonnabel, S. 和 Rigollet, P. (2022年). 通过Wasserstein梯度流进行变分推断. *arXiv* 预印本 *arXiv:2205.15902*. Liero, M., Mielke, A. 和 Savaré, G. (2018年). 最优熵传输问题及正测度之间一种新的 Hellinger–Kantorovich 距离. 《数学创作》，211(3):969–1117. Lindsay, B. G. (1983年). 混合似然的几何: 一般理论. 《统计年刊》，第86–94页. Lindsay, B. G. 和 Roeder, K. (1993年). 混合模型中估计的唯一性与可识别性.《加拿大统计学杂志》，21(2):139–147. Lu, Y., Lu, J. 和 Nolen, J. (2019a). 利用生灭动力学加速朗之万采样. arXiv. Lu, Y., Lu, J. 和 Nolen, J. (2019b). 利用生灭动力学加速朗之万采样. *arXiv* 预印本 *arXiv:1905.09863*. Lu, Y.、Slepčev, D. 和 Wang, L. (2022). 用于采样的生灭动力学：全局收敛、近似及其渐近性质. arXiv. Mei, S., Montanari, A. 和 Nguyen, P.-M. (2018年). 双层神经网络损失景观的平均场视角. 《美国国家科学院院刊》，115(33):E7665–E7671. Otto, F. (2001). 耗散演化方程的几何：多孔介质方程. 《偏微分方程通讯》，26(1-2):101–174. Polyanskiy, Y. 和 Wu, Y. (2020年). 混合模型中非参数最大似然估计量的自正则化性质. *arXiv* 预印本 *arXiv:2008.08244*. Saha, S. 和 Guntuboyina, A. (2020年). 关于高斯位置混合密度的非参数最大似然估计量及其在高斯去噪中的应用. 《统计年刊》，48(2):738–762. Salim, A., Korba, A. 和 Luise, G. (2020年). Wasserstein 近端梯度算法. 收录于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. 和 Lin, H.（编），神经信息处理系统进展，第33卷，第12356–12366页. Curran Associates 公司. Sander, M. E., Ablin, P., Blondel, M. 和 Peyré, G. (2022). Sinkformers：采用双随机注意力的 Transformer. *AISTATS*. Santambrogio, F. (2017). {欧几里得、度量与Wasserstein} 梯度流：综述. 《数学科学通报》，7(1):87–154.

Villani, C.（2009年）。最优传输：旧与新，第338卷。施普林格。Wainwright, M. J.（2019年）。高维统计学：非渐近视角，第48卷。剑桥大学出版社。Zhang, C.-H.（2009年）。正态混合密度的广义极大似然估计。《中华统计学志》，第1297–1318页。Zhang, Y.、Cui, Y.、Sen, B. 和 Toh, K.-C.（2022年）。关于混合模型中非参数最大似然估计量的高效且可扩展计算。*arXiv*预印本 *arXiv:2208.07514*。