基于温度分布流的高效多模态采样

Yixuan Qiu

上海财经大学统计与管理学院，中国上海 200433，qiuyixuan@sufe.edu.cn

Xiao Wang

统计系，普渡大学

西拉法叶，印第安纳州 47907，美国，wangxiao@purdue.edu

arXiv:2304.03933v1 [stat.ME] 8 Apr 2023

摘要

从高维分布中进行采样是统计研究与实践中的一个基本问题。然而，当目标密度函数未归一化且包含孤立众数时，会面临巨大挑战。我们通过拟合参考概率测度与目标分布之间的可逆变换映射（称为传输映射）来解决这一困难，从而可以通过将参考样本经由传输映射向前推送，实现从目标分布中采样。我们利用Wasserstein梯度流理论对现有基于传输的采样方法的局限性进行了理论分析，并提出了一种名为 TemperFlow 的新方法，以解决多峰问题。TemperFlow 自适应地学习一系列温度分布，使其逐步逼近目标分布，并证明了该方法能够克服现有方法的局限性。各种实验表明，与传统方法相比，这种新型采样器具有更优越的性能；此外，我们还展示了其在现代深度学习任务（如图像生成）中的应用。数值实验的编程代码可在https://github.com/yixuan/temperflow获取。

关键词：深度神经网络、梯度流、马尔可夫链蒙特卡罗、归一化流、并行回火

1 引言

从概率分布中采样是统计研究与实践中的一项基础任务，在贝叶斯数据分析（Gelman 等人，2014）、不确定性量化（Sullivan，2015）以及生成式机器学习模型（Salakhutdinov，2015；Bond-Taylor 等人，2021）等众多领域都有广泛应用。长期以来，这一直是统计计算与模拟中的核心课题，并发展出了许多标准方法，包括拒绝采样、重要性采样、马尔可夫链蒙特卡罗（MCMC）等。一般的采样问题可以描述如下。设 $p(x)$ $x \in \mathbb{R}^d$ 为我们希望从中采样的目标密度函数，且假设其形式未归一化。换言之，我们只能访问能量函数$E(x) = -\log p(x) + C$，其中 C 是与 x 无关的未知常数。采样的目标是生成一组相互独立的点样本 $X_1, \ldots, X_n$，每个点都服从 $p(x)$ 分布。

尽管标准采样算法已广泛应用，但当目标分布具有复杂结构（包括高维、孤立众数以及各分量之间的强相关性）时，仍会面临巨大挑战。因此，许多采样方法并不能完全适用于这类复杂问题。例如，拒绝采样要求感兴趣的密度函数具有严格上界，而这通常难以求得，更不用说未归一化的情形了。另一方面，重要性采样主要用于近似积分，而非生成独立样本，并且其方差可能随变量维数呈指数级增长（见Owen的2013年版Gilks 等人，1995年；Brooks 等人，2011；Dunson 和 Johndrow，2020）。为此，马尔可夫链蒙特卡罗（MCMC）成为面向高维分布最受欢迎的通用采样器之一，具有良好的统计性质，并且有许多可用的变体（Gilks 等人，1995年；Brooks 等人，2011；Dunson 和 Johndrow，2020）。

然而，马尔可夫链蒙特卡罗方法并非没有局限性。首先，马尔可夫链蒙特卡罗生成的随机变量

彼此相关，导致在近似计算期望时有效样本量降低。此外，这类方法不适用于要求样本相互独立的场景。其次，在大多数马尔可夫链蒙特卡罗算法中，用户通常需要手动调整超参数，这需要丰富的专业知识和经验。此外，检验马尔可夫链是否收敛并非易事，因此很难确定停止规则。最后但同样重要的是，目标密度函数通常包含许多孤立众数，这使得马尔可夫链蒙特卡罗样本对初始状态高度敏感，因为采样点可能会被困在某个众数中，难以逃脱。所有这些因素都严重阻碍了马尔可夫链蒙特卡罗在实践中的应用。

近来，薛定谔–福尔默采样器（Huang 等人，2021年）是一种基于扩散过程的新型采样方法，可在时间零点将服从退化分布的粒子输运至时间一点的目标分布。薛定谔–福尔默采样器易于实现，并且在多峰高斯混合模型上表现出良好的前景。然而，扩散过程的漂移项并非总有闭式解，可能需要通过蒙特卡罗方法进行近似，从而进一步增加计算负担。

为克服上述困难，本文倡导一种用于高效统计抽样和模拟的新框架。我们的方法受到 Marzouk et al. (2016) 提出的测度输运框架的启发，其核心思想是估计基准概率测度与目标分布之间的确定性输运映射。基准测度通常是一个固定且便于使用的分布，随后只需通过传输映射将一个独立同分布的参考样本向前推送，即可获得来自目标分布的独立随机变量。尽管这一框架颇具吸引力，但 Marzouk et al. (2016) 中仍有许多尚未解决的挑战，其中最令人关注的是映射的灵活性和计算难度。更重要的是，我们在第 2.3 节指出，这种方法存在内在局限우우우우우우

当目标分布具有多个众数时的现有方法。特别是，我们展示了这些方法即使对于诸如双峰正态混合分布这样的简单分布也会失效。

为了揭示此类失败的原因，我们使用概率测度梯度流（也称为Wasserstein梯度流（Ambrosio等人，2008年；Santambrogio，2017年）分析现有的测度传输方法。Wasserstein梯度流可以被视为概率测度的连续演化，是研究涉及分布的优化问题的收敛性质的强大工具。为了解决多模态采样问题，我们借鉴模拟退火和并行退火算法的思想（Kirkpatrick等人，1983；Swendsen和Wang，1986；Geyer，1991；Marinari和Parisi，1992；Geyer和Thompson，1995年；Neal，1996），并提出一种新方法，用于估计基准测度与目标多模态分布之间的传输映射。我们证明，在非常温和的条件下，所提出的算法能够快速收敛，并从目标分布中生成高质量样本。

值得一提的是，所提出的采样方法极大地受益于深度学习的最新进展。具体而言，由于深度神经网络具有近似高度非线性关系的卓越表达能力，我们主要使用深度神经网络（DNNs，Goodfellow等人，2016）来构建测度之间的传输映射。DNNs已广泛应用于统计建模（Yuan等人，2020；Pang等人，2020；Qiu和 Wang，2021年；Sun等人，2021年；Liu 等人，2021年）以及给定训练数据的采样（Romano 等人，2020；Zhou 等人，2021年）；在本文中，我们还展示了它们在给定能量函数的情况下进行采样的巨大潜力。结合随机逼近（Robbins和Monro，1951）和自适应梯度优化（Kingma和Ba，2015）等强大的优化技术，所提出的方法可扩展到复杂的高维分布，并且能够借助图形处理器（GPU）等现代计算硬件加速。

本文其余部分的安排如下。在第2 节中，我们介绍背-

基于Marzouk 等人 (2016提出的测度输运框架，我们介绍现有的采样器，讨论其问题与挑战，并阐述它们与马尔可夫链蒙特卡罗的联系。第 3节专门介绍 Wasserstein 梯度流的理论框架，并在该框架下分析现有采样器。我们在第 4节提出新的 TemperFlow采样器，并在第 5节介绍其理论性质。第 6 节开展了多项数值实验，以展示所提出采样器的有效性，并在第7节展示其在现代深度生成模型中的应用。最后，我们在第8节总结本文并进行讨论。

2 背景与相关工作

2.1 测度传输

令 μ： $\mathcal{B}(\mathbb{R}^d) \to [0,1]$ 为我们希望从中进行采样的概率测度，该测度定义在 $\mathbb{R}^d$ 上的博雷尔 σ-代数上。本文聚焦于从连续分布中进行采样，因此假设 μ 关于 λ具有密度函数 $p(x)$，其中 λ是 $\mathbb{R}^d$上的勒贝格测度。在下文中，我们将交替使用 μ 和 $p(x)$来表示目标分布。令 μ_0： $\mathcal{B}(\mathbb{R}^d) \to [0,1]$ 为另一个可以轻松生成独立样本的概率测度。在大多数情况下， μ_0 可以选为简单的固定分布，例如标准多元正态分布 $N(0, I_d)$，我们也遵循这一约定。如果对于任意集合 $A \in \mathcal{B}(\mathbb{R}^d)$，映射 T： $\mathbb{R}^d \to \mathbb{R}^d$ 满足 $\mu(A) = \mu_0(T^{-1}(A))$，则称其将 μ_0 推前至 μ ，记为 $T_\sharp \mu_0 = \mu$。在这种情况下， T 也称为传输映射。在全文中，我们假设传输映射是可逆且可微的，即， T 是一个微分同胚。

当 μ和 μ_0相对于 λ绝对连续时，从μ_0 到 μ 的传输映射总是存在的，但不一定唯一（Villani，2009年）。要从目标分布 μ中生成随机样本，只需模拟 Z_1、……、$Z_n \overset{iid}{\sim} \mu_0$,

由此可得 $T(Z_1)$……， $T(Z_n) \overset{iid}{\sim} \mu$。因此，该框架的关键步骤是根据能量函数 $E(x)$估计传输映射 T。

Marzouk 等人(2016）提出求解 T 所需的以下优化问题

$$\min_{T} \ \mathrm{KL}(T_\sharp \mu_0 \| \mu) \quad \text{subject to} \quad \det \nabla T > 0, \ T \in \mathcal{T}, \tag{1}$$

其中，$\mathrm{KL}(\nu\|\mu) = \int \log(\mathrm{d}\nu/\mathrm{d}\mu)\mathrm{d}\nu$ 是从 μ 到 ν的 Kullback–Leibler（KL）散度， ∇T是 T的雅可比矩阵， $\mathcal{T}$是一个合适的微分同胚集合。在 Marzouk 等人（2016）中， $\mathcal{T}$ 是基于 Knothe–Rosenblatt 重排的三角映射，这意味着 T 的第 i 个输出变量仅取决于前 i 个输入变量。第2节的第2.2节表明， $\mathcal{T}$也有更一般的定义。条件 $\det \nabla T > 0$ 保证推前密度 $T_\sharp \mu_0$ 在 μ的支撑集上为正。值得注意的是，优化 (1) 只需要 μ的能量函数，因为 (1) 等价于

$$\min_{T} \int p_0(x) \ [E(T(x)) - \log \det \nabla T(x)] \, \mathrm{d}x \text{ 满足 } \det \nabla T > 0, \ T \in \mathcal{T}, \text{ 其中 } p_0(x) \tag{2}$$

是 μ_0 的密度函数。在实践中， T 由有限维向量 θ 参数化，并且可以通过蒙特卡洛样本对（2）中的积分进行无偏估计。例如，Marzouk 等人（2016）使用多项式逼近 $\mathcal{T}$ 空间，并通过随机梯度下降优化（2）。然而，对于多项式而言，要保持 T 的可逆性并满足约束 $\det \nabla T > 0$ 并非易事，因此 Marzouk 等人（2016）中的方法适用性受到严重限制。

2.2 神经传输采样器

为克服这些问题，Hoffman 等人 (2019通过使用逆自回归流扩展了 Marzouk 等人 (2016) 的工作，以表示传输映射（Kingma 等人，2016）， T从而实现更具可扩展性的计算。逆自回归流属于更广泛的归一化流类别（Tabak 和 Vanden-Eijnden，2010；Tabak 和 Turner，2013；Rezende

和 Mohamed，2015），它可以描述为通过一系列可逆映射对概率密度进行变换。归一化流还包括许多其他变体，例如仿射耦合流（Dinh 等人，2014；Dinh 等人，2016）、掩码自回归流（Papamakarios 等人，2017年）、神经样条流（Durkan 等人，2019）和线性有理样条流（Dolatabadi 等人，2020）。这些模型通常使用神经网络来构造可逆映射，因此在文献中也被称为可逆神经网络。需要说明的是，归一化流中的“流”与梯度流中的“流”在概念上有所不同，本文关注的是后者。因此，为避免歧义，下文将使用“可逆神经网络”指代归一化流模型。补充材料的第 A.1 节对可逆神经网络进行了更详细的介绍。

需要指出的是，上述所有可逆神经网络都通过恰当设计其神经网络架构，保证为满足 $\det \nabla T > 0$ 的微分同胚。因此，它们是构建传输映射 T （1）的自然且强大的工具。我们使用术语 神经传输采样器来表示使用可逆神经网络构建传输映射 T 的测度传输采样方法。由于在（1）中， T 是通过最小化KL散度学习得到的，因此，当需要强调这一目标函数时，我们也将此类方法称为 KL 神经输运采样器，简称 KL采样器。KL采样器的概要如算法1所示。

2.3 问题与挑战

尽管（1）的全局极小值点保证满足 $T_\sharp \mu_0 = \mu$ 前提是 $\mathcal{T}$足够丰富，但在实践中，如果 μ 包含多个孤立众数，收敛可能会极其缓慢。下面，我们使用一个示例来说明这一问题。具体而言，我们将神经传输采样器应用于两个简单的单变量分布：第一个具有单峰且对数凹的密度函数 $p_u(x) = \exp\{x - \exp(x/3)\}/6$，第二个

算法 1测度输运框架下的 KL 神经输运采样器。输入：具有能量函数 $E(x)$ 的目标分布 μ，具有初始参数值 $\theta^{(0)}$ 的可逆神经网络 T_θ，批量大小 M，步长 $\{\alpha_k\}$ ^输出：神经网络参数 θ，使得 $T_{\hat{\theta}\sharp}\mu_0 \approx \mu$1：对于 $k = 1$，2，.. .执行2：生成 Z_1，.. .，$Z_M \overset{iid}{\sim} \mu_0$3：定义 $L(\theta) = M^{-1}\sum_{i=1}^{M}[E(T_\theta(Z_i)) - \log\det \nabla_z T_\theta(z)|_{z=Z_i}]$4：设置 $\theta^{(k)} \leftarrow \theta^{(k-1)} - \alpha_k \nabla_\theta L(\theta)|_{\theta=\theta^{(k-1)}}$5：如果 $L(\theta)$ 收敛，则 6：返回 $\hat{\theta} = \theta^{(k)}$7：结束如果 8：结束循环

是两个正态分布的混合 $p_m(x) \sim 0.7 \cdot N(1, 1) + 0.3 \cdot N(8, 0.25)$。基准测度为 $\mu_0 = N(0, 1)$，T初始化为恒等映射，并使用随机梯度下降来更新 T 中的参数。令 $T^{(k)}$ 表示经过 k 次迭代后估计得到的传输映射，然后在图 1 中绘制 $T_\sharp^{(k)}\mu_0$ 在 $k = 0$、10、50、100、500 时的密度函数。

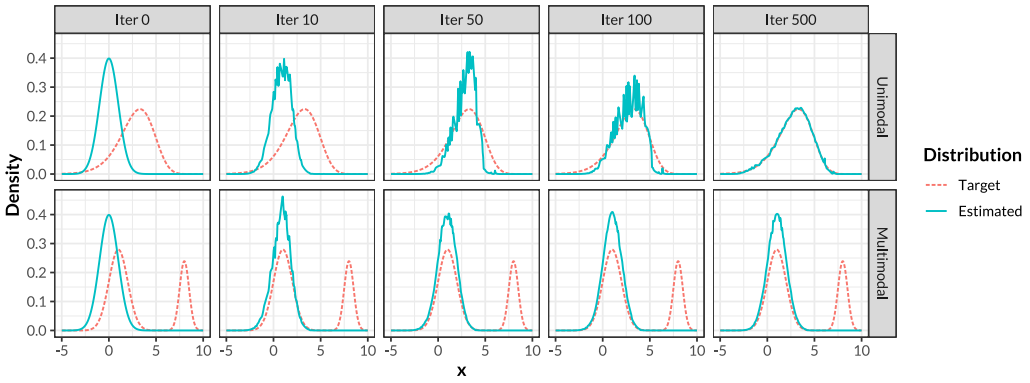


图 1：将 KL采样器应用于两个分布。第一行展示单峰分布的估计密度演变，第二行展示双峰分布的估计密度演变。每一列对应优化过程中的一次迭代。

显然，对于单峰密度 $p_u(x)$，$T_\sharp^{(k)}\mu_0$ 收敛得非常快，并且很好地逼近目标密度，而在双峰正态混合分布的情况下，$T_\sharp^{(k)}\mu_0$ 只能捕获第一个模态，并且在五十次迭代之后几乎没有进展。这个启发性示例表明，当目标密度 $p(x)$ 为多峰时，优化问题（2）具有内在的困难。因此，我们需要仔细分析

优化过程，以揭示失败的原因。

2.4 与马尔可夫链蒙特卡罗的联系

马尔可夫链蒙特卡罗被列为20世纪顶尖算法之一（Dongarra和Sullivan，2000）。然而，对于远非对数凹且具有许多孤立众数的分布，当前的马尔可夫链蒙特卡罗技术还需要额外的方法。例如，模拟退火（Marinari和Parisi，1992）就是这类颇具吸引力的方法。模拟退火会在不同温度下的原始马尔可夫链变体之间进行切换。其背后的直觉是，处于较高温度下的马尔可夫链能够更容易地跨越能量势垒，因此可以充当孤立众数之间的桥梁。对于这一启发式方法，可证明的结果寥寥无几。Zheng（2003年）以及Woodard等人（2009年）建立了通用模拟链谱隙的下界，但Woodard等人（2009年）中的谱隙界关于众数数量呈指数级微小。对于朗之万动力学（Bhattacharya，1978），Raginsky等人（2017年）、Cheng等人（2018）以及Vempala和Wibisono（2019）分析了任意非对数凹分布中离散化的影响，但混合时间通常是指数级的。Ge等人（2018）将朗之万扩散与模拟退火相结合，并对同一形状的对数凹分布混合上的马尔可夫链进行了理论分析。特别是，Ge等人（2018）证明了在这一设定下多模态分布能够快速混合。其他相关技术包括并行回火（Geyer，1991；Falcioni和Deem，1999）以及模拟退火（Kirkpatrick等人，1983）。

我们强调，测度传输和马尔可夫链蒙特卡罗彼此互为补充。具体而言，测度传输需要进行大量训练才能获得传输映射，而马尔可夫链蒙特卡罗无需任何训练即可生成样本。但是，在获得传输映射后，对于测度传输而言，生成步骤就变得简单且极其快速；相反，马尔可夫链蒙特卡罗需要迭代运行马尔可夫链。

3 理论性　神经传输采样器的解析分析

3.1 瓦瑟斯坦梯度流

在本节中，我们将介绍 Wasserstein梯度流，作为分析神经传输采样器的工具。假设传输映射 T 属于希尔伯特空间 $\mathbb{H}=\{T:\int\|T(x)\|^2\mathrm{d}\mu_0(x)<\infty\}$，该空间配备内积 $\langle T,M\rangle_{\mathbb{H}}=\int\langle T(x)\ M(x)\rangle\mathrm{d}\mu_0(x)$，其中 $\|\cdot\|$ 和 $\langle\cdot,\cdot\rangle$ 分别表示通常的欧几里得范数和内积。对于泛函 $\mathcal{G}(T)$，将其泛函导数 $\delta\mathcal{G}/\delta T$ 定义为映射 $M:\mathbb{R}^d\to\mathbb{R}^d$，使得对于任意 $\Psi\in\mathbb{H}$，都有 $\mathrm{d}\,\mathrm{d}_\varepsilon\mathcal{G}(T+\varepsilon\Psi)\big|_{\varepsilon=0}=\langle M,\Psi\rangle_{\mathbb{H}}$，并假设相关量均有定义。于是，主优化问题 (1) 可通过梯度下降方法求解如下。对于目标泛函 $\mathcal{G}(T)=\mathrm{KL}(T_\sharp\mu_0\|\mu)$，在第 k 次迭代时，我们对当前传输映射 $T^{(k)}$ 施加一个小扰动，得到新映射$T(k+1)=T(k)-\alpha_kM(k)$，其中 $\alpha_k>0$是步长，而 $M(k)=(\delta\mathcal{G}/\delta T)|_{T=T^{(k)}}$是泛函 $\mathcal{G}(T)$在 $T(k)$处的泛函导数。在命题1中，我们将证明，在温和的光滑性假设下，$\delta\mathcal{G}/\delta T$ 存在且具有简单的形式。

命题*1*。令 $p(x)$和 $q(x)$分别为 μ 和 $T_\sharp\mu_0$ 的密度函数，并将 $\|\cdot\|_{\mathrm{op}}$ 表示一个矩阵的算子范数。对于任意 T、$\Psi\in\mathbb{H}$，如果 *(a)* $\|\nabla(\Psi\circ T^{-1})(x)\|_{\mathrm{op}}\le C_\Psi$，$T$且 $\|\nabla^2\log p(x)\|_{\mathrm{op}}\le c$对于某些不依赖于 x 的常数C_Ψ、T、$c>0$；*(b)* $[\nabla(\log p-\log\ q)]\circ T\in\mathbb{H}$；以及 *(c)*$\lim_{\|x\|\to\infty}\|q(x)\Psi(T^{-1}(x))\|=0$，则

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{G}(T+\varepsilon\Psi)\bigg|_{\varepsilon=0}=\left\langle\frac{\delta\mathcal{G}}{\delta T},\Psi\right\rangle_{\mathbb{H}},$$

其中 $\delta\mathcal{G}/\delta T$是 $\mathcal{G}(T)$关于 T的泛函导数，给定为 $(\delta\mathcal{G}/\delta T(x)=M(T(x))$，$M(x)=\nabla[\log q(x)-\log p(x)]$。

在命题1中，假设 (a) 用于控制映射和能量函数的平滑性，假设 (b) 确保计算得到的导数定义良好，

而假设（c）基本上意味着 q 的密度在无穷远处趋于零。它们并未对我们接下来要分析的对象形式施加实质性限制。

为了进一步深入了解优化算法并避免不必要的技术细节，我们考虑步长为无穷小的情况，此时优化过程变为连续时间动力学。令 $T^{(t)}$ 为时间 t 时的传输映射，则 $T^{(t)}$ 按照一个微分方程演化，

$$\frac{\partial}{\partial t}T^{(t)}(x) = -\left(\frac{\delta \mathcal{G}}{\delta T}\bigg|_{T=T^{(t)}}\right)(x) = \mathbf{u}_t(T^{(t)}(x)), \quad \mathbf{u}_t(x) = -\nabla\left[\log p_t(x) - \log p(x)\right], \quad (3)$$

其中 $p_t(x)$ 是 $T_\sharp^{(t)}\mu_0$ 的密度函数，整个过程 $T_{(t)}$ 通常称为关于泛函 $\mathcal{G}$的梯度流。

然而，向量场 $\mathbf{u}_t$ 间接依赖于 $T^{(t)}$，途径是 $p_t(x)$，这使得 $T^{(t)}$ 的收敛性质难以分析。为此，我们引入概率测度空间中的 Wasserstein梯度流概念。概括而言，我们不再研究传输映射 T，而是将（1）视为关于概率测度的优化问题$T_\sharp\mu_0$，并为该测度建立一个与（3）相当的微分方程。

记 $\mathscr{P}(\mathbb{R}^d)$ 为 $\mathbb{R}^d$ 上具有有限二阶矩的博雷尔概率测度空间。Ambrosio等人的定理 8.3.1(2008 年表明，任意绝对连续曲线 $\mu_t : (0,T) \to \mathscr{P}(\mathbb{R}^d)$ 都是连续性方程的一个解

$$\frac{\partial}{\partial t}\mu_t + \nabla \cdot (\mathbf{v}_t \mu_t) = 0 \quad (4)$$

对于 ome 向量场 $\mathbf{v}_t : \mathbb{R}^d \to \mathbb{R}^d$。（4）在分布的意义下成立，即，

$$\int_0^T \int_{\mathbb{R}^d} \left(\partial_t \varphi(x,t) + \langle \mathbf{v}_t(x), \nabla_x \varphi(x,t)\rangle\right) \mathrm{d}\mu_t \mathrm{d}t = 0, \quad \forall \varphi \in C^\infty(\mathbb{R}^d \times (0,T)),$$

其中 C^∞ 是无穷可微函数的类别。在一些温和条件下，曲线 μ_t 由向量场 $\mathbf{v}_t$唯一确定，反之亦然（Ambrosio等人，

2008年, 命题 8.1.8, 定理 8.3.1).

设 $\mathcal{F}$: $\mathscr{P}(\mathbb{R}^d) \to \mathbb{R}$ 为如下形式的泛函: $\mathcal{F}(\rho) = \int F(x, \rho(x)\ \nabla\rho(x))\mathrm{d}x$, 其中 ρ是定义在 $\mathbb{R}^d$ 上的概率测度, 并配备可微密度函数 $\rho(x)$, 而 F是关于其三个自变量的光滑函数。将 $\mathcal{F}$ 在 ρ处的一阶变分$(\delta\mathcal{F}/\delta\nu)(\rho)$定义为满足以下条件的函数: 对于任意扰动 ψ, $\mathrm{d}\,\mathrm{d}_\varepsilon{}^{\mathcal{F}}(\rho+\varepsilon\psi)\big|_{\varepsilon=0} = \int(\delta\mathcal{F}/\delta\nu)(\rho)\mathrm{d}\psi$ (San- tambrogio, 2017年)。那么, 如果满足以下条件, 就称 μ_t 是相对于泛函 $\mathcal{F}$ 的Wasserstein梯度流。

$$\mathbf{v}_t = -\nabla\left(\frac{\delta\mathcal{F}}{\delta\nu}(\mu_t)\right). \tag{5}$$

总而言之, Wasserstein梯度流 μ_t 通过连续性方程 (4) 刻画概率测度随时间的演化, 其向量场$\mathbf{v}_t$ 由基于方程 ($\mathcal{F}$ 5) 决定。

对于测度输运问题 (1), $\mathcal{F}(\cdot) = \mathcal{F}_{\mathrm{KL}}(\cdot) := \mathrm{KL}(\cdot\|\mu)$。接着, Gao 等人 (2019) 的引理 2.1 表明, 对于 $\mathcal{F}_{\mathrm{KL}}$, $\mathrm{v}_t(x) = -\nabla\left[\log\ p_t(x) - \log\ p(x)\right],$, 其中 $p_t(x)$ 是 μ_t 的密度函数。值得注意的是, 这个 v_t 与 (3) 中的 u_t 完全相同。事实上, Wasserstein梯度流理论指出, 在 v_t 满足适当正则性条件时, 传输映射的梯度流$T(t)$ 与概率测度 μ_t 的 Wasserstein梯度流等价, 即 $T^{(t)}$ 和 μ_t 通过关系 $\mu_t = T^{(t)}_\sharp\mu_0$ 相联系 (Ambrosio等人, 2008年, 引理 8.1.4, 命题 8.1.8)。因此, 可以通过研究 μ_t 来理解 $T^{(t)}$ 的行为, 反之亦然。本文关注后者, 因为这为分析优化算法的收敛行为提供了很大便利。

3.2 When the neural transportsampler fails

算法 1 中的神经传输采样器使用离散时间梯度下降法来更新采样器分布 $\{T^{(k)}\mu_0\}$, 这可以视为关于目标泛函 μ_t 的 Wasserstein 梯度流 $\mathcal{F} = \mathcal{F}_{\mathrm{KL}}$的离散化版本。对于

理论分析，我们关注连续的 μ_t，尽管在实际实现中必须将其离散化。为方便起见，我们也将 μ_t 称为时间 t 处的采样器分布。

Wasserstein梯度流的一个重要事实是，曲线 μ_t 由（5）中的向量场支配，会随时间推移而降低泛函 $\mathcal{F}$ （Ambrosio 等人，2008年，第233页）：

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) = -\int p_t(x)\|\mathbf{v}_t(x)\|^2 \mathrm{d}x \leq 0, \tag{6}$$

这意味着如果 $\mathcal{F} \geq 0$， μ_t 最终会收敛。然而，其收敛速度会受到目标密度 $p(x)$ 形状的显著影响。回想图 1，采样器分布只需数百次迭代就能准确估计目标分布 $p_u(x)$。事实上，这一优良性质适用于所有对数凹密度，且收敛速率可以量化。令 $\chi^2(\nu\|\mu) = \int (\mathrm{d}\nu/\mathrm{d}\mu)^2 \mathrm{d}\mu - 1$ 和 $H^2(\mu,\nu) = \frac{1}{2}\int(\sqrt{\mathrm{d}\mu/\mathrm{d}\lambda} - \sqrt{\mathrm{d}\nu/\mathrm{d}\lambda})^2\mathrm{d}\lambda$ 分别表示 μ 与 ν 之间的散度和平方海灵格距离。于是我们得到以下结果。

定理1。 假设 $p(x)$ 是对数凹的，即 $\log p(x)$ 是凹函数。设 Σ 为 $p(x)$的协方差矩阵，并将 Σ的最大特征值记为 σ^2 。那么对于任意 $t > 0$:

1. $\mathcal{F}_{\mathrm{KL}}(\mu_t) \leq e^{-t/(2\sigma^2 C(d))}\chi^2(\mu_0\|\mu)$，其中 $C(d) = \exp\left\{c\sqrt{l_d}\cdot[\log(l_d) + 1]\right\}, l_d = \log(d)+1$，且 $c > 0$ 是一个通用常数。

2. $$\left|\frac{\mathrm{d}\mathcal{F}_{\mathrm{KL}}(\mu_t)}{\mathrm{d}t}\right| \geq [\sigma^2 C(d)]^{-1}\cdot[2 - H^2(\mu_t,\mu)]\cdot H^2(\mu_t,\mu) \geq [\sigma^2 C(d)]^{-1} H^2(\mu_t,\mu).$$

定理的第一部分1 表明，对于对数凹密度，沿着梯度流，采样器分布与目标分布之间的差异会以指数速度衰减，这解释了图1中神经传输采样器对于单峰分布的快速收敛。第二部分对这一现象提供了更为细致的解释。具体而言，它表明时间导数 $\mathrm{d}\mathcal{F}_{\mathrm{KL}}(\mu_t)/\mathrm{d}t$ 只要 μ_t 不同于 μ就不会消失，我们称之为非零梯度性质。直观地说，

这意味着速度场$\mathbf{v}_t$不断推动当前分布 μ_t 趋近于μ，而变化率（以 $|\mathrm{d}\mathcal{F}_{\mathrm{KL}}(\mu_t)/\mathrm{d}t|$衡量）不会太小，除非 μ_t已经接近 μ。如果出现后一种情况，那么我们实际上已经实现了采样目标。因此，对于对数凹密度，导数的绝对值较小意味着 μ_t 与 μ之间的海灵格距离较小。

然而，当目标密度包含多个模态时，梯度不消失性质可能不再成立。我们考虑一个简单但富有启发性的模型来说明这一问题。设 $h(x) = e^{-V(x)}$ 是一个基准密度函数，并将目标混合密度定义为 $p(x) = \alpha h(x) + (1-\alpha)h(x-\mu)$，其中 $\alpha \in (0,1)$ 是混合概率，且$\|\mu\|$ 量化了两个分量之间的距离。假设在某个时刻 t^*，采样器分布 μ_{t^*} 也呈现为混合形式，但具有不同的混合概率 $\gamma \neq \alpha$，即 $p_{t^*}(x) = \gamma h(x) + (1-\gamma)h(x-\mu)$。接下来我们将证明，梯度流 μ_t 表现出与对数凹情形几乎相反的性质，而且这一结论甚至可以推广到更广泛的目标泛函类别。定义一个基于一般 ϕ散度的泛函 $\mathcal{F}_\phi(\cdot) = \mathcal{D}_\phi(\cdot\|\mu)$，其中 $\mathcal{D}_\phi(\nu\|\mu) = \int \phi(\mathrm{d}\nu/\mathrm{d}\mu)\mathrm{d}\mu$，并且 ϕ 是一个满足 $\phi(1) = 0$ 的凸函数。容易证明，KL散度是 ϕ散度的一个特例，其中 $\phi(x) = x\log(x)$。于是我们得到以下结果。

定理*2*。设 X 是服从 $h(x) = e^{-V(x)}$的随机向量，并定义$Y_{\mu,1} = V(X-\mu) - V(X)$, $Y_{\mu,2} = V(X) - V(X+\mu)$。假设：*(a)* 对任意固定的 $x \in \mathbb{R}^d$, $h(x-\mu) \to 0$，当$\|\mu\| \to \infty$时；*(b)* 存在常数 $c_1, C_1, k \geq 0$，使得对于所有满足 $\|x-y\| \geq c_1$的 $x, y \in \mathbb{R}^d$，$\|\nabla V(x) - \nabla V(y)\| \leq C_1\|x-y\|^k$；*(c)* 存在常数 $C_2 > 0$，使得当 $\|\mu\| \to \infty$, $i = 1$趋于*2*时，$C_1\|\mu\|^k \cdot P(\|\mu\|^{-1}|Y_{\mu,i}| \leq C_2) \to 0$。于是我们有以下结论。

1.$\lim_{\|\mu\|\to\infty} \mathcal{F}_\phi(\mu_{t^*}) = \alpha\phi\left(\frac{\gamma}{\alpha}\right) + (1-\alpha)\phi\left(\frac{1-\gamma}{1-\alpha}\right) > 0$. 2. $\lim_{\|\mu\|\to\infty} \mathrm{d}\mathcal{F}_\phi(\mu_t)/\mathrm{d}t\big|_{t=t^*} = 0$.

定理2 指出，当 $p(x)$ 的两个分量彼此相距很远时，存在一种 μ_t 配置，使其不同于目标分布，但与此同时梯度为零。由于梯度刻画了当前目标泛函在瞬间下降的速度，因此梯度消失意味着收敛速度极其缓慢。定理2 也解释了图1中的多峰示例。在第 100 次迭代时，采样器分布集中在第一个模态上，形成混合分布 $p_{t^*}(x)$ ，其中 $\gamma \approx 1$，因此定理2 成立。此外，由于该结果适用于一般的 ϕ-散度，因此，神经传输采样器的内在局限性无法通过简单地更改散度度量轻易解决。

总之，定理 1 和定理 2表明，目标密度的形状会极大地影响神经传输采样器的收敛速度。不幸的是，现实问题中的大多数目标密度都不是对数凹的，因此神经传输采样器需要进行修改，以适应更复杂的采样问题。

4 TemperFlow采样器

4.1 概述

第 3.2 节中的分析表明，关于 μ_t 的 ϕ-散度 $\mathcal{D}_\phi(\cdot\|\mu)$ 的Wasserstein梯度流在近似目标分布 μ 时 $p(x)$为多峰分布时，由于梯度消失现象而效率低下。因此，找到一条能够快速收敛到 μ的替代曲线至关重要。受优化中的模拟退火方法和马尔可夫链蒙特卡罗的启发，考虑 μ的退火密度函数，其定义为

$$q_{\beta_t}(x) = \frac{1}{Z(\beta_t)} e^{-\beta_t E(x)}, \tag{7}$$

其中， β_t 是一个递增连续函数，满足 $\beta_0 > 0$ 且 $\lim_{t\to\infty} \beta_t = 1$； $Z(\beta_t)$ 是归一化常数，使 $q_{\beta_t}(x)$成为密度函数。在模拟

退火文献中，β_t 通常称为逆温度参数。设 ρ_t 为与 $q_{\beta_t}(x)$对应的概率测度曲线，则容易发现，当 $t \to \infty$时，ρ_t收敛于 μ，因为若 $\beta_t = 1$，$q_{\beta_t}(x)$就会化简为 $p(x)$。

回火曲线 ρ_t 具有几个良好性质。首先，对于任意 $t > 0$，$q_{\beta_t}(x)$与 $p(x)$具有相同的局部极大值，因此 $p(x)$的所有众数都沿着曲线 ρ_t得以保留。这一点很重要，因为众数丢失是许多现有采样方法的常见问题，正如数值实验中将要展示的那样。其次，对于 $t < s$，$q_{\beta_t}(x)$是一个比 $q_{\beta_s}(x)$更“平坦”的分布，众数之间具有更好的连通性，因此从中间分布 $q_{\beta_t}(x)$进行采样通常比直接从 $p(x)$进行采样容易得多。第三，与 Wasserstein梯度流类似，回火曲线 ρ_t 相对于目标分布也会降低 KL散度，如定理3所示。

定理3。令 ρ_t 为与 $q_{\beta_t}(x)$ 在（7）中定义的概率测度，则 $\mathrm{d}\mathcal{F}_{\mathrm{KL}}(\rho_t)/\mathrm{d}t \leq 0$。等号成立当且仅当 $\rho_t = \mu$。

然而，追踪 ρ_t 的精确路径并不可行，因此更现实的做法是让采样器学习从回火曲线ρ_t中提取的分布序列 $\{r_k\}$。具体做法是将 β_t 离散化为序列 $\{\beta_k\}$，并定义

$$r_k(x) = q_{\beta_k}(x),\ k = 0, 1, \ldots, K,\ \beta_0 < \beta_1 < \cdots < \beta_K = 1. \tag{8}$$

然后，我们寻求传输映射来连接每个 r_{k-1} 与 r_k。我们将序列 $\{r_k\}$ 称为关于 $p(x)$的退火分布流，以体现这一思想，并将我们提出的算法简称为TemperFlow采样器。TemperFlow的核心思想是学习一系列传输映射 $T^{(k)}, k = 0,1,\ldots, K$，使得 $T(k)\sharp\ \mu_0 \approx r_k$，其中每个 $T(k)$都使用 $T^{(k-1)}$ 作为热启动进行优化。TemperFlow采样器的概要见算法 2，其中子程序 kl_采样器将算法1应用于温度化能量函数$\beta_0 E(x)$，而l2_采样器和estimate_beta将在后续章节中介绍。

算法 2 TemperFlow采样器的概要。

输入： 目标分布 μ，其能量函数为 $E(x)$；可逆神经网络 T_θ，由参数化；初始逆温度 β_0

输出： 估计的传输映射 $\hat{T}$，满足 $\hat{T}_\sharp\mu_0 \approx \mu$

初始化 $\theta \leftarrow \theta^{(-1)}$，使其满足 $T^{(-1)} := T_{\theta^{(-1)}} \approx I$，即恒等映射
计算 $\theta^{(0)} \leftarrow$ `kl_sampler`$(\beta_0 E, \theta^{(-1)})$ (算法 1)
对于 $k = 1, 2, \ldots$ **do**
  计算 $\beta_k \leftarrow \min\{1,$ `estimate_beta`$(\beta_{k-1}, \theta^{(k-1)})\}$ (算法 4)
  计算 $\theta^{(k)} \leftarrow$ `l2_sampler`$(\beta_k E, \theta^{(k-1)})$ (算法 3)
  if $\beta_k = 1$ **then**
    返回 $\hat{T} = T_{\theta^{(k)}}$
  结束if
结束循环

4.2 温度之间的传输

平滑地改变 $T^{(k-1)}$ 至 $T^{(k)}$ 是 TemperFlow 的核心部分，可视为寻找一条连接 r_{k-1} 与 r_k的曲线。理论上，可以使用算法 1 中的现有 KL采样器，但第 3.2 节的分析表明，对于多峰分布，它可能会遭遇梯度消失问题。为此，我们提出了一种名为 L^2 采样器的新方法，可显著提升采样质量。

L^2 采样器背后的思路很简单。与算法 1 中最小化采样器分布和目标分布之间的KL散度不同，L^2采样器以两个密度之间的平方 L^2-距离作为目标函数，定义为 $\mathcal{L}(g) = \int (g - f)^2 \mathrm{d}\lambda$，其中 f 是目标密度，g 表示采样器密度。给定参数化传输映射 T_θ（通常是可逆神经网络），L^2 采样器使用随机梯度下降最小化 $\mathcal{L}(p_\theta)$，其中 p_θ 是 $T_{\theta\sharp}\mu_0$的密度函数。我们将在第 5 节讨论 L^2 采样器的理论分析，此处重点关注计算方面。

计算 $\mathcal{L}(g)$ 的主要挑战在于求取密度函数 f，因为它通常以未归一化的形式 $f(x) = e^{-E(x)}/U$ 给出，其中 U 是未知的

归一化常数。幸运的是， U 可以通过重要性采样高效估计

$$U = \int \exp\{-E(x)\}\mathrm{d}x = \mathbb{E}_{X\sim h(x)} \exp\{-E(X) - \log h(X)\},$$

其中， $h \approx f$ 是一个接近目标分布的提议分布，精确期望可通过蒙特卡洛样本近似。在 TemperFlow 中，我们取 $f = r_k$ 和 $h = r_{k-1}$，因此，只要退火分布流 $\{r_k\}$ 中相邻分布彼此接近， U 的估计量就不会受到高方差的影响。于是我们得到

$$\begin{aligned}\mathcal{L}(g) &= \int [g(x) - f(x)]^2\mathrm{d}x = \int [g(x)]^2\mathrm{d}x - 2\int f(x)g(x)\mathrm{d}x + \int [f(x)]^2\mathrm{d}x \\ &= \mathbb{E}_{X\sim g(x)}\,[g(X) - 2f(X)] + \int [f(x)]^2\mathrm{d}x,\end{aligned}$$

其中，最后一项不依赖于 g。完整算法见算法3。

算法 3 TemperFlow 采样器。输入：目标分布 f，其能量函数为 $E(x)$；具有初始参数值 $\theta^{(0)}$的可逆神经网络 T_θ；批量大小 M；步长 $\{\alpha_k\}$ ˆ 输出：神经网络参数 θ $T_{\hat{\theta}\sharp}\mu_0 \approx f$，使得：令 ( ) 为 $T_{\theta^{(0)}\sharp}\mu_0$

1 h x

2 Z_1 ... $X_i \leftarrow T_{\theta^{(0)}}$ Z_i $i = 1$ ... M 的密度函数：生成 , ,

3 $U \leftarrow M^{-1}\sum_{i=1}^{M}$ $\{-E$ X_i $-$ h X_i $\}$

$Z_M \overset{iid}{\sim} \mu_0$ 并设置 ( ), , ,：设置 ˆ exp ( ) log ( ) 4：对于 $k = 1, 2, \ldots$ 执行 5：令 $g_\theta(x)$ 为 $T_{\theta\sharp}\mu_0$ 的密度函数 6：生成 $Z_1, \ldots, Z_M \overset{iid}{\sim} \mu_0$ 并定义 $X_i = T_\theta(Z_i)$, $i = 1, \ldots, M$ 7：定义 $L(\theta) = M^{-1}\sum_{i=1}^{M}\left[g_\theta(X_i) - 2\exp\{-E(X_i)\}/\hat{U}\right]$ 8：设置 $\theta^{(k)} \leftarrow \theta^{(k-1)} - \alpha_k\nabla_\theta L(\theta)|_{\theta=\theta^{(k-1)}}$ 9：如果 $L(\theta)$ 收敛，则 10：返回 $\hat{\theta} = \theta^{(k)}$ 11：结束 if 12：结束 for

备注。 在实际实现中，优化 $\mathcal{L}(g)$ 的对数有助于增强数值稳定性，尤其是在 f 的维度较高时。在这种情况下，我们

可以再次应用重要性采样以获得

$$\log[\mathcal{L}(g)] = \log\left[\mathbb{E}_{X\sim h(x)}\frac{[g(X) - U^{-1}\exp\{-E(X)\}]^2}{h(X)}\right] \approx \log\left[M^{-1}\sum_{i=1}^{M}\exp\{W_i\}\right],$$

$$\begin{aligned} W_i &= 2\log\left|g(X_i) - U^{-1}\exp\{-E(X_i)\}\right| - \log h(X_i) \\ &= 2\log g(X_i) - \log h(X_i) + 2\log\left|1 - U^{-1}\exp\{-E(X_i) - \log g(X_i)\}\right|, \end{aligned}$$

其中 $X_i \overset{iid}{\sim} h(x)$。

4.3 自适应温度选择

仍然存在如何选择 β_k （8）中的序列的问题。显然，使用稠密的 β_k 会使从 r_{k-1} 到 r_k 的传输更加容易，但也会增加计算时间。另一方面，使用由 β_k 个点组成的粗网格会减少转移次数，但会导致更困难的传输。

对于并行回火马尔可夫链蒙特卡罗，关于并行链温度选择的讨论十分广泛，其中一种广泛使用的准则是使不同链之间交换的接受概率保持一致（Kofke，2002；Kone and Kofke，2005；Earl andDeem，2005）。根据正态分布的计算，这会产生一个几何间隔温度阶梯；此外，还有一些自适应方案可动态调整温度（Vousden et al.，2016）。然而，这些启发式方法并不能直接应用于 TemperFlow，因为我们的框架中没有接受概率。

对于 TemperFlow，我们设计了一种自适应方案，以根据当前的β_k确定 β_{k+1} 。作为起点，定理3指出，回火曲线会随着时间推移降低 KL散度，因此，稳定优化过程的一种自然方式是选择能够平稳降低 KL散度的 β_k。令 $\ell(\beta)$ =KL($q_\beta\|\mu$)，定理3表明，$\ell(\beta)$是 β的递减函数。假设给定 r_k 和 β_k，那么一种自然的 β_{k+1} 选择方式是使得 $\ell(\beta_{k+1}) = \alpha\ell(\beta_k)$，其中 $0 < \alpha < 1$ 是预先设定的折扣

因子，例如例如，0.9。当然， $\ell(\beta_{k+1})$是未知的，但通过泰勒展开可得

$$\ell(\beta_{k+1}) - \ell(\beta_k) \approx \left.\frac{\mathrm{d}\ell(\beta)}{\mathrm{d}(\log\beta)}\right|_{\beta=\beta_k} (\log\beta_{k+1} - \log\beta_k)\,.$$

随着 t 通过重新参数化 γ =log(β) 并重新排列各项，我们得到

$$\gamma_{k+1} \approx \gamma_k - (1-\alpha)\left\{\frac{\mathrm{d}[\log\ell(e^{\gamma})]}{\mathrm{d}\gamma}\right\}^{-1}\Bigg|_{\gamma=\gamma_k}. \tag{9}$$

在补充材料中，我们展示了

$$\frac{\mathrm{d}[\log\ell(e^{\gamma})]}{\mathrm{d}\gamma} = \frac{-\beta(1-\beta)\left\{\int q_\beta(x)E^2(x)\mathrm{d}x - \left[\int q_\beta(x)E(x)\mathrm{d}x\right]^2\right\}}{\int q_\beta(x)\left[\log q_\beta(x) + E(x)\right]\mathrm{d}x + \log\int q_\beta(x)\exp\left[-E(x) - \log q_\beta(x)\right]\mathrm{d}x}, \tag{10}$$

因此（9）可以基于 r_k 和 β_k进行估计。计算 β 的总体方法见算法4。

算法4 TemperFlow采样器中 β 参数的自适应选择。

输入： 当前 β_k 和神经网络参数值 $\theta^{(k)}$，批量大小 M，折扣因子 α

输出： 下一个参数 β_{k+1}

1 $Z_1 \ldots X_i = T_{\theta^{(k)}} Z_i \quad i = 1 \ldots M$: 生成，

2 $c_1 \leftarrow M^{-1}\sum_{i=1}^{M} E\ X_i \quad c_2 \leftarrow M^{-1}\sum_{i=1}^{M}[E\ X_i\ ]^2$

，$Z_M \overset{iid}{\sim} \mu_0$并令 ()，，，：设置 ()，()3: 计算 U_i =log p_θ(X_i) + E(

X_i)，$i = 1$，.. .，M4: 设置 $c_3 \leftarrow M^{-1}\sum_{i=1}^{M} U_i$, $c_4 \leftarrow$log

$\left[M^{-1}\sum_{i=1}^{M} e^{-U_i}\right]$5: 设置 $\gamma^{(k+1)} \leftarrow \log\beta_k + (1-\alpha)\beta_k^{-1}(1-\beta_k)^{-1}(c_3 + c_4)/$

($c_2 - c_1^2$)6: **返回** β_{k+1} =exp{$\gamma^{(k+1)}$}

5 TemperFlow 的收敛性质

如第 4.1 节所述，退火分布流 $\{r_k\}$可视为回火曲线 ρ_t的离散化版本，其收敛性质由定理 3给出。为了连接序列 $\{r_k\}$中的相邻分布，我们在第 L^2 4.2 节提出了采样器，它是研究 TemperFlow 理论性质的关键。关于平方 L^2距离的 Wasserstein 梯度流已

被用于从数据中估计生成模型（Gao 等人，2022），下面我们将展示，它也可用于分析 L^2 采样器。具体而言，我们从理论上证明，这种 Wasserstein 梯度流不存在梯度消失问题。

令 g_t 为关于 $\mathcal{L}$ 的 Wasserstein梯度流。假设目标概率测度 f 和每个 g_t 都关于勒贝格测度 λ 绝对连续，因此它们存在密度。为简单起见，我们用相同的符号 f 或 g_t 同时表示概率测度和密度函数。假设 f 和 g_t 可微、支撑集为 $\mathbb{R}^d$，且属于 $L^2(\mathbb{R}^d)$ 空间，因此 $\int f^2 \mathrm{d}\lambda < \infty$ 和 $\int g_t^2 \mathrm{d}\lambda < \infty$。用 $\mathcal{Q}$ 表示 $\mathbb{R}^d$ 中球的集合；对于可测集 $A \subset \mathbb{R}^d$，令$|A| = \lambda(A)$。为分析 g_t 的性质，我们首先作出以下假设：

假设 1. 对于每个 $\varepsilon \in (0, \varepsilon_0]$，其中 $\varepsilon_0 < 1$ 是一个固定常数，都存在互不相交的球 $Q_1^\varepsilon, \ldots, Q_K^\varepsilon \in \mathcal{Q}$，它们可能取决于 ε，使得对于所有 $i = 1, \ldots, K$：

(a) $\inf_{Q_i^\varepsilon} f \le \varepsilon^{1/4}$且 $\int_{Q^\varepsilon} f \mathrm{d}\lambda \ge 1 - \varepsilon^{1/4}$，其中 $Q^\varepsilon = \bigcup_{i=1}^K Q_i^\varepsilon$。 (b) $\sup_{\substack{Q \subset Q_i^\varepsilon \\ Q \in \mathcal{Q}}} |Q|^{2/d} \left[\frac{1}{|Q|} \int_Q f^{-2} \mathrm{d}\lambda \right]^{1/2} \le M_1 \varepsilon^{-1/4} [\log(1/\varepsilon)]^c$ 对于某个 c， $M_1 > 0$成立。

(c) $\int_{Q_i^\varepsilon} \left[\varepsilon^{1/4} - f(x) \right]_+^2 \mathrm{d}x \ge M_2 \varepsilon^{3/4} [\log(1/\varepsilon)]^c$ 对于充分大的 $M_2 > 0$成立。

直观地说，假设1（a）意味着我们可以使用 K 个球覆盖目标分布的众数，而球外的概率质量极少。假设1（b）和（c）是关于每个众数尾部行为的技术条件。作为起点，我们首先证明标准正态分布满足假设1。

命题 2. 令 $f(x) = (2\pi)^{-1/2} e^{-x^2/2}$ 为标准正态分布。则对于某个 $\varepsilon_0 < 1$ 以及每个 $\varepsilon \in (0, \varepsilon_0]$, $Q^\varepsilon = [-x_\varepsilon - M\varepsilon^{1/12}, x_\varepsilon + M\varepsilon^{1/12}]$ 满足假设*1*，其中 $x_\varepsilon = \sqrt{-(\log \varepsilon)/2 - \log(2\pi)}$ 且 M 是一个足够大的常数。

命题2可以轻松推广到更复杂的分布。事实上，假设1 主要关注分布的尾部行为，因此密度

函数可以在紧集内进行修改，而不违反这些假设。此外，对于多峰分布，只要各个球彼此不相交，就可以用大小合适的球覆盖每个众数；或者根据密度函数的形状，用一个球覆盖多个众数。

假设2。 存在一个常数 $\alpha > 0$ ，使得对于所有 $g_t(x) \geq \alpha f(x)$ 以及 $x \in Q^{\varepsilon}$ 和 $t > 0$, 均成立。

假设2 实质上要求采样器密度与目标密度具有相同的支撑集。在退火分布流中，这一假设是自然的，因为在这种情况下， g_t 通常是比 f更平坦的分布。在这些条件下，我们证明 $\mathcal{L}(g_t)$ 的梯度不会消失。

定理 4。 假设*1* 和*2* 成立。则存在一个常数 $C > 0$，使得当 $|\mathrm{d}\mathcal{L}(g_t)/\mathrm{d}t| := \varepsilon_t \leq \varepsilon_0$时，我们有

$$\left|\frac{\mathrm{d}\mathcal{L}(g_t)}{\mathrm{d}t}\right| = \varepsilon_t \geq C \cdot \max_{i=1,\ldots,K} \left\{ \left[\int_{Q_i^{\varepsilon_t}} (g_t - f)^2 \mathrm{d}\lambda \right]^2 \cdot \min(|Q_i^{\varepsilon_t}|, |Q_i^{\varepsilon_t}|^{-1}) \right\}. \tag{11}$$

命题2表明， $(|Q_i^{\varepsilon_t}|$ 项在（11）中通常为 $O(\sqrt{\log(1/\varepsilon_t)})$ 的数量级，因此在大多数情况下可以将其视为常数。此外，由于根据假设1， $Q_i^{\varepsilon_t}$的并集几乎涵盖了 f 的全部概率质量，（11）的右端可以视为 L^2距离的平方，即 f 与 g_t 之间的平方距离，差一个常数因子。总体而言，定理4表明， L^2 Wasserstein梯度流具有梯度不消失性质。

为直观展示基于 KL 的采样与基于 L^2 的采样之间的差异，我们使用这两种方法对图 p_m 1 中的双峰分布进行采样，并将初始采样器分布设置为 $0.9 \cdot N(1, 1) + 0.1 \cdot N(8, 0.25)$。如图2所示，KL采样器的行为符合定理 2的预测，即使经过 100 次迭代，进展仍然非常有限。相反， L^2 采样器在 50 步后便成功估计出目标分布。

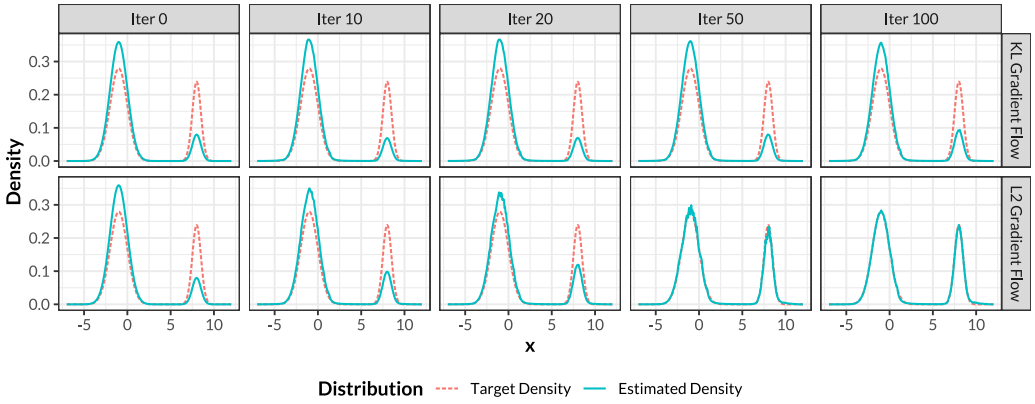


图 2：基于 KL 和 L^2 的采样器比较。目标密度为$p_m(x)$ 见第 2.3 节，初始采样器分布与 $p_m(x)$ 具有相同的成分，但混合概率不同。每一列代表优化过程中的一次迭代。

6 模拟研究

本节通过模拟实验，将 TemperFlow 与广泛使用的 MCMC采样器进行比较，包括 Metropolis–Hastings算法（Metropolis等人，1953；Hastings，1970）、哈密顿蒙特卡罗（Brooks 等人，2011）以及并行回火。我们还考虑了 TemperFlow算法的一个变体，该变体通过拒绝采样对 TemperFlow 样本进行后处理优化。由于我们关注的是多峰分布，因此首先考虑若干个易于可视化其众数的二元高斯混合模型，然后基于 Copula 研究具有任意维度的多元分布。算法超参数设置的详细信息见补充材料第 A.2.2 节。用于从其他方面验证所提出方法的补充实验见第 A.3节。

6.1 高斯混合模型

我们研究了图3最后一列所示的三个高斯混合模型。对于每种方法，我们根据生成点与 Gretton等人（2006）提出的1-Wasserstein距离和最大均值差异（MMD）来计算采样误差。

真实样本之间的距离。我们将这些指标称为调整后的瓦瑟斯坦距离和调整后的最大均值差异，其精确定义见补充材料第A.2.1 节。这两个指标可以取负值，数值越小表示采样质量越高。图3 展示了一次模拟运行的结果。

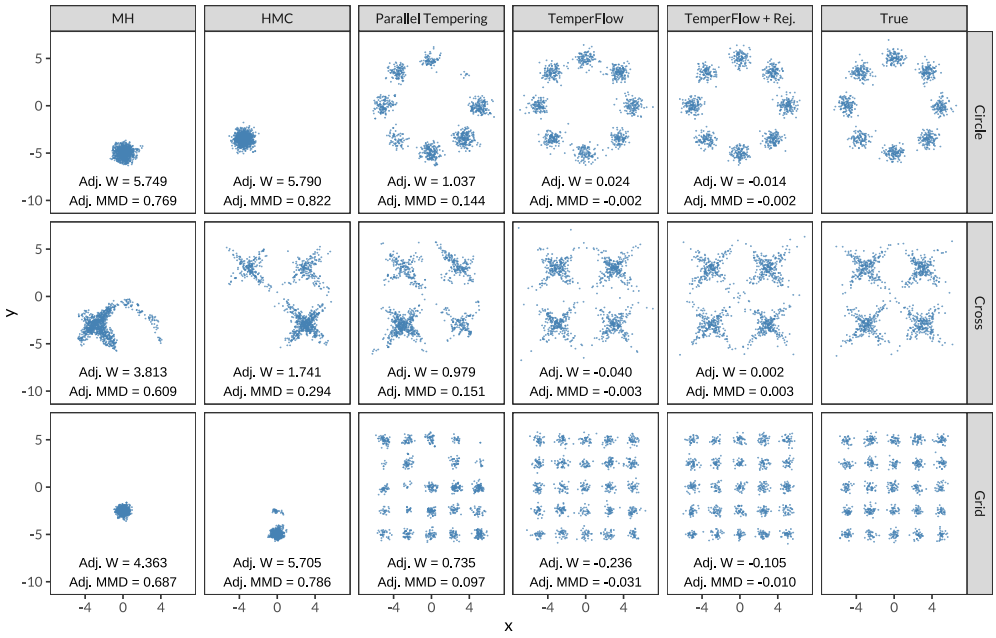


图3：三种高斯混合模型上不同采样方法的比较。MH 表示 Metropolis–Hastings 马尔可夫链蒙特卡罗方法，HMC 表示哈密顿蒙特卡罗，而“TemperFlow +Rej.”表示通过拒绝采样对 TemperFlow 样本进行后处理。每个面板下方的数字分别表示生成点与真实样本之间调整后的瓦瑟斯坦距离和调整后的最大均值差异。

显然，对于 Metropolis–Hastings 和哈密顿蒙特卡罗等基本马尔可夫链蒙特卡罗方法，采样结果会严重遗漏大多数众数。并行回火大幅提升了样本质量，但仍难以使各个模态的比例达到正确值。相比之下，TemperFlow 能够成功捕获所有众数，而通过拒绝采样进行细化还可以进一步提升样本质量。为考虑采样中的随机性，我们对每种方法和每个目标分布分别重复实验 100 次，并在图 4 中总结其采样误差。结果与图 3 一致：在所有方法中，TemperFlow 及其细化版本的误差始终小于其他方法。

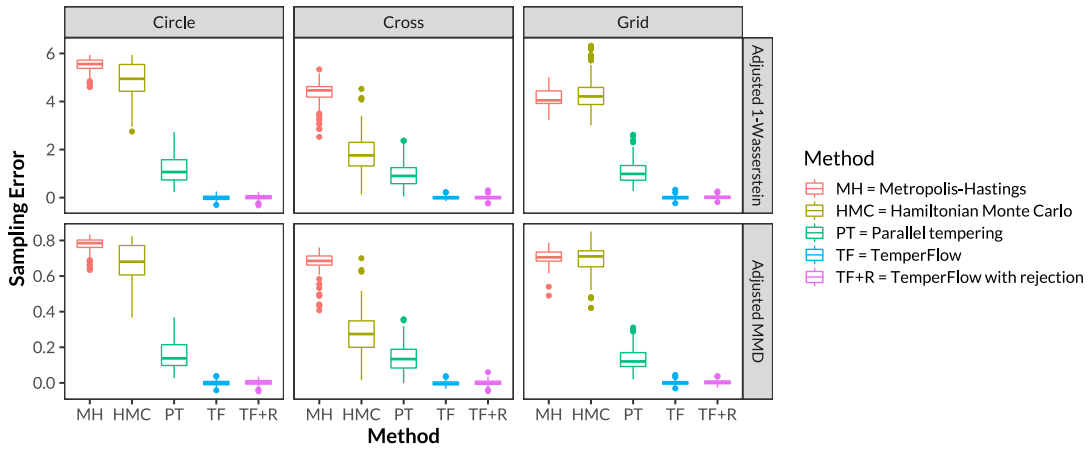


图4：不同方法在三种正态混合分布上的采样误差比较。箱线图基于100次模拟运行绘制。

6.2 Copula 生成的分布

为了研究更一般的多峰分布，我们使用 Copula 来定义具有任意维度的多元密度。关于 Copula 建模的介绍，请参见 Nelsen（2006）。具体而言，我们将目标分布函数定义为 $F(x_1, \ldots, x_d) = C(F_1(x_1) \ldots, F_d(x_d))$，其中 $F_i(x)$ 是每个分量 X_i的边际分布函数，而 $C(u_1, \ldots, u_d)$是Copula函数。在我们的实验中，我们令前 s 个边际分布为正态分布混合，$F_i \sim 0.7 \cdot N(-1, 0.22) + 0.3 \cdot N(1, 0.22), i = 1, \ldots, s$，其余边际分布为正态分布 $N(0, 0.25)$。函数$C(u_1, \ldots, u_d) = (u_1^{-\theta} + \cdots + u_d^{-\theta} - d + 1)^{-1/\theta}$是一个 Clayton Copula，其中 $\theta = 2$。图5展示了 (X_i, X_j) 的散点图和密度图，$i, j \leq s, i \neq j$。

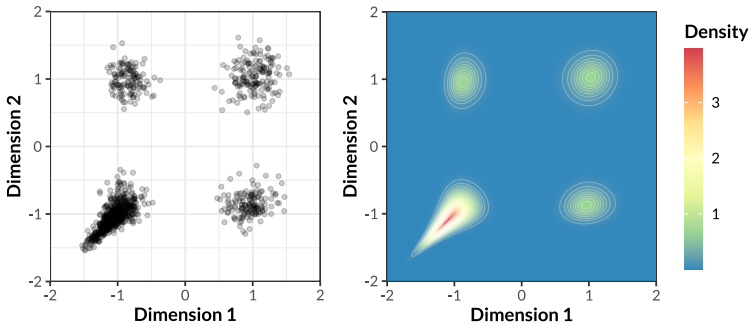


图5：与 Clayton Copula 结合的正态混合边际分布。

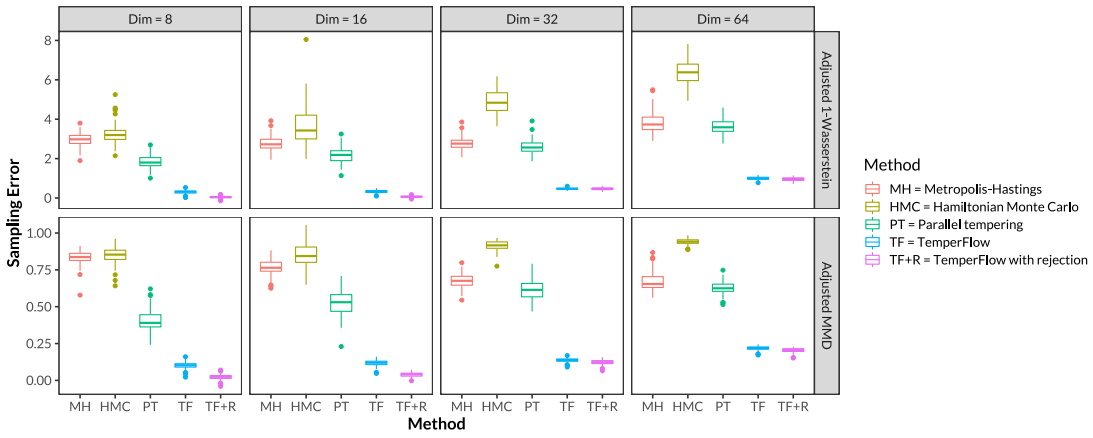


图 6：不同方法在由 Copula 生成的分布上的采样误差。

在这种设计下，目标分布 $F(x_1, \ldots, x_d)$ 有 2^s 个众数，其数量随 s 呈指数增长，而且每一对 (X_i, X_j) 都是相关的。因此，即使 s 和 d 适中，这也是一种极具挑战性的采样分布。我们首先通过计算 F 的偏导数，推导出密度函数 $f(x_1, \ldots, x_d)$，然后使用不同的方法从 f 中进行采样，其中 $s = 8$ 和 $d = 8$,16,32,64。也就是说，每个分布总共有 256 个众数，维度各不相同。

与双变量实验类似，我们计算真实样本与生成样本之间调整后的瓦瑟斯坦距离和调整后的最大均值差异，并根据 100 次重复实验汇总比较结果，如图6 所示。补充材料中的图 S5 还展示了并行回火和 TemperFlow 生成样本的成对散点图和密度等高线图。正如预期，TemperFlow 的样本质量良好，接近真实值；而其他方法则因众数数量庞大且维度较高而面临严重问题。

当然，TemperFlow 和马尔可夫链蒙特卡罗都是迭代算法，因此它们的采样误差会受到迭代次数的影响，进而影响总体计算时间。我们在补充材料的第 A.3.4 节中报告了更多比较结果，并在表 S1 中展示了不同算法的计算时间。我们指出，马尔可夫链蒙特卡罗只有生成成本，而 TemperFlow 还具有训练成本。该

主要区别在于，马尔可夫链蒙特卡罗的生成成本与马尔可夫链迭代的次数成正比，而对于TemperFlow 而言，生成成本几乎是固定的，并且极低。相反，TemperFlow 的训练成本会随优化迭代的次数呈线性增长。

7 应用：深度生成模型

现代深度学习技术吸引了统计学研究人员和实践者的极大关注，其中深度生成模型是一类重要的无监督学习方法（Salakhutdinov，2015；Bond-Taylor 等人，2021年）。深度生成模型试图利用深度神经网络对高维数据的统计分布进行建模，并广泛应用于图像合成、文本生成等领域。

一类通用的深度生成模型具有如下形式： $X = G(Z)$，其中 $X \in \mathbb{R}^p$ 是高维数据点，例如图像； $Z \in \mathbb{R}^d$ 是具有 $d \ll p$ 的潜在随机向量； $G : \mathbb{R}^d \to \mathbb{R}^p$ 是一个深度神经网络（DNN），通常称为生成器。Z 的分布由能量函数 $E : \mathbb{R}^d \to \mathbb{R}$ 刻画，该函数也是一个 DNN。从这个意义上说，$(E_d, G_{d,p})$ 这一对定义了一个深度生成模型，其中下标 d 和 p 表示维度。关于 $(E_d, G_{d,p})$ 的建模与估计，已有大量文献讨论；更多详情请参见 Che 等人 (2020)；Pang 等人 (2020)。在本节中，我们将函数 E 和 G 视为已知，并重点讨论对 $p(z) \propto \exp\{-E(z)\}$ 的采样，因为这是生成 X 的新数据点的关键。

我们首先考虑 Fashion-MNIST数据集的生成模型（Xiao 等人，2017年）。Fashion-MNIST数据包含一个由 60,000 张图像组成的训练集和一个由 10,000 张图像组成的测试集，每张图像由 28×28 灰度像素组成。图 7 左侧的图显示了来自该数据集的 60 张图像。然后，我们基于现有文献（（E_2，$G_{2,\ 784}$）构建一个 $d = 2$深度生成模型，其Che 等人，2020），生成的图像显示在图7右侧的图中。

由于生成图像需要独立样本，测度传输采样

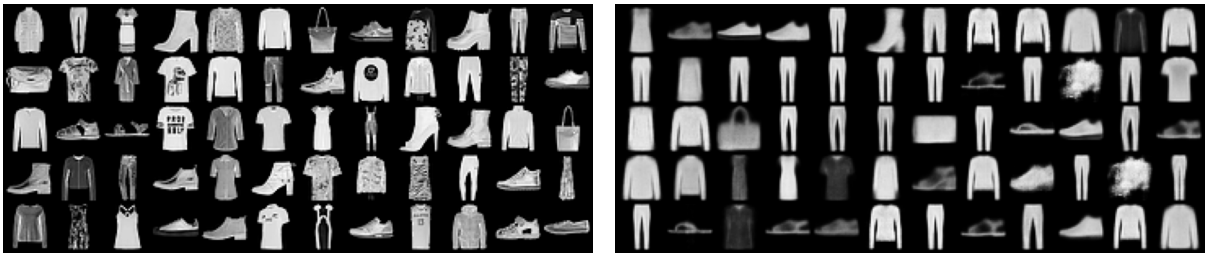


图7：左：来自 Fashion-MNIST 数据集的真实图像。右：由潜在维度为 $d = 2$ 的深度生成模型生成的图像。

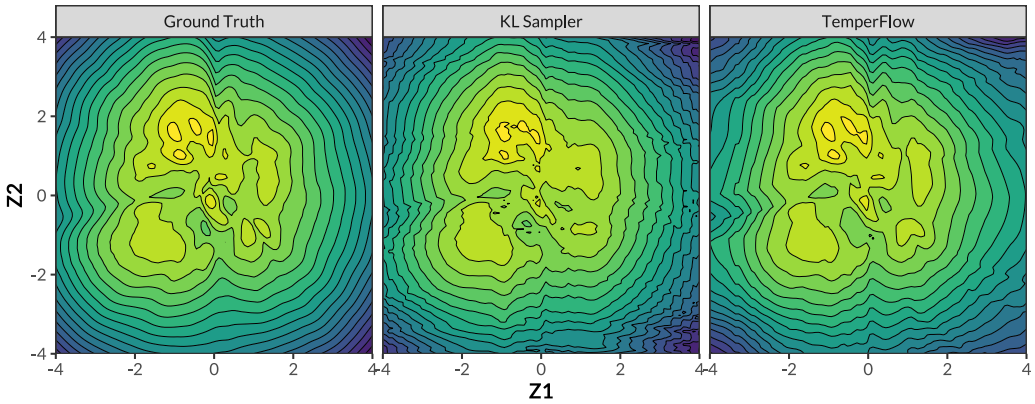


图 8：真实对数密度函数（左）、KL采样器学习到的函数（中）以及 TemperFlow 学习到的函数（右）的等高线图。

这种方法对于此任务尤其有用。为此，我们比较现有的 KL采样器与我们提出的 TemperFlow 在从 $p(z)$ 中采样的表现。训练完两个采样器后，我们通过图 $\log p(z)$ 中的等高线图将它们估计的对数密度函数可视化 8，并将其与真实值进行比较。

显然，真实对数密度函数 $\log p(z)$ 包含许多孤立众数，但 KL采样器只能捕捉其中少数几个。相比之下，TemperFlow 给出的分布非常接近真实值。随着潜在维度 d 增加，这种差异会更加明显。我们拟合了另一个生成模型 $(E_{32}, G_{32,784})$，其潜在维度 $d = 32$更大，并在图 9左侧的图中展示生成的图像。与 $d = 2$ 情形类似，我们使用能量函数 $E(z)$训练 KL采样器和 TemperFlow 采样器。在更高维的情形下，我们无法再直接可视化密度函数。相反，我们模拟潜变量



图9：左：由具有潜在维度 $d = 32$的深度生成模型生成的图像。中：使用KL采样器从模型中采样的图像。右：使用TemperFlow采样器采样的图像。

$Z_1, \ldots, Z_{100} \in \mathbb{R}^{32}$，分别从两个采样器中采样，并将其输入生成器 G 以生成图像 $X_1 = G(Z_1) \ldots, X_{100} = G(Z_{100}) \in \mathbb{R}^{784}$。图 9 中间和右侧的图分别展示了 KL采样器和 TemperFlow 生成的图像。显然，KL采样器只能生成类似外套和类似裤子的图像，这表明它丢失了真实分布中的许多众数。另一方面，TemperFlow 成功保留了 $p(z)$的主要众数，进一步验证了所提出的采样器具有优越性能。

最后，我们展示一个使用 CelebA数据集训练的大型生成模型（Liu 等人，2015）。CelebA 中的每个数据点都是一张尺寸为 64×64 的彩色人脸图像，我们从原始数据集中选取一个子集，构成以下四个类别：戴眼镜的女性（2677 张图像）、不戴眼镜的女性（20000 张图像）、戴眼镜的男性（10346 张图像）以及不戴眼镜的男性（20000 张图像）。一个深度生成模型 $(E_{100}, G_{100,\ 12288})$ 在此子集上进行训练，图 10(a) 展示了该模型生成的图像。自然，潜在能量函数 E 会反映这四个主要模态，我们通过观察 KL采样器、MCMC采样器和 TemperFlow 生成的图像来测试它们的性能。

图10(b)-(f)展示了结果。显然，KL采样器几乎只输出一个类别，即不戴眼镜的女性，而TemperFlow则涵盖了全部四个类别。另一方面，马尔可夫链蒙特卡罗样本几乎无法生成清晰可见的人脸。总而言之，所有示例



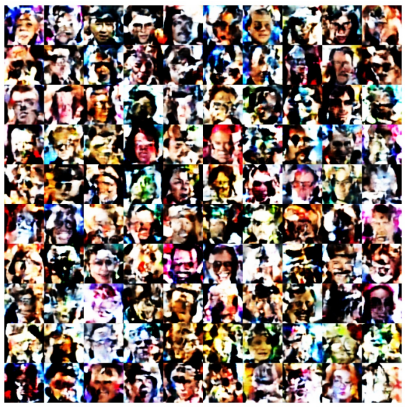
(a) 真实模型样本



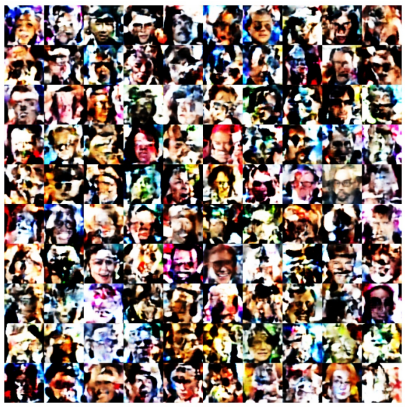
(b) KL采样器



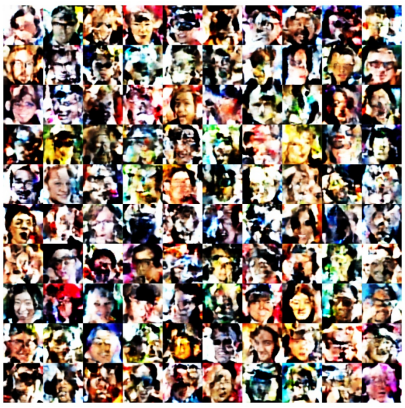
(c) TemperFlow



(d) Metropolis–Hastings



(e) 哈密顿蒙特卡罗



(f) 并行回火

图10： (a)使用潜在维度 $d = 100$的深度生成模型生成的图像，该模型从CelebA数据集训练得到。(b)-(f)：使用不同采样器从模型中采样得到的图像。

提出了一个 上述结果展示了 TemperFlow 采样的准确性和可扩展性 r.

8 结论与讨论

本文提出了一种通用采样器 TemperFlow，用于从具有挑战性的统计分布中采样。TemperFlow 的灵感源自Marzouket al.(2016) 提出的测度输运框架，但克服了该框架在目标分布呈多峰时的关键弱点。我们使用 Wasserstein梯度流理论对二者的根本差异进行了严格分析，并通过各种数值实验加以清晰展示。

TemperFlow 的一个有趣特性是，它将采样问题转化为一系列优化问题，这带来了以下影响。首先，许多现代优化技术都可以直接用于加速传输映射的训练。其次，通常而言，诊断优化问题的收敛情况比诊断马尔可夫链的收敛情况更容易。第三，由于我们主要依赖基于梯度的优化方法，TemperFlow的计算具有高度并行性，并且可以从现代计算硬件（如图形处理器）中大受裨益。所有这些方面都体现了 TemperFlow 在计算效率方面的巨大优势。

作为一种通用采样方法，TemperFlow可以在许多方面与马尔可夫链蒙特卡罗进行比较，但最明显的差异之一是，TemperFlow具有训练阶段，而马尔可夫链蒙特卡罗则直接生成随机变量。这一差异有助于指导我们在需要采样时选择使用哪种方法。具体而言，当需要大量相互独立的随机变量时，TemperFlow尤其有用；另一方面，在计算持续变化的分布的期望时，马尔可夫链蒙特卡罗可能更合适，例如在蒙特卡洛期望最大化算法（Wei和Tanner，1990年；Levine和Casella，2001）中。为此，我们并不将TemperFlow定位为马尔可夫链蒙特卡罗的替代方法。相反，我们将其视为马尔可夫链蒙特卡罗和其他基于粒子的方法的补充。我们预计，这些方法可以结合起来开发新的高效采样器；我们将其留作未来的研究方向。

另一个有前景的方向是评估模型在观测数据上的拟合优度。令 X_1，...， $X_n \overset{iid}{\sim} q$（$x$）是来自TemperFlow采样器的随机样本。回顾一下，目标密度为 p（x） $\propto e^{-E(x)}$。我们可以进行单样本拟合优度检验H_0： $q = p$。其中的主要困难源于复杂的高维模型，以及 p（x）的未知归一化常数。传统的拟合优度检验，例如χ^2检验和柯尔莫哥洛夫–斯米尔诺夫检验，几乎无法应用。基于斯坦因差异的非参数单样本检验由Ch wialkowski等人（2016）进行了研究。请注意

，TemperFlow采样器的拟合优度检验具有一些不同于先前问题的特殊特征。特别是，给定归一化流模型后，密度 q 具有显式公式。这一特征使我们能够提出更强大的检验，例如基于Fisher散度的检验。Fisher散度比许多其他散度更强，例如全变差、海灵格距离和Wasserstein距离（Ley和Swan，2013）。

A 附录

A.1 可逆神经网络

可逆神经网络（INNs）是一类相对于其输入可逆的神经网络模型。INN 可以被视为一种映射 $T:\ \mathbb{R}^d \to \mathbb{R}^d$ 使得 T^{-1} 存在且能够被高效计算，并且 T 通常是更简单映射的组合，$T = T_K \circ T_{K-1} \circ \cdots \circ T_1$，其中每个 T_i 都是可逆的。INNs 主要用于实现归一化流模型（Tabak 和 Vanden-Eijnden，2010；Tabak 和 Turner，2013；Rezende 和 Mohamed，2015），这类模型可以描述为通过一系列可逆映射对概率密度进行变换。

归一化流最初被开发为非参数密度估计器（Tabak 和 Vanden-Eijnden，2010；Tabak 和 Turner，2013），当时所使用的映射是表达能力有限的简单函数。归一化流被引入深度学习领域后，研究者开发了许多强大的基于可逆神经网络的模型，包括仿射耦合流（Dinh 等人，2014；Dinh 等人，2016）、掩码自回归流（Papamakarios 等人，2017年）、逆自回归流（Kingma 等人，2016）、神经样条流（Durkan 等人，2019）以及线性有理样条流（Dolatabadi 等人，2020）等。值得一提的是，归一化流中的“流”这一术语与梯度流中的“流”在概念上存在差异，后者正是本文的关注重点。因此，为避免歧义，本文使用“可逆神经网络”来指代归一化流模型，尽管还存在其他形式的归一化流，例如基于常微分方程的连续归一化流（Chen 等人，2018），以及使用自由形式连续动力学的扩展形式（Grathwohl 等人，2019）。

以逆自回归流为例。它是由一系列可逆映射复合而成的，$T = T_K \circ T_{K-1} \circ \cdots \circ T_1$。对于每个 $i = 1, \ldots, K$，令 $x = (x_1, \ldots, x_d)'$

为输入向量，并将 $y=(y_1,\ldots,y_d)'=T_i(x)$ 记为输出向量。那么 y 具有以下形式：

$$
\begin{aligned}
y_1 &= \mu_{i1} + \sigma_{i1} x_1, \\
y_j &= \mu_{ij}(x_{1:j-1}) + \sigma_{ij}(x_{1:j-1}) x_j, \quad j = 2,\ldots,d,
\end{aligned}
$$

其中，$x_{1:r}$ 表示（$x_1,\ldots,x_r$）′，μ_{i1}、σ_{i1} 是标量，而 μ_{ij},σ_{ij}：$\mathbb{R}^{j-1}\to\mathbb{R}$ 是用于 $j\geq 2$ 的神经网络。很容易验证 T_i 是可逆的，因为

$$
\begin{aligned}
x_1 &= (y_1-\mu_{i1})/\sigma_{i1}, \\
x_2 &= (y_2-\mu_{i2}(x_1))/\sigma_{i2}(x_1) = (y_2-\mu_{i2}((y_1-\mu_{i1})/\sigma_{i1}))/\sigma_{i2}((y_1-\mu_{i1})/\sigma_{i1}), \\
x_j &= (y_j-\mu_{ij}(x_{1:j-1}))/\sigma_{ij}(x_{1:j-1}), \quad j=3,\ldots,d,
\end{aligned}
$$

其中，每个 x_j 都是 y_j 和 $x_{1:j-1}$ 的函数，而 $x_{1:j-1}$ 可以递归地化简为 $y_{1:j-1}$ 的函数。此外，如果神经网络 $\mu_{ij}(\cdot)$ 和 $\sigma_{ij}(\cdot)$ 可微（使用平滑激活函数即可轻松实现），那么每个 T_i 以及整个 T 映射也都是可微的。从这个意义上说，在非常宽松的条件下，INNs 从设计上就是微分同胚。

此外，大多数 INNs 都具有一个理想特性：雅可比矩阵 ∇T 是三角矩阵，因此其行列式 $\det\nabla T$ 只是对角元素的乘积。再次使用上面的示例，我们可以发现，

$$
\nabla T = \begin{pmatrix}
\frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_d} \\
\frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_d} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial y_d}{\partial x_1} & \frac{\partial y_d}{\partial x_2} & \cdots & \frac{\partial y_d}{\partial x_d}
\end{pmatrix} = \begin{pmatrix}
\sigma_{i1} & 0 & \cdots & 0 \\
* & \sigma_{i2}(x_1) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
* & * & \cdots & \sigma_{id}(x_{1:d-1})
\end{pmatrix},
$$

因此 $\det\nabla T=\sigma_{i1}\cdot\sigma_{i2}(x_1)\cdots\sigma_{id}(x_{1:d-1})$，且当 $\sigma_{ij}>0$ 时，$\det\nabla T>0$ 自动成立。

在实践中，会在每一对 T_k 和 T_{k-1} 之间插入一个置换算子 $P_\pi(x_1,\ldots,x_d) = (x_{\pi(1)},\ldots,x_{\pi(d)})'$，其中 $(\pi(1)\ldots,\pi(d))$ 是 $(1,\ldots,d)$ 的一个置换。这是因为 T_{ij} 是给定 $x_{1:j-1}$ 时 x_j 的仿射映射；如果变量的顺序不变，T 的表达能力就会受到限制。

A.2 数值实验的详细信息

A.2.1 指标的定义

给定数据点 $X = (X_1,\ldots,X_n)$ 和 $Y = (Y_1,\ldots,Y_m)$，定义 X 和 Y 之间的离散1-Wasserstein距离为

$$W(X,Y) = \min_{P\in\Pi} \langle P, C\rangle, \quad \Pi = \{P \in \mathbb{R}_+^{n\times m} : P\mathbf{1}_m = n^{-1}\mathbf{1}_n, P'\mathbf{1}_n = m^{-1}\mathbf{1}_m\},$$

其中，$C = (C_{ij})$ 是一个具有 $C_{ij} = \|X_i - Y_j\|_1$ 的 $n\times m$矩阵。此外，定义 X 和 Y 之间的经验MMD为

$$\mathrm{MMD}(X,Y) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K(X_i,X_j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i} K(Y_i,Y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K(X_i,Y_j),$$

其中，$K(\cdot,\cdot)$: $\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ 是一个正定核函数。

然后，给定目标分布 $p(x)$ 以及由某种采样算法生成的点 $X = (X_1,\ldots,X_n)$，定义 X 和 p 之间的调整后的1-Wasserstein距离和调整后的最大均值差异，如下：

$$\tilde{W}(X,p) = W(X,Y) - W(Y,\tilde{Y}),$$

$$\tilde{\mathrm{MMD}}(X,p) = \mathrm{MMD}(X,Y) - \mathrm{MMD}(Y,\tilde{Y}),$$

$\tilde{}$分别，其中 Y，$Y\sim p(x)$ 是来自真实 $p(x)$ 分布的两个大小为 n 的独立随机样本。

A.2.2 超参数设置

高斯混合模型 对于所有 MCMC采样器，初始状态均由标准正态分布生成，前 200 步作为预热期丢弃，随后收集 1000 个点作为样本。Metropolis–Hastings 使用标准差为 $\sigma_{\mathrm{MH}} = 0.2$ 的正态分布进行随机游走。哈密顿蒙特卡罗使用步长 $\varepsilon = 0.2$ ，每次迭代执行五个蛙跳步。并行回火方法使用五条并行链，逆温度参数在对数刻度上从 0.1 到 1 等距分布。对于每个逆温度 β，并行回火使用标准差为 $\sigma_{\mathrm{PT},\ \beta} = 0.2/\sqrt{\beta}$ 的正态分布进行随机游走。

TemperFlow 由标准正态分布和 $\beta_0 = 0.1$初始化。自适应 β 选择的折扣因子设置为 $\alpha = 0.5$。对于每个 β_k，如果 $\beta_k < 0.5$，则 L^2 采样器运行 2000 次迭代，否则运行 1000 次迭代。

Copula 生成的分布超参数设置与高斯混合模型的类似，但 $\sigma_{\mathrm{MH}} = \varepsilon = 0.1$、$\sigma_{\mathrm{PT},\ \beta} = 0.1/\sqrt{\beta}$，和 $\alpha = 0.7$。

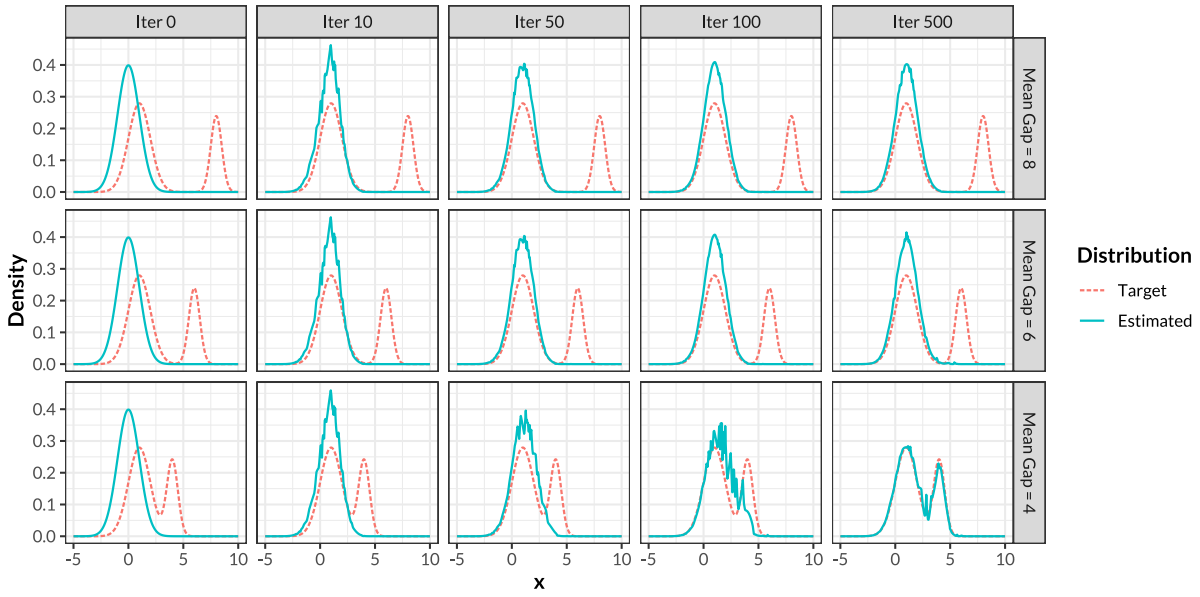
对于表中的计算时间基准测试S1，每个MCMC采样器运行10200次迭代，并将最初的200步作为预热期舍弃。

A.3 补充实验

A.3.1 KL采样器用于正态混合分布

在本实验中，我们扩展了第 2.3 节中的示例，并研究众数之间的间隔对采样质量的影响。图S1的第一行与原始示例相同，其中将 KL采样器应用于混合分布 $p_m(x) \sim 0.7 \cdot N(1,1) + 0.3 \cdot N(\mu, 0.25)$，其中 $\mu = 8$。图 S1将平均间隔 μ 分别减小到 6 和 4。可以观察到，除非两个众数是

当它们彼此足够接近时，KL采样器将只捕获一个模态。这一发现也验证了所提出框架中使用温度分布的合理性，因为温度化处理具有类似的效果，可以使众数之间的连接更加紧密。

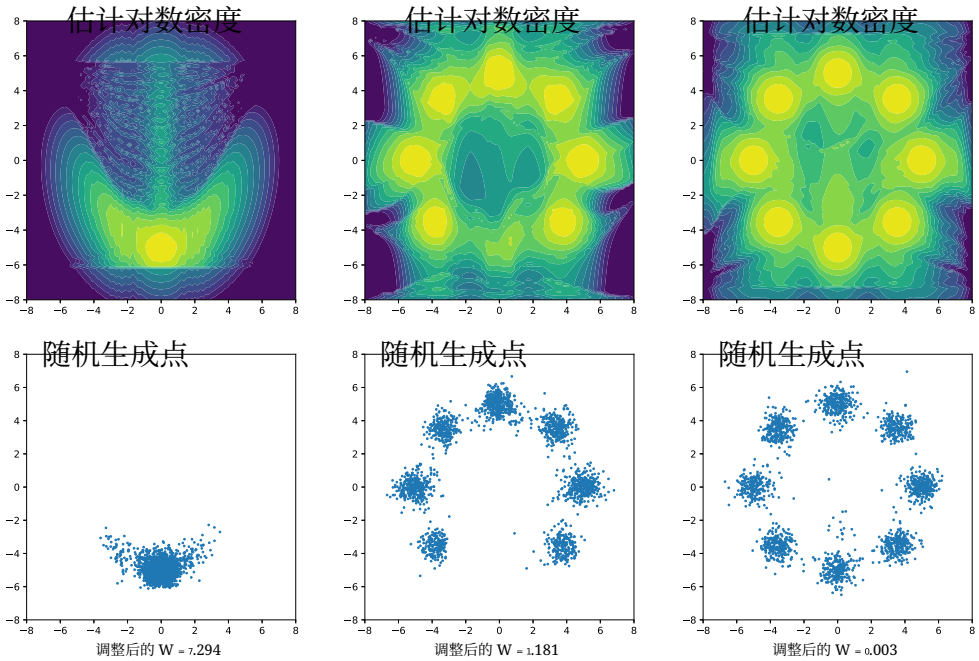


图S1：将 KL采样器应用于具有不同均值间隔的正态混合分布。每一列代表优化过程中的一次迭代。

在 TemperFlow 中用 KL采样器替换 L^2 采样器

在本实验中，我们测试结合温度化方法和KL采样器是否能够达到与TemperFlow相同的性能。这可以视为一项消融实验，将 L^2 采样器替换为TemperFlow中的KL采样器。我们重新考察第6.1节中的高斯混合模型，并分别使用TemperFlow及其修改版本学习采样器。我们特意让两种方法使用相同的 β序列进行因子控制，同时纳入未进行温度化的KL采样器结果作为参考。结果如图S2、S3和S4所示。

从图中我们可以发现，总体而言，温度调节大幅提升了KL采样器的质量。然而，即使采用温度调节，KL采样器仍倾向于错误地估计每个模态的概率质量，这与图2中的发现一致。这



(a) (b) (c)

图 S2：使用 (a) KL采样器、(b) 带回火的 KL采样器和 (c) TemperFlow 从“圆形”分布中采样。

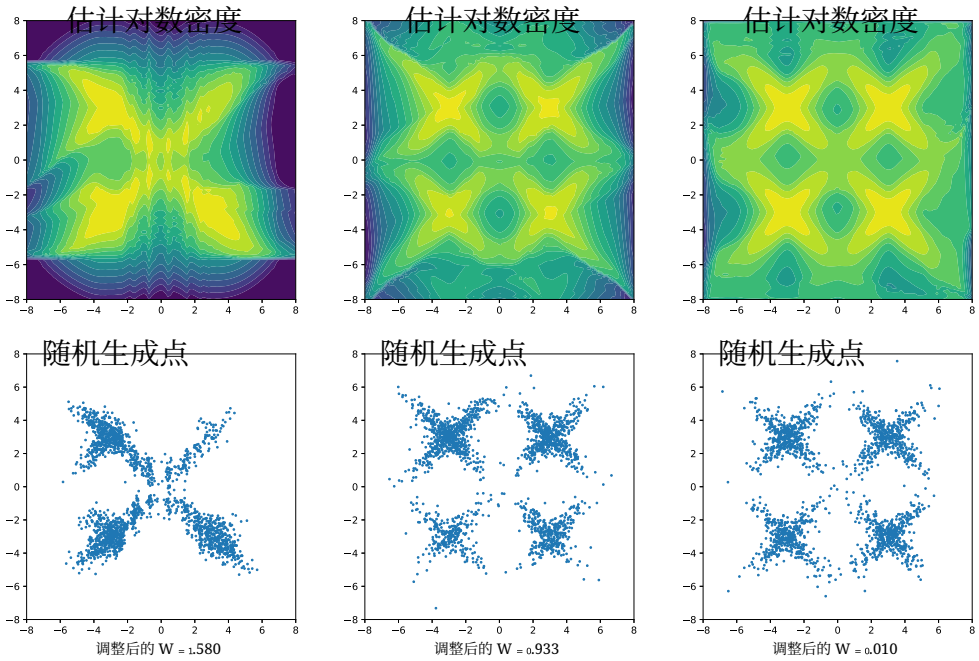
这表明 L^2 采样器在 TemperFlow 中发挥着重要作用，无法简单地由 KL采样器替代。

A.3.3 由 Copula 生成的分布的随机样本

图 S5分别展示了第 6.2 节实验中，通过并行回火和 TemperFlow 生成的样本的成对散点图和密度等高线图，实验采用$s = d = 8$。

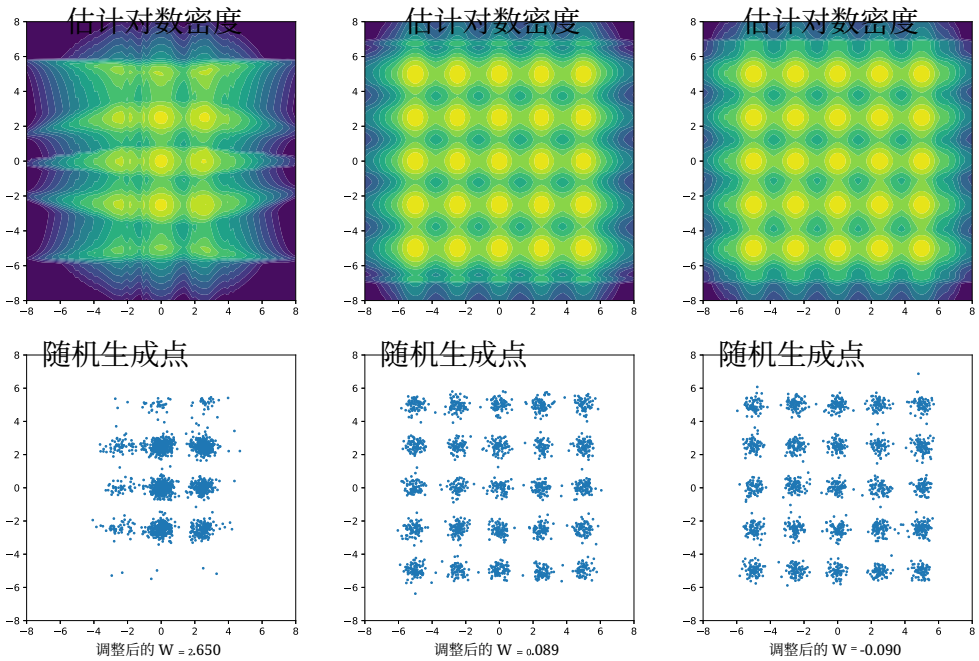
A.3.4 增加 MCMC 采样器的迭代次数 s

基于第6.2节中的实验，我们增加 MCMC采样器的迭代次数，并再次将其结果与 TemperFlow 进行比较。与之前的设置类似，我们舍弃最初的 200 步作为预热期，随后每 10 次迭代采集一个数据点。最终收集到 1000 个数据点，因此 MCMC采样器的总迭代次数为 10200。图S6 展示了这一新设置下的采样误差。



(a) (b) (c)

图 S3：使用 (a) KL采样器、(b) 采用退火的 KL采样器和 (c) TemperFlow 从“Cross”分布中采样。



(a) (b) (c)

图 S4：使用 (a) KL采样器、(b) 采用退火的 KL采样器和 (c) TemperFlow 从“Grid”分布中采样。

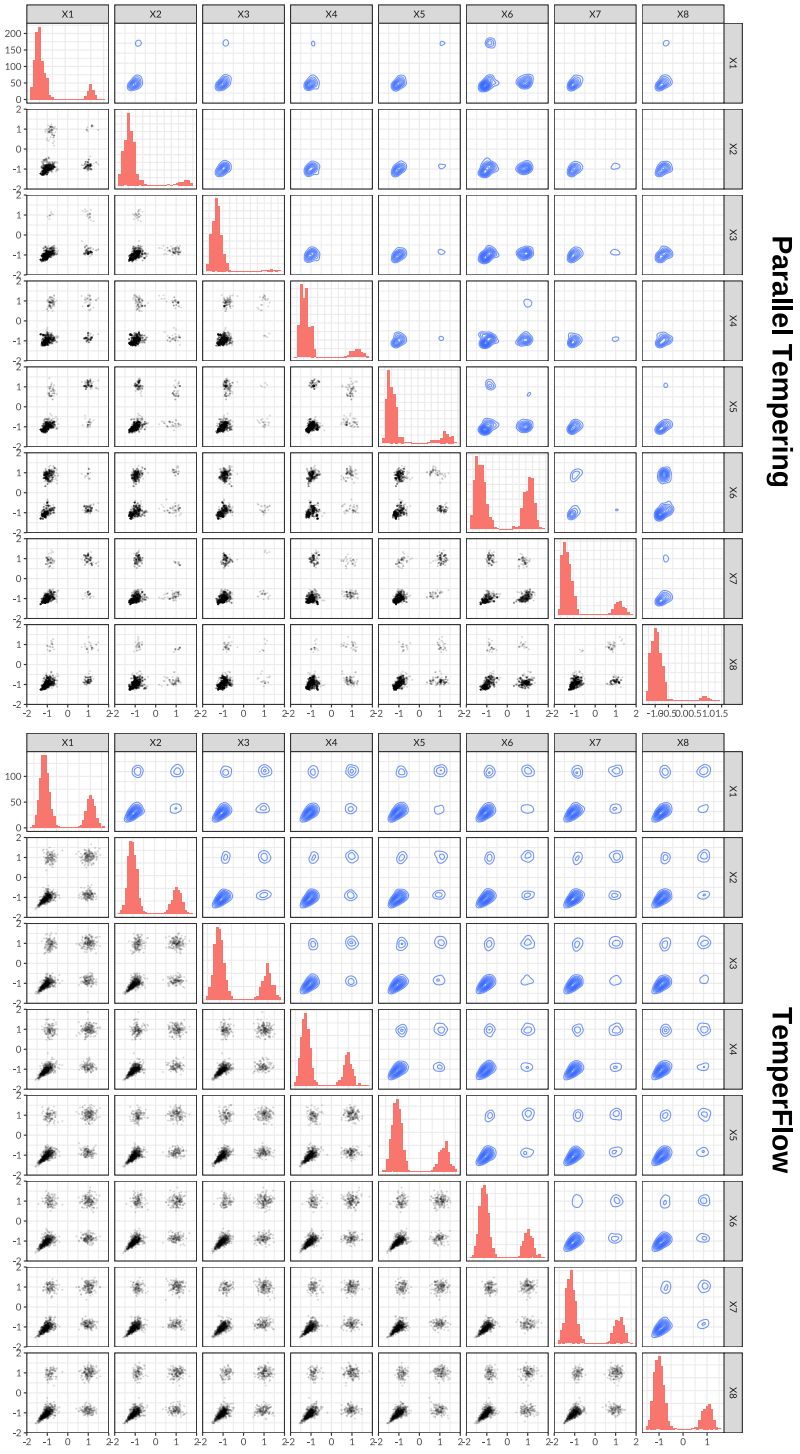


图 S5：并行回火和 TemperFlow 对由 Copula 生成的分布进行采样的结果，其中 $s = d = 8$。下三角部分显示生成样本的成对散点图，上三角部分为密度等高线图。对角线上的图为边缘分布的直方图。

与图 6相比，我们可以观察到整体模式非常相似，尽管现在马尔可夫链蒙特卡罗的迭代次数是之前的十倍。

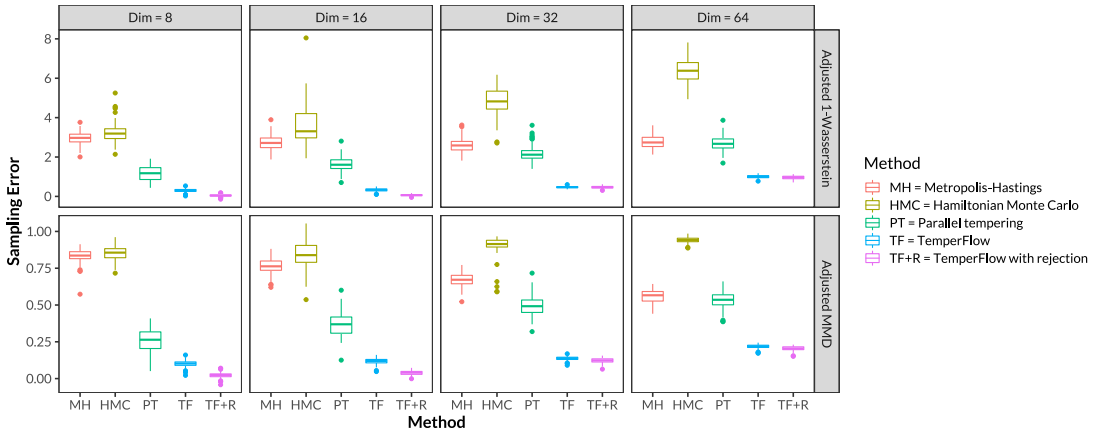


图 S6：不同方法在由 Copula 生成的分布上的采样误差，其中马尔可夫链蒙特卡罗的运行时间是图 S6中的十倍。

A.3.5 采样方法的计算时间

我们在表 S1 中报告了第 6.2 节中各个采样器在这些分布上的计算时间。所有采样器均在图形处理器上实现并运行。正如文章中所介绍的，TemperFlow包含一个训练阶段，其计算时间列于表 S1 的第一行。有趣的是，实证结果表明，训练时间大致随目标分布的维度线性增长。采样器训练完成后，即使执行后处理拒绝采样步骤，TemperFlow的生成时间也几乎可以忽略。相比之下，MCMC方法无需训练成本，但其生成时间与马尔可夫链迭代的次数成正比。还必须强调的是，表S1 中MCMC方法的时间仅反映固定迭代次数之后的计算成本，但图 6 和 S6 显然表明，它们的采样误差远高于TemperFlow。因此，在实践中，MCMC采样器需要大得多的迭代次数，这将大幅增加总计算时间。

表 S1：不同采样方法和维度下的计算时间。第一行显示 TemperFlow 的训练时间，括号中的数字是算法中使用的自适应 β'2。其余各行显示每种算法生成 10,000 个点所需的时间。所有时间值均以秒为单位。

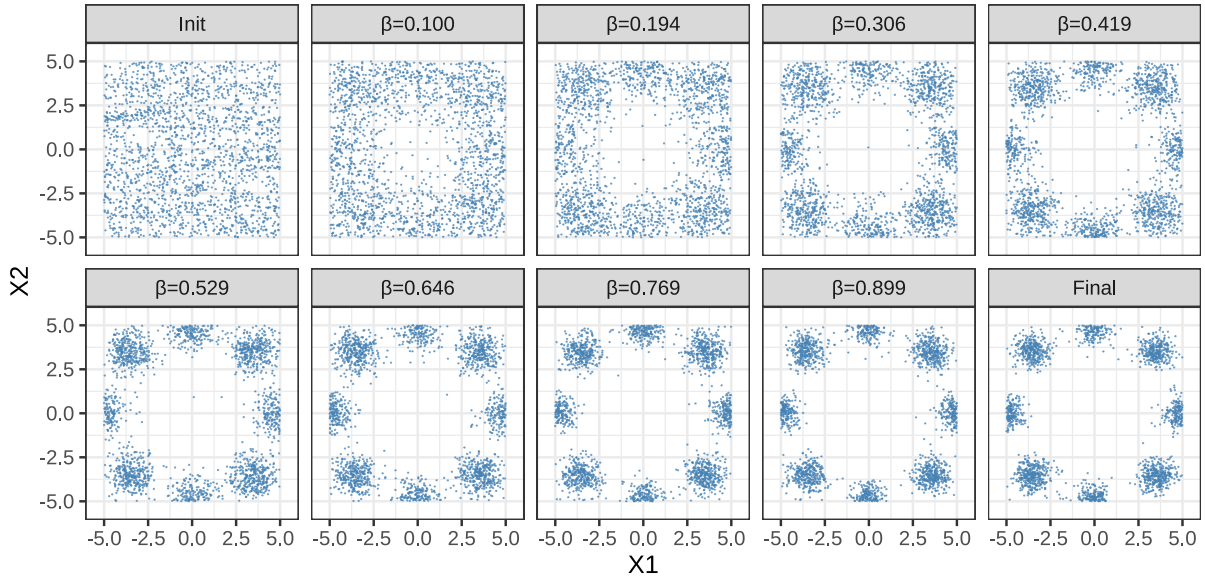
	$p = 8$	$p = 16$	$p = 32$	$p = 64$
（TemperFlow 训练）	133.7 (17)	309.9 (21)	598.6 (22)	1475.5 (30)
TemperFlow/10^4	0.00206	0.00449	0.00899	0.0169
TemperFlow+拒绝./10^4	0.0148	0.0263	0.0547	0.0928
MH/10^4	11.8	16.0	15.8	15.6
HMC/10^4	80.9	133.0	131.0	130.0
并行回火/10^4	24.9	32.2	32.5	31.9

A.3.6 更改 TemperFlow 中的基准测度

当目标分布 $p(x)$支撑于整个欧几里得空间 $\mathbb{R}^d$时，TemperFlow 默认使用标准多元正态分布作为基准测度 μ_0。如果 $p(x)$ 已知支撑于一个紧集上，例如超立方体 $[-B, B]^d$，那么 μ_0 也可以取为支撑于 $[-B, B]^d$上的分布，例如均匀分布。在图 S7 中，我们展示了这样一个例子。目标分布与第 6.1节中的"Circle"高斯混合模型相同，但我们将其密度限制在 $[-5, 5]^2$ 正方形内。然后，我们选择 μ_0 作为 $[-5, 5]^2$上的均匀分布，并使用 TemperFlow估计传输映射 T。图 S7 展示了$T_\sharp^{(k)}\mu_0$ 对应不同逆温度参数的退火分布流 β_k。可以发现，TemperFlow 再次很好地估计了目标分布。

A.3.7 Comparingdifferent INN architectures

从理论上讲，TemperFlow 可以利用任意变体的 INNs 来构建 T；在实践中，我们发现基于样条的模型（例如线性有理样条（Dolatabadi 等人，2020））表现非常出色，因此我们在 TemperFlow 中默认采用这类模型。为了测试不同 INN 架构的性能，我们首先考察第6.1节研究的二维分布，并选取三种具有代表性的 INN 进行比较：



图S7：将 TemperFlow 应用于支撑集为闭正方形的分布。基准测度 μ_0 取为均匀分布。

逆自回归流（Kingma 等人，2016）、线性有理样条（Dolatabadi 等人，2020），以及基于二次有理样条的神经样条流（Durkan 等人，2019）。图 S8 展示了使用不同 INN 架构生成的样本。显然，逆自回归流能够捕捉目标分布的轮廓，但在细节方面不如另外两种方法。因此，对于高维问题，我们只研究这两种基于样条的 INN。

图 S9 总结了第 6.2 节中的实验，但分别比较了基于线性有理样条和二次有理样条的 TemperFlow 采样器。表 S2 还展示了每个采样器的训练和采样时间。可以观察到，两个INNs的表现相似；在线性有理样条和二次有理样条的比较中，前者的采样误差略低，但训练时间稍长。

A.3.8 比较不同的折扣因子 α

在本节中，我们研究折扣因子 α 对 TemperFlow中自适应温度选择的影响。我们再次考虑第 节中的由 Copula 生成的分布。

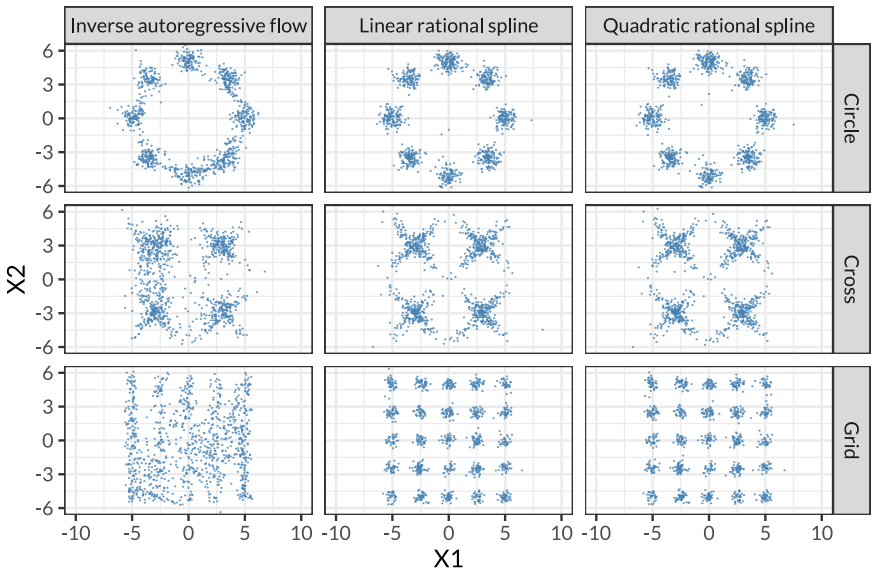


图 S8：基于不同 INN 架构的 TemperFlow 样本。

表 S2：基于两种不同的 INN 架构，TemperFlow 在由 Copula 生成的分布上的计算时间。对这些数字的解释与表 S1.

		$p = 8$	$p = 16$	$p = 32$	$p = 64$
LR	训练	135.6 (17)	320.6 (21)	642.0 (23)	1534.7 (30)
	采样/10^4	0.00215	0.00465	0.00892	0.0173
	采样+拒绝采样/10^4	0.0131	0.0189	0.0401	0.0783
QR	训练	122.9 (17)	284.1 (21)	567.7 (23)	1366.2 (30)
	采样/10^4	0.00156	0.00406	0.00806	0.0156
	采样+拒绝/10^4	0.0119	0.0199	0.0358	0.0779

6.2，并使用 $\alpha = 0.5$、0。6、0。7、0。8、0。9 来估计传输映射。图 S10 展示了各方法的最终采样误差，表 S3 给出了训练时间和所使用的 β 数量。使用 $\alpha = 0.9$ 的采样器在维度 32 和 64 下结果异常，因为我们将 β 的最大数量设为 100，且其计算尚未完成。

一般来说，通过设置更大的 α 数值，可以以增加迭代次数和延长训练时间为代价，获得更小的采样误差。对于 α没有明确的最优选择，但图 S10表明，TemperFlow 对 α 的选择并不十分敏感

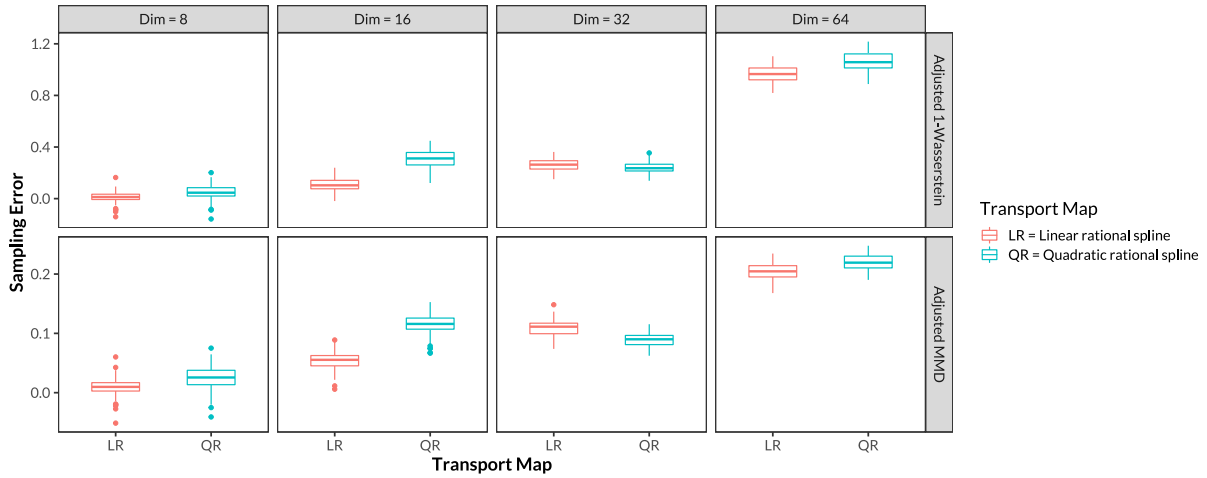


图 S9：基于不同 INN 架构的 TemperFlow 对由 Copula 生成的分布的采样误差。

$\alpha \geq 0.6$。此外，我们的实证结果表明，对于简单且低维的分布，α 可以设得较小，而对于更具挑战性的情况，则应设得较大。

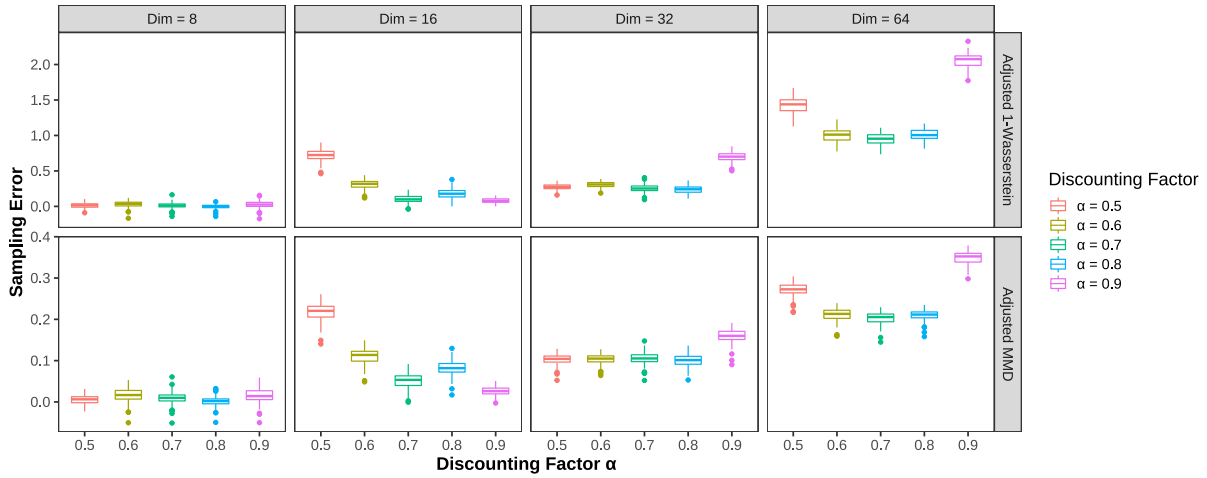


图 S10：基于不同 INN 架构的 TemperFlow 对由 Copula 生成的分布的采样误差。

A.3.9 重要性采样估计量的方差

在 TemperFlow 中，我们依靠重要性采样来估计温度化密度函数的归一化常数。一般而言，重要性采样的方差在很大程度上

表S3：基于不同 α 值的由Copula生成的分布上的 TemperFlow 训练时间。括号中的数字是所使用的自适应 β的数量，见算法 2；“最大值”表示 TemperFlow 已使用最大数量的 β，该数量在实验中设置为 100。

	$p=8$	$p=16$	$p=32$	$p=64$
$\alpha=0.5$	86.3 (10)	198.4 (12)	418.4 (14)	938.3 (17)
$\alpha=0.6$	106.1 (13)	236.5 (15)	486.9 (17)	1170.8 (22)
$\alpha=0.7$	135.6 (17)	320.6 (21)	642.0 (23)	1534.7 (30)
$\alpha=0.8$	206.4 (27)	459.7 (32)	984.8 (38)	2404.4 (50)
$\alpha=0.9$	488.6 (73)	1104.2 (87)	2356.0（最大值）	4730.1（最大值）

取决于提议分布。例如，为了估计归一化常数

$$Z_\beta = \int \exp\{-\beta E(x)\}\mathrm{d}x = \mathbb{E}_{X\sim h(x)} \exp\{-\beta E(X) - \log h(X)\},$$

ˆ 重要性采样估计量 Z_β 定义为

$$\hat{Z}_\beta = \frac{1}{M}\sum_{i=1}^{M} \frac{e^{-\beta E(X_i)}}{h(X_i)} := \frac{1}{M}\sum_{i=1}^{M} w(X_i), \quad X_i \sim h(x),$$

其中 $h(x)$ i ˆ s 是提议分布。容易验证， $\mathbb{E}(Z_\beta) = Z_\beta$，an d

$$\mathrm{Var}(\hat{Z}_\beta) = \frac{1}{M}\int \frac{[e^{-\beta E(x)} - Z_\beta h(x)]^2}{h(x)}\mathrm{d}x = \frac{1}{M}\int [w(x) - Z_\beta]^2 h(x)\mathrm{d}x.$$

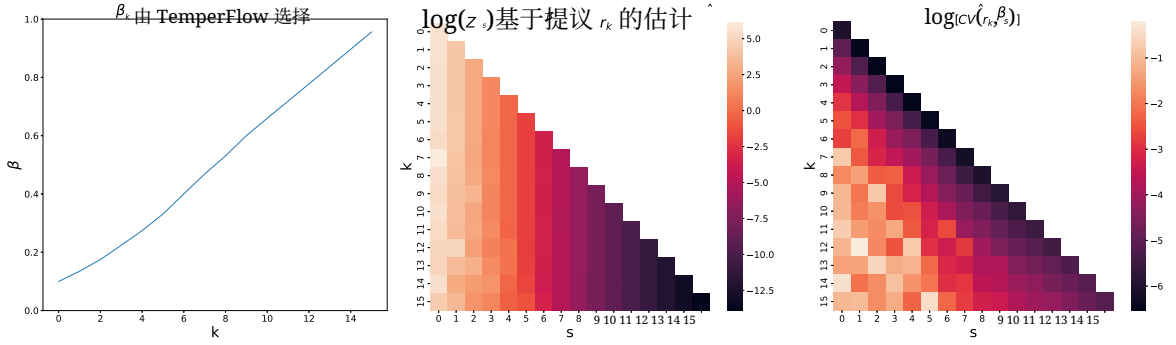
显然，如果 $h(x) \propto e^{-\beta E(x)}$，那么 Var(ˆ$Z_\beta$) $=0$。在 TemperFlow 中，我们使用 ˆ$r_s(x)$ $\approx Z_{\beta_s}^{-1} e^{-\beta_s E(x)}$ 作为提议分布来估计 $Z_{\beta_{s+1}}$，其中 ˆ$r_s = T_\sharp^{(s)}\mu_0$ 是估计得到的温度化 ˆ 分布。因此，只要 $\beta_s \approx \beta_{s+1}$， $Z_{\beta_{s+1}}$ 的方差就会很小。

为验证这一说法，我们考虑第 6.2 节中 $d=8$对应的实验，并从训练过程中提取 {ˆ r_k} 分布。然后，我们使用以下变异系数来量化重要性采样估计量的不确定性：

$$\mathrm{CV}(\hat{r}_k, \beta_s) = \frac{\sqrt{\widehat{\mathrm{Var}}(\hat{Z}_{\beta_s})}}{\hat{Z}_{\beta_s}} \quad \text{with } h(x) = \hat{r}_k(x), \quad s = 0, 1, \ldots, K,\ 0 \le k \le s,$$

$\widehat{\ }$ ˆ其中 Var(Z_β) $= M^{-1}(M-1)^{-1}\sum_{i=1}^{M}[w(X_i) - \hat{Z}_\beta]^2$是 Var(ˆ$Z_\beta$) 的一个估计量。直观而言，

如果 k 接近 s，我们预期 CV($\hat{r}_k, \beta_s$) 会较小。图 S11(a) 展示了 TemperFlow 选出的十六个 β 值 $\hat{}$ $\beta_0, \ldots, \beta_{15}$，图 S11(b) 则展示了以 $\hat{r}_k$ 作为提议分布时 log(Z_{β_s}) 的取值。如预期所示，矩阵的每一列中 log(Z_{β_s}) 的取值 $\hat{}$ $\hat{}$ 都相近，因为 Z_{β_s} 的期望不取决于提议分布 $\hat{}$。然而，不同的 $\hat{r}_k$会导致 Z_{β_s}的方差不同。图 S11(c) 展示了不同 (k, s) 组合下 log[CV($\hat{r}_k, \beta_s$)] 的取值，由此可观察到，对应于 $k = s$的对角元素具有最小的不确定性。这与重要性采样的理论分析一致。在 TemperFlow 中，我们只有 $\hat{r}_k$, $k < s$ 来估计 Z_{β_s}, 因此 $\hat{r}_{s-1}$ 在某种程度上是提议分布的最优选择。



(a) (b) (c)

$\hat{}$ 图 S11：（a）TemperFlow 选取的 β 值。（b）以 $\hat{r}_k$ 作为提议分布时，log(Z_{β_s}) 的值。（c）不同 (k, s) 下，log[CV($\hat{r}_k, \beta_s$)] 的值。

A.4 命题证明1

根据定义，$\mathcal{G}(T) = \mathcal{G}_1(T) - \mathcal{G}_2(T)$，其中 $\mathcal{G}_1(T) = \int q(x) \log q(x) \mathrm{d}x$，且 $\mathcal{G}_2(T) = \int q(x) \log p(x) \mathrm{d}x$。令 $\omega = T_\sharp \mu_0$ 且 $\nu = A_\sharp \omega = (A \circ T)_\sharp \mu_0$ 对于某个映射 $A\colon \mathbb{R}^d \to \mathbb{R}^d$，则 ν 的密度函数，记为 r，由 $r = (q/\det(\nabla A)) \circ A^{-1}$ 给出。对于任意函数 $f\colon \mathbb{R} \to \mathbb{R}$，我们有

$$\int f(r(x))\mathrm{d}x = \int f\left(\frac{q(A^{-1}(y))}{\det((\nabla A \circ A^{-1})(y))}\right)\mathrm{d}y = \int f\left(\frac{q(x)}{\det(\nabla A(x))}\right)\det(\nabla A(x))\mathrm{d}x.$$

取 $f(x) = x\log(x)$，然后我们得到

$$\begin{aligned}\mathcal{G}_1(A\circ T) &= \int r(x)\log r(x)\mathrm{d}x = \int q(x)\left[\log q(x) - \log\det(\nabla A(x))\right]\mathrm{d}x \\ &= \mathcal{G}_1(T) - \int \log\det(\nabla A(x))\mathrm{d}\omega(x).\end{aligned}$$

令 $I\colon \mathbb{R}^d \to \mathbb{R}^d$ 表示恒等映射，并令 I_d 表示 $d\times d$ 恒等矩阵。对于任意 $\Psi\in\mathbb{H}$，考虑$A = I+\varepsilon\Psi\circ T^{-1}$，则 $\nabla A(x) = I_d+\varepsilon G(x)$，其中 $G = \nabla(\Psi\circ T^{-1})$。

令 $l(\varepsilon; x) = \log\det(I_p + \varepsilon G(x))$，然后根据泰勒定理，

$$l(\varepsilon;x) = l(0;x) + l'(0;x)\varepsilon + \frac{1}{2}\int_0^\varepsilon l''(t;x)(\varepsilon - t)\mathrm{d}t.$$

可以验证，$l'(\varepsilon; x) = \mathrm{tr}\{U(\varepsilon;\ x)\}$，且 $l''(\varepsilon;\ x) = -\mathrm{tr}\{[U(\varepsilon;\ x)]^2\}$，其中 $U(\varepsilon;\ x) = [I_d + \varepsilon G(x)]^{-1}G(x)$, so

$$\int \log\det(\nabla A(x))\mathrm{d}\omega(x) = \varepsilon\int \mathrm{tr}\{G(x)\}\mathrm{d}\omega(x) - \frac{1}{2}\int\left[\int_0^\varepsilon \mathrm{tr}\{[U(t;x)]^2\}(\varepsilon - t)\mathrm{d}t\right]\mathrm{d}\omega(x).$$

令 $\sigma_1(M)\geq\cdots\geq\sigma_d(M)$ 为矩阵 M 的有序奇异值，根据冯·诺依曼迹不等式，我们有

$$\left|\mathrm{tr}\{[U(t;x)]^2\}\right| \leq \sum_{i=1}^d [\sigma_i(U(t;x))]^2 = \sum_{i=1}^d [\sigma_i(M_1^{-1}M_2)]^2,$$

其中 $M_1 = I_d+tG(x)$ 且 $M_2 = G(x)$。此外，Horn and Johnson(1991）的定理 3.3.14(b)表明

$$\sum_{i=1}^d [\sigma_i(M_1^{-1}M_2)]^2 \leq \sum_{i=1}^d [\sigma_i(M_1^{-1})\sigma_i(M_2)]^2 = \sum_{i=1}^d \left[\frac{\sigma_i(M_2)}{\sigma_{d-i+1}(M_1)}\right]^2.$$

根据假设 (a)，$\sigma_i(M_2)\leq\sigma_1(M_2) = \|G(x)\|_{\mathrm{op}}\leq C_{\Psi,\ T}$；而 Hornand Johnson （1991）中的定理 3.3.16(c)表明，$\sigma_i(M_1)\geq 1-|t|\sigma_1(G(x))\geq 1-|t|C_{\Psi,\ T}$。因此，对于

所有 $|t| < 1/(2C_{\Psi,\ T})$，都有 $\sigma_i(M_1) \geq 1/2$，因此

$$\left|\mathrm{tr}\{[U(t;x)]^2\}\right| \leq \sum_{i=1}^{d} \left[\frac{\sigma_i(M_2)}{\sigma_{d-i+1}(M_1)}\right]^2 \leq 4dC_{\Psi,T}^2.$$

因此，对于 $0 \leq \varepsilon \leq 1/(2C_{\Psi,T})$,

$$\left|\int \left[\int_0^{\varepsilon} \mathrm{tr}\{[U(t;x)]^2\}(\varepsilon - t)\mathrm{d}t\right] \mathrm{d}\omega(x)\right| \leq 4dC_{\Psi,T}^2 \int_0^{\varepsilon} (\varepsilon - t)\mathrm{d}t = 2dC_{\Psi,T}^2\varepsilon^2,$$

以及 $\varepsilon < 0$ 的情形也可类似地证明。这意味着

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\int \log\det(\nabla A(x))\mathrm{d}\omega(x)\Big|_{\varepsilon=0} = \int \mathrm{tr}\{G(x)\}\mathrm{d}\omega(x),$$

其中，由于 $|\mathrm{tr}(G(x))| \leq \sum_{i=1}^{d} \sigma_i(G(x)) \leq dC_{\Psi,T}$，因此存在 $\int \mathrm{tr}\{G(x)\}\mathrm{d}\omega(x)$。因此,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{G}_1(T + \varepsilon\Psi)\Big|_{\varepsilon=0} = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{G}_1(A \circ T)\Big|_{\varepsilon=0} = -\int \mathrm{tr}\left\{G(x)\right\}\mathrm{d}\omega(x) = -\int \nabla\cdot(\Psi \circ T^{-1})\mathrm{d}\omega,$$

其中 $\nabla \cdot F$ 是向量场 F的散度。对于标量值函数 φ，散度算子满足乘积法则 $\nabla\cdot(\varphi F) = (\nabla\varphi)\cdot F + \varphi(\nabla\cdot F)$，因此

$$\begin{aligned}
\int \nabla\cdot(\Psi \circ T^{-1})\mathrm{d}\omega &= \int q(x)\left(\nabla\cdot(\Psi \circ T^{-1})\right)(x)\mathrm{d}x \\
&= \int \nabla\cdot(q(\Psi \circ T^{-1}))\mathrm{d}\lambda - \int \langle\nabla q, \Psi \circ T^{-1}\rangle\mathrm{d}\lambda \\
&= \int \nabla\cdot(q(\Psi \circ T^{-1}))\mathrm{d}\lambda - \int \langle\nabla \log q, \Psi \circ T^{-1}\rangle\mathrm{d}\omega \\
&= \int \nabla\cdot(q(\Psi \circ T^{-1}))\mathrm{d}\lambda - \int \langle\nabla \log q \circ T, \Psi\rangle\mathrm{d}\mu_0.
\end{aligned}$$

在 $\lim_{\|x\|\to\infty} \|q(x)\Psi(T^{-1}(x))\| = 0$ 这一假设下，我们有 $\int \nabla\cdot(q(\Psi\circ T^{-1}))\mathrm{d}\lambda = 0$，因此 $\delta\mathcal{G}_1/\delta T = \nabla \log\ q \circ T$。

对于第二部分，根据定义，我们有 $\mathcal{G}_2(T) = \int \log\ p(T(x))\mathrm{d}\mu_0(x)$

推前算子。根据泰勒定理,

$$\log p(T+\varepsilon\Psi)-\log p(T)=\varepsilon\langle\nabla\log p\circ T,\Psi\rangle+\frac{1}{2}\varepsilon^2\Psi'\left[\nabla^2(\log p)\right](T+c\varepsilon\Psi)\Psi$$

f或某个 $c\in(0,1)$。由于我们假设 $\|\nabla^2(\log p)\|_{\mathrm{op}}\leq c$，因此得到 t

$$\int\left|\Psi'\left[\nabla^2(\log p)\right](T+c\varepsilon\Psi)\Psi\right|\mathrm{d}\mu_0\leq c\int\|\Psi(x)\|^2\mathrm{d}\mu_0(x)<\infty,$$

然后

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{G}_2(T+\varepsilon\Psi)\Big|_{\varepsilon=0}&=\lim_{\varepsilon\to0}\frac{\mathcal{G}_2(T+\varepsilon\Psi)-\mathcal{G}_2(T)}{\varepsilon}\\&=\lim_{\varepsilon\to0}\varepsilon^{-1}\int\left[\varepsilon\langle\nabla\log p\circ T,\Psi\rangle+\frac{1}{2}\varepsilon^2\Psi'\left[\nabla^2(\log p)\right](T+c\varepsilon\Psi)\Psi\right]\mathrm{d}\mu_0\\&=\int\langle\nabla\log p\circ T,\Psi\rangle\mathrm{d}\mu_0=\langle\nabla\log p\circ T,\Psi\rangle_{\mathbb{H}}.\end{aligned}$$

因此， $\delta\mathcal{G}_2/\delta T=\nabla\log p\circ T$。

综合上述结果，容易得出 $\delta\mathcal{G}/\delta T=\delta\mathcal{G}_1/\delta T-\delta\mathcal{G}_2/\delta T=\nabla(\log q-\log p)\circ T$。

A.5 定理1的证明

定理的第一部分是若干经典及近期结果的推论。我们首先证明，若概率测度 μ 具有对数凹密度 $p(x)$，则其满足 Poincaré不等式

$$\int f^2\mathrm{d}\mu-\left(\int f\mathrm{d}\mu\right)^2\leq C_P\int\|\nabla f\|^2\mathrm{d}\mu\tag{12}$$

对所有局部 Lipschitz $f\in L^2(\mu)$，其中 $C_P>0$ 与 f无关。事实上，Cheeger(1970) 表明（12）在 $C_P=4/\psi_\mu^2$，其中

$$\psi_\mu=\inf_{S\subset\mathbb{R}^d}\frac{\int_{\partial S}p(x)\mathrm{d}x}{\min\{\mu(S),\mu(S^c)\}}$$

被称为 μ的等周系数。此外，Chen （2021年）的定理1证明，对于任意整数 d， $l \geq 1$，，

$$\psi_\mu \geq \frac{1}{[c \cdot l(\log(d)+1)]^{l/2} \cdot d^{16/l} \cdot \sqrt{\sigma^2}}$$

对于某个通用常数 $c > 0$，其中 σ^2 是 μ的协方差矩阵的谱范数。为方便起见，令 $l_1 \equiv l_1($ $d) = \log(d)+1$， $l_2 = \log(l_1)+1$， $L_1 = \sqrt{l_1/l_2}$，以及 $L_2 = \sqrt{l_1 l_2}$。取 $l = \lceil L_1 \rceil$，其中 $\lceil x \rceil$ 是不小于 x的最小整数。容易验证， L_1 关于 d 和 $L_1 \geq 1$递增，因此我们必有 $L_1 \leq l = \lceil L_1 \rceil \leq 2L_1$。另外注意， $L_1 \leq \sqrt{l_1}$ 和

$$\log(L_1) \leq \frac{1}{2}\log(l_1) < \frac{1}{2}l_2, \qquad L_1 \log(l_1) < L_1 l_2 = L_2, \qquad L_1 \log(L_1) < L_1 \cdot \frac{1}{2}l_2 = \frac{1}{2}L_2,$$

so

$$\begin{aligned}
\log\left\{[c \cdot l(\log(d)+1)]^{l/2} \cdot d^{16/l}\right\} &= \frac{l}{2} \cdot \log(c \cdot l \cdot l_1) + \frac{16}{l}\log(d) \\
&\leq L_1 \left[\log(2c) + \log(L_1) + \log(l_1)\right] + \frac{16}{L_1} \cdot l_1 \\
&\leq \log(2c) L_1 + \frac{3}{2}L_2 + 16L_2.
\end{aligned}$$

显然， $L_1 \leq L_2$，因此存在一个常数 $c' > 0$，使得 $\log\left\{[c \cdot l(\log(d)+1)]^{l/2} \cdot d^{16/l}\right\} \leq c' L_2$ 和

$$\psi_\mu \geq \frac{1}{e^{c' L_2} \cdot \sqrt{\sigma^2}}.$$

综上所述，我们可以取 $C_P = 4\sigma^2 e^{2c' L_2}$。

接下来，Chewi 等人（2020）的 $p(x)$ 满足 Poincaré 不等式（12），那么 $\{\nu_t\}$ 朗之万扩散的分布

$$\mathrm{d}X_t = \nabla_x \log p(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

其中 B_t 是 $\mathbb{R}^d$上的布朗运动，满足 $\mathcal{F}(\nu_t) \leq e^{-2t/C_P}\chi^2(\nu_0\|\mu)$。已知朗之万扩散的分布是KL散度的梯度流（Jordan等人，1998），因此 μ_t 和 ν_t 具有相同的边缘分布。因此，我们还有$\mathcal{F}(\mu_t) \leq e^{-2t/C_P}\chi^2(\mu_0\|\mu)$。将 C_P的表达式代入，并将常数吸收到c'中，即得所述结果。

对于第二部分，（6）以及方程 $\mathbf{v}_t(x) = -\nabla\left[\log p_t(x) - \log p(x)\right]$ 表明

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) = -\int \left\|\nabla \log \frac{\mathrm{d}\mu_t}{\mathrm{d}\mu}\right\|^2 \mathrm{d}\mu_t.$$

令 $f = \sqrt{\mathrm{d}\mu_t/\mathrm{d}\mu}$，则

$$\left\|\nabla \log \frac{\mathrm{d}\mu_t}{\mathrm{d}\mu}\right\|^2 = \left\|\nabla \log(f^2)\right\|^2 = 4\left\|\nabla \log f\right\|^2 = \frac{4}{f^2}\left\|\nabla f\right\|^2.$$

因此，

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) = -\frac{4\mathrm{d}\mu}{\mathrm{d}\mu_t}\int \|\nabla f\|^2 \mathrm{d}\mu_t = -4\int \|\nabla f\|^2 \mathrm{d}\mu.$$

根据 Poincaré不等式（12），我们有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) &\leq -\frac{4}{C_P}\left[\int f^2 \mathrm{d}\mu - \left(\int f \mathrm{d}\mu\right)^2\right] = -\frac{4}{C_P}\left[\int \frac{\mathrm{d}\mu_t}{\mathrm{d}\mu}\mathrm{d}\mu - \left(\int \sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\mu}}\mathrm{d}\mu\right)^2\right]. \\ &= -\frac{4}{C_P}\left[1 - \left(\int \sqrt{\frac{\mathrm{d}\mu_t}{\mathrm{d}\lambda}\cdot\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}}\mathrm{d}\lambda\right)^2\right].\end{aligned}$$

请注意

$$\begin{aligned}2H^2(\mu,\nu) &= \int (\sqrt{\mathrm{d}\mu/\mathrm{d}\lambda} - \sqrt{\mathrm{d}\nu/\mathrm{d}\lambda})^2 \mathrm{d}\lambda = \int \mathrm{d}\mu + \int \mathrm{d}\nu - 2\int \sqrt{\mathrm{d}\mu/\mathrm{d}\lambda}\cdot\sqrt{\mathrm{d}\nu/\mathrm{d}\lambda}\mathrm{d}\lambda \\ &= 2\left[1 - \int \sqrt{\mathrm{d}\mu/\mathrm{d}\lambda}\cdot\sqrt{\mathrm{d}\nu/\mathrm{d}\lambda}\mathrm{d}\lambda\right],\end{aligned}$$

so

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) \le -\frac{4}{C_P}\left[1-\left(1-H^2(\mu_t,\mu)\right)^2\right] = -\frac{4}{C_P}\left[2-H^2(\mu_t,\mu)\right]\cdot H^2(\mu_t,\mu) \le -\frac{4}{C_P}H^2(\mu_t,\mu).$$

最后的不等式成立，这是因为 $0 \le H(\mu,\nu) \le 1$。最后，代入 C_P的表达式即可得到结果。

A.6 定理证明2

为简化记号，我们令 $q(x) \equiv p_{t^*}(x)$，且 $\mathcal{F}$ 表示基于 ϕ-散度的泛函 $\mathcal{F}_\phi$。根据定义，

$$\begin{aligned}
\mathcal{F}(\mu_t) = \mathcal{D}_\phi(\mu_t\|\mu) &= \int p(x)\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x \\
&= \alpha\int h(x)\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x + (1-\alpha)\int h(x-\mu)\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x \\
&= \alpha\int h(x)\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x + (1-\alpha)\int h(x)\phi\left(\frac{q(x+\mu)}{p(x+\mu)}\right)\mathrm{d}x, \\
\frac{q(x)}{p(x)} &= \frac{\gamma h(x)+(1-\gamma)h(x-\mu)}{\alpha h(x)+(1-\alpha)h(x-\mu)} = 1+\frac{(\gamma-\alpha)(h(x)-h(x-\mu))}{\alpha h(x)+(1-\alpha)h(x-\mu)}.
\end{aligned}$$

使用元素 tary 不等式 $(a+b)/(c+d) \le a/c+b/d$ 对任意 $a,b,c,d \ge 0$，我们有

$$\frac{p(x)}{q(x)} \le \frac{\alpha}{\gamma}+\frac{1-\alpha}{1-\gamma} \coloneqq c_1^{-1}(\alpha,\gamma), \quad \frac{q(x)}{p(x)} \le \frac{\gamma}{\alpha}+\frac{1-\gamma}{1-\alpha} \coloneqq c_2(\alpha,\gamma).$$

因此，对于所有 $x \in \mathbb{R}^d$，$0 < c_1 \le q(x)/p(x) \le c_2 < \infty$，我们将 $D_1 = \text{最大值}_{x\in[c_1,c_2]}|\phi(x)|$ 和 $D_2 = \text{最大值}_{x\in[c_1,c_2]}|\phi''(x)|$ 记作。因此，

$$\left|\phi\left(\frac{q(x)}{p(x)}\right)\right| \le D_1, \quad \left|\phi\left(\frac{q(x+\mu)}{p(x+\mu)}\right)\right| \le D_1, \quad \forall x,\mu \in \mathbb{R}^d.$$

此外，根据假设 (a)，对于任意固定的 x，有 $\lim_{\|\mu\|\to\infty}\phi(q(x)/p(x)) = \phi(\gamma/\alpha)$，且 $\lim_{\|\mu\|\to\infty}\phi(q(x+\mu)/p(x+\mu)) = \phi((1-\gamma)/(1-\alpha))$。由于 $\int Dh(x)\mathrm{d}x = D < \infty$，根据

控制收敛定理，可得

$$\lim_{\|\mu\|\to\infty}\int h(x)\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x=\int h(x)\phi\left(\frac{\gamma}{\alpha}\right)\mathrm{d}x=\phi\left(\frac{\gamma}{\alpha}\right),$$
$$\lim_{\|\mu\|\to\infty}\int h(x)\phi\left(\frac{q(x+\mu)}{p(x+\mu)}\right)\mathrm{d}x=\int h(x)\phi\left(\frac{1-\gamma}{1-\alpha}\right)\mathrm{d}x=\phi\left(\frac{1-\gamma}{1-\alpha}\right),$$

由此得到第一个结果。

对于第二部分，根据定义，

$$\begin{aligned}\nabla\left(\frac{q(x)}{p(x)}\right)&=(\alpha-\gamma)\cdot\nabla\left(\frac{h(x-\mu)-h(x)}{p(x)}\right)\\&=\frac{\alpha-\gamma}{[p(x)]^2}\cdot\{p(x)[\nabla h(x-\mu)-\nabla h(x)]-[h(x-\mu)-h(x)]\nabla p(x)\}\\&=\frac{\alpha-\gamma}{[p(x)]^2}\cdot[h(x)\nabla h(x-\mu)-h(x-\mu)\nabla h(x)]\\&=(\alpha-\gamma)\left[\frac{h(x)}{p(x)}\right]^2\nabla\left(\frac{h(x-\mu)}{h(x)}\right).\end{aligned}$$

定义 $\delta(x,\ \mu)=V(x-\mu)-V(x)$，于是

$$\frac{h(x)}{p(x)}=\frac{h(x)}{\alpha h(x)+(1-\alpha)h(x-\mu)}=\frac{1}{\alpha+(1-\alpha)\exp\{-\delta(x,\mu)\}}.$$

取 $c=\min\{\alpha,1-\alpha\}$ 并定义 $\sigma(x)=1/(1+e^{-x})$，于是 $h(x)/p(x)\le\sigma(\delta(x,\ \mu))/c$。此外，$\nabla(h(x-\mu)/h(x))=\nabla(e^{-\delta(x,\mu)})=-e^{-\delta(x,\mu)}\nabla_x\delta(x,\ \mu)$。根据假设 (b)，$\|\nabla_x\delta(x,\ \mu)\|=\|\nabla V(x-\mu)-\nabla V(x)\|\le C_1\|\mu\|^k$ 当 $\|\mu\|$ 足够大时成立，因此

$$\left\|\nabla\left(\frac{q(x)}{p(x)}\right)\right\|\le\frac{|\alpha-\gamma|}{c^2}\cdot[\sigma(\delta(x,\mu))]^2e^{-\delta(x,\mu)}\cdot C_1\|\mu\|^k\le\frac{C_1|\alpha-\gamma|}{c^2}\cdot\|\mu\|^ke^{-|\delta(x,\mu)|},\quad(13)$$

其中，第二个不等式是由于 $[\sigma(x)]^2e^{-x}\le e^{-|x|}$ 这一事实。

接下来，Gao 等人 (2019) 的引理 2.1 表明 $(\delta\mathcal{F}/\delta\nu)(\rho)=\phi'(\mathrm{d}\rho/\mathrm{d}\mu)$，因此 $\mathbf{v}_t(x)=-\phi''(p_t(x)/p(x))\nabla[p_t(x)/p(x)]$。代入 (13)，可得

$$\|\mathbf{v}_{t^*}(x)\|\le\frac{C_1D_2|\alpha-\gamma|}{c^2}\cdot\|\mu\|^ke^{-|\delta(x,\mu)|}.$$

因此,

$$0 \leq -\left.\frac{\mathrm{d}\mathcal{F}(\mu_t)}{\mathrm{d}t}\right|_{t=t^*} = \int q(x)\|\mathbf{v}_{t^*}(x)\|^2 \mathrm{d}x \leq \frac{C_1 D_2|\alpha-\gamma|}{c^2} \cdot \|\mu\|^k \int q(x) e^{-|\delta(x,\mu)|}\mathrm{d}x.$$

根据定义,

$$\begin{aligned}
\int q(x)e^{-|\delta(x,\mu)|}\mathrm{d}x &= \gamma \int h(x)e^{-|\delta(x,\mu)|}\mathrm{d}x + (1-\gamma)\int h(x-\mu)e^{-|\delta(x,\mu)|}\mathrm{d}x \\
&= \gamma \int h(x)e^{-|\delta(x,\mu)|}\mathrm{d}x + (1-\gamma)\int h(x)e^{-|\delta(x+\mu,\mu)|}\mathrm{d}x \\
&= \gamma \mathbb{E}e^{-|Y_{\mu,1}|} + (1-\gamma)\mathbb{E}e^{-|Y_{\mu,2}|}.
\end{aligned}$$

令 $g_1(y)$ 为 $|Y_{\mu,1}|$ 的密度函数，然后

$$\begin{aligned}
\mathbb{E}e^{-|Y_{\mu,1}|} &= \int_0^{C_2\|\mu\|} g_1(y)e^{-y}\mathrm{d}x + \int_{C_2\|\mu\|}^{+\infty} g_1(y)e^{-y}\mathrm{d}x \\
&\leq \int_0^{C_2\|\mu\|} g_1(y)\mathrm{d}x + \int_{C_2\|\mu\|}^{+\infty} g_1(y)e^{-C_2\|\mu\|}\mathrm{d}x \\
&= P(|Y_{\mu,1}| \leq C_2\|\mu\|) + e^{-C_2\|\mu\|}P(|Y_{\mu,1}| > C_2\|\mu\|) \\
&\leq P(|Y_{\mu,1}| \leq C_2\|\mu\|) + e^{-C_2\|\mu\|}.
\end{aligned}$$

同样，我们可以验证 $\mathbb{E}e^{-|Y_{\mu,2}|} \leq P(|Y_{\mu,2}| \leq C_2\|\mu\|) + e^{-C_2\|\mu\|}$。最后，根据假设（c），我们得到

$$\begin{aligned}
C_1\|\mu\|^k \int q(x)e^{-|\delta(x,\mu)|}\mathrm{d}x \leq C_1\|\mu\|^k e^{-C_2\|\mu\|} + \gamma C_1\|\mu\|^k P(|Y_{\mu,1}| \leq C_2\|\mu\|) + \\
(1-\gamma)C_1\|\mu\|^k P(|Y_{\mu,2}| \leq C_2\|\mu\|) \to 0
\end{aligned}$$

作为 $\|\mu\| \to \infty$，这得出了所要求的结论。

A.7 定理证明 3 和（10）

令 $q_\beta(x) = e^{-\beta E(x)}/Z(\beta)$，其中 $Z(\beta) = \int e^{-\beta E(x)}\mathrm{d}x$。定义函数 $\ell(\beta) = \mathrm{KL}(q_\beta \| p)$，我们将证明 $\ell(\beta)$ 随着 β的增加而递减。根据定义，

$$\ell(\beta) = \int q_\beta(x) \log \frac{q_\beta(x)}{p(x)} \mathrm{d}x = \int q_\beta(x)[\log q_\beta(x) - \log p(x)]\mathrm{d}x.$$

为简洁起见，令 $\zeta(\beta) = \log Z(\beta)$，则 $\log\ q_\beta(x) = -\beta E(x) - \zeta(\beta)$ 且 $\log\ p(x) = -E(x) - \zeta(1)$。因此，

$$\begin{aligned}\ell(\beta) &= \int q_\beta(x)[E(x) + \zeta(1) - \beta E(x) - \zeta(\beta)]\mathrm{d}x \\ &= (1-\beta)\int q_\beta(x)E(x)\mathrm{d}x + \zeta(1) - \zeta(\beta),\end{aligned}$$

and

$$\ell'(\beta) = -\int q_\beta(x)E(x)\mathrm{d}x + (1-\beta)\int \frac{\partial q_\beta(x)}{\partial \beta}E(x)\mathrm{d}x - \zeta'(\beta).$$

请注意，

$$\zeta'(\beta) = \frac{Z'(\beta)}{Z(\beta)} = \frac{1}{Z(\beta)}\int e^{-\beta E(x)}[-E(x)]\mathrm{d}x = -\int q_\beta(x)E(x)\mathrm{d}x,$$

因此 $\ell'(\beta) = (1-\beta)\int[\partial q_\beta(x)/\partial\beta]E(x)\mathrm{d}x$。此外，由于 $\partial \log\ q_\beta(x)/\partial\beta = -E(x) - \zeta'(\beta)$，我们有

$$\frac{\partial q_\beta(x)}{\partial \beta} = q_\beta(x)\cdot\frac{\partial \log q_\beta(x)}{\partial\beta} = -q_\beta(x)[E(x) + \zeta'(\beta)].$$

因此，

$$\begin{aligned}\ell'(\beta) &= -(1-\beta)\int q_\beta(x)[E(x) + \zeta'(\beta)]E(x)\mathrm{d}x \\ &= -(1-\beta)\left\{\int q_\beta(x)E^2(x)\mathrm{d}x - \left[\int q_\beta(x)E(x)\mathrm{d}x\right]^2\right\}. \qquad (14)\end{aligned}$$

根据詹森不等式，$\int q_\beta(x)E^2(x)\mathrm{d}x \geq \left[\int q_\beta(x)E(x)\mathrm{d}x\right]^2$，且等号当且仅当 $E(x)$ 是常数时成立。由于能量函数不可能在 $\mathbb{R}^d$ 上为常数，因此不等式是严格的。因此，当 $0<\beta<1$ 时，$\ell'(\beta)<0$，且仅当 $\beta=1$ 时，$\ell'(\beta)=0$。由于 β_t 在 t 中递增，因此可知 $\mathcal{F}(\rho_t)=\ell(\beta_t)$ 关于 t 递减。

最后，（10）可以通过注意到 $\mathrm{d}[\log \ell(e^\gamma)]/\mathrm{d}\gamma = \beta\cdot\ell'(\beta)/\ell(\beta)$，其中 $\ell'(\beta)$ 在（14）中给出，下面我们给出 $\ell(\beta)$ 的表达式。根据定义，

$$\ell(\beta)=\int q_\beta(x)[\log q_\beta(x)-\log p(x)]\mathrm{d}x=\int q_\beta(x)[\log q_\beta(x)+E(x)+\zeta(1)]\mathrm{d}x,$$

其中

$$\zeta(1)=\log Z(1)=\log\int \exp\{-E(x)\}\mathrm{d}x=\log\int q_\beta(x)\exp\{-E(x)-\log q_\beta(x)\}\mathrm{d}x,$$

于是该方程成立。

A.8 命题的证明2

由于 $f(x_\varepsilon)=\varepsilon^{1/4}$ 且 $f(x)$ 对于 $x\geq 0$ 是递减的，我们有 $\inf_{Q^\varepsilon} f\leq \varepsilon^{1/4}$。此外，$\int_{Q^\varepsilon} f\mathrm{d}\lambda = 2\Phi(x_\varepsilon+M\varepsilon^{1/12})-1$，其中 $\Phi(\cdot)$ 是标准正态分布的累积分布函数。对于充分小的 ε，$x_\varepsilon=\sqrt{-(\log\varepsilon)/2-\log(2\pi)}>\sqrt{\log(1/\varepsilon)/3}>2$。然后根据众所周知的不等式 $1-\Phi(x)<f(x)/x$ for $x>0$，我们有

$$\int_{Q^\varepsilon} f\mathrm{d}\lambda > 2\Phi(x_\varepsilon)-1>1-2f(x_\varepsilon)/x_\varepsilon = 1-2\varepsilon^{1/4}/x_\varepsilon>1-\varepsilon^{1/4}$$

如果 ε 足够小。因此假设 1（a）成立。

为方便起见，令 $\delta=M\varepsilon^{1/12}$。由于 $f^{-2}(x)=2\pi e^{x^2}$ 关于零对称，且在 $x>0$ 上递增，因此假设 1(b) 中的上确界在 $Q=Q^\varepsilon$ 时取得。已知 $h(x)=e^{-x^2}\int_0^x e^{y^2}\mathrm{d}y<0.6$，所以

$$\int_{Q^\varepsilon} f^{-2}(x)\mathrm{d}x = 4\pi\int_0^{x_\varepsilon+\delta} e^{x^2}\mathrm{d}x = 4\pi e^{(x_\varepsilon+\delta)^2}h(x_\varepsilon+\delta)<2.4\pi e^{(x_\varepsilon+\delta)^2}.$$

设 $l(x) = e^{x^2/2}$，然后 $l'(x) = xl(x)$ 以及 $l''(x) = l(x)(x^2+1)$，因此

$$e^{(x_\varepsilon+\delta)^2/2} = e^{x_\varepsilon^2/2}\left[1 + x_\varepsilon\delta + (1+x_\varepsilon^2)\delta^2 + O((x_\varepsilon\delta)^3)\right] < Ce^{x_\varepsilon^2/2} = C\varepsilon^{-1/4}$$

对于某个常数 $C > 0$。因此，

$$\sup_{\substack{Q\subset Q^\varepsilon \\ Q\in\mathcal{Q}}} |Q|^{2/d}\left[\frac{1}{|Q|}\int_Q f^{-2}\mathrm{d}\lambda\right]^{1/2} = |Q^\varepsilon|^{3/2}\left[\int_{Q^\varepsilon} f^{-2}(x)\mathrm{d}x\right]^{1/2} < \sqrt{2.4\pi}C\varepsilon^{-1/4}|Q^\varepsilon|^{3/2}.$$

$|Q^\varepsilon|$ 的阶为 $\sqrt{\log(1/\varepsilon)}$，因此假设 1(b) 在 $c = 3/4$ 下成立。

最后，注意到

$$\int_{Q^\varepsilon}\left[\varepsilon^{1/4} - f(x)\right]_+^2 \mathrm{d}x = 2\int_{x_\varepsilon}^{x_\varepsilon+\delta}\left[\varepsilon^{1/4} - f(x)\right]^2\mathrm{d}x.$$

令 $g(x) = \left[\varepsilon^{1/4} - f(x)\right]^2$，则有 $g'(x) = -2\left[\varepsilon^{1/4} - f(x)\right]f'(x)$，且 $g''(x) = 2[f'(x)]^2 - 2\left[\varepsilon^{1/4} - f(x)\right]f''(x)$。由于 $g(x_\varepsilon) = g'(x_\varepsilon) = 0$并且

$$g''(x_\varepsilon) = 2[f'(x_\varepsilon)]^2 = 2x_\varepsilon^2[f(x_\varepsilon)]^2 = 2x_\varepsilon^2\varepsilon^{1/2},$$

根据泰勒展开，我们有

$$\begin{aligned}\frac{1}{2}\int_{Q^\varepsilon}\left[\varepsilon^{1/4} - f(x)\right]_+^2\mathrm{d}x &= g(x_\varepsilon)\delta + \frac{1}{2}g'(x_\varepsilon)\delta^2 + \frac{1}{6}g''(x_\varepsilon)\delta^3 + O(\delta^4) \\ &= \frac{1}{3}x_\varepsilon^2\varepsilon^{1/2}\delta^3 > \frac{M^3}{9}\varepsilon^{3/4}\log(1/\varepsilon) > \frac{M^3}{9}\varepsilon^{3/4}[\log(1/\varepsilon)]^{3/4}\end{aligned}$$

当 ε充分小时，这验证了假设1(c)。

A.9 技术引理

令 λ 表示 $\mathbb{R}^d$ 上的勒贝格测度。若函数 w 在 $\mathbb{R}^d$ 上非负且局部可积，则称其为权。权 w 诱导出一个测度，定义为 $w(A) = \int_A w\mathrm{d}\lambda$。用 $\mathcal{Q}$ 表示 $\mathbb{R}^d$ 中球的集合。若对每个 $Q \in \mathcal{Q}$，权 w 都满足 $w(2Q) \le Cw(Q)$，则称其为倍增权，其中 $2Q$ 表示与 Q 同心且半径为其两倍的球，而 C

称为 w 的倍增常数。对于可测集 $A \subset \mathbb{R}^d$，令 $|A| = \lambda(A)$，并用 $\|f\|_{L^p_w(A)}$ 表示向量值函数 $f\colon \mathbb{R}^d \to \mathbb{R}^{d'}$ 在 A 上的 w 加权 L^p 范数，即，

$$\|f\|_{L^p_w(A)} = \left[\int_A \|f(x)\|^p w(x)\mathrm{d}x\right]^{1/p},$$

其中 $\|\cdot\|$ 是欧几里得范数。

引理 1 (定理 2.14，Chua，1993)。 设 $1 < p \le q < \infty$ 并设 $\sigma = v^{-1/(p-1)}$ 其中 v是一个权。假设 w 是一个倍增权。那么对于所有 $Q_0 \in \mathcal{Q}$ 及定义在 Q_0上的 *Lipschitz* 连续函数 f，

$$\|f - f_{Q_0,w}\|_{L^q_w(Q_0)} \le A(v,w,Q_0)\|\nabla f\|_{L^p_v(Q_0)}$$

其中

$$A(v,w,Q_0) = C(p,q,C_0) \sup_{\substack{Q\subset Q_0\\ Q\in\mathcal{Q}}} |Q|^{1/d-1} w(Q)^{1/q}\sigma(Q)^{1/p'}$$

当 $p < q$ 时，且

$$A(v,w,Q_0) = C(p,r,C_0) \sup_{\substack{Q\subset Q_0\\ Q\in\mathcal{Q}}} |Q|^{1/d}\left[\frac{1}{|Q|}\int_Q w^r\mathrm{d}\lambda\right]^{1/(pr)}\left[\frac{1}{|Q|}\int_Q \sigma^r\mathrm{d}\lambda\right]^{1/(p'r)}$$

当 $p = q$ 时，对于任意 $r > 1$。C_0 是 w 的倍增常数，$p' = p/(p-1)$，且$f_{Q_0,w} = w(Q_0)^{-1}\int_{Q_0} f w\mathrm{d}\lambda$。

引理 2。 设 $g(x)$ 是定义在 $\mathbb{R}^d$ 上的密度函数。那么对于任意球 $Q_0 \in \mathcal{Q}$ 以及定义在 Q_0 上的 *Lipschitz* 连续函数 h，以下不等式成立，

$$\int_{Q_0} |h(x) - h_{Q_0}|^2\,\mathrm{d}x \le A(g,d,Q_0)\int_{Q_0}\|\nabla h(x)\|^2 g(x)\mathrm{d}x,$$

其中

$$h_{Q_0} = \frac{1}{|Q_0|}\int_{Q_0} h\mathrm{d}\lambda, \quad A(g,d,Q_0) = C(d)\sup_{\substack{Q\subset Q_0\\ Q\in\mathcal{Q}}} |Q|^{2/d}\left[\frac{1}{|Q|}\int_Q g^{-2}\mathrm{d}\lambda\right]^{1/2},$$

且 $C(d)$ 是一个与 g、 h 和 Q_0 无关的常数。

证明。 容易验证， $w \equiv 1$ 是一个倍增权，其倍增常数为 $C_0 = 2^d$。然后令 $v = g$, $w = 1$, $p = q = r = 2$ 代入引理 1，结论即成立。 □

引理 3.设 $f(x)$是定义在紧集 $S \subset \mathbb{R}^d$上的连续函数。假设$\inf_S f(x) < c$对于某个常数 $c > 0$，并记 $M = \int_S [c - f(x)]_+^2 dx$，其中 g_+ =最大值$(0, g)$ 表示函数 g 的正部。则对于满足$\int_S [e(x)]^2 dx \le M$ 的任意连续函数$e(x)$，有 $\inf_S [f(x) + e(x)] \le c$。

证明。 由于任何 $e(x)$ 都可以替换为 $|e(x)|$ ，从而在相同条件 $\int_S [e(x)]^2 \mathrm{d}x \le M$ 下获得更大的 $\inf_S [f(x) + e(x)]$，因此不失一般性，我们可以假设 $e(x) \ge 0$ 。

令 $h_c(x)$ =最大值$(c, f(x))$，$e^*(x) = h_c(x) - f(x) = [c - f(x)]_+$，且 $u^*(x) = [e^*(x)]^2$。于是容易看出， $h_c(x)$ 在 $T := \{x$ 上： $f(x) \le c\}$ ，并且 $e^*(x) = u^*(x) = 0$ 在 $S \backslash T$ 上，这意味着 $M = \int_S u^*(x) \mathrm{d}x = \int_T u^*(x) \mathrm{d}x$。对于任意连续的 $u(x) \ge 0$ ，使得 $\int_S u(x) \mathrm{d}x \le M$，我们断言必然存在某个点 $x_0 \in T$，使得 $u(x_0) \le u^*(x_0)$。如果这不成立，那么

$$\int_S u(x)\mathrm{d}x > \int_T u^*(x)\mathrm{d}x + \int_{S\backslash T} u(x)\mathrm{d}x \ge \int_T u^*(x)\mathrm{d}x + \int_{S\backslash T} u^*(x)\mathrm{d}x = \int_S u^*(x)\mathrm{d}x = M,$$

这就导致了矛盾。

令 $e(x) = \sqrt{u(x)}$，那么对于某个 $x_0 \in T$， $e(x_0) \le e^*(x_0)$ 也成立，这意味着

$$\inf_S [f(x) + e(x)] \le f(x_0) + e(x_0) \le f(x_0) + e^*(x_0) = h_c(x_0) = c.$$

□

A.10 定理 4

容易证明， $\mathcal{L}$ 在 g 处的一阶变分为 $(\delta\mathcal{L}/\delta g)(g) = 2(g-f)$，然后连续性方程 $\partial g_t/\partial t + \nabla\cdot(\mathbf{v}_t g_t) = 0$ 的向量场 $\mathbf{v}_t$ 由$\mathbf{v}_t = 2\nabla(f-g_t)$给出。令 $h_t = g_t - f$，然后

$$\frac{\mathrm{d}\mathcal{L}(g_t)}{\mathrm{d}t} = -\int \|2\nabla h_t\|^2 g_t \mathrm{d}\lambda = -4\int \|\nabla h_t\|^2 g_t \mathrm{d}\lambda.$$

在下文中，只要不会引起混淆，我们就省略 g_t 中的下标 t，以及 Q_i^ε 中的上标 ε。由于 $|\mathrm{d}\mathcal{F}(g_t)/\mathrm{d}t| \le \varepsilon$，我们有 $\int \|\nabla h\|^2 g \mathrm{d}\lambda \le \varepsilon/4$，因此

$$\max_{1\le i\le K}\int_{Q_i} \|\nabla h\|^2 g \mathrm{d}\lambda \le \varepsilon/4.$$

根据引理 2，我们有

$$\int_{Q_i} |h(x) - h_{Q_i}|^2 \,\mathrm{d}x \le A(g,d,Q_i)\varepsilon/4, \tag{15}$$

其中

$$A(g,d,Q_i) = C(d)\sup_{\substack{Q\subset Q_i\\ Q\in\mathcal{Q}}} |Q|^{2/d}\left[\frac{1}{|Q|}\int_Q g^{-2}\mathrm{d}\lambda\right]^{1/2}.$$

根据假设1(b) 和 2，我们有

$$A(g,d,Q_i) \le \alpha^{-1}C(d)\sup_{\substack{Q\subset Q_i\\ Q\in\mathcal{Q}}} |Q|^{2/d}\left[\frac{1}{|Q|}\int_Q f^{-2}\mathrm{d}\lambda\right]^{1/2} \le \alpha^{-1}C(d)M_1\varepsilon^{-1/4}[\log(1/\varepsilon)]^c.$$

由于 $M_2 > 0$ 在假设1(c) 中是一个足够大的常数，我们可以取 $M_2 = (4\alpha)^{-1}C(d)M_1$，于是（15）可得

$$\int_{Q_i} |h(x) - h_{Q_i}|^2 \,\mathrm{d}x \le M_2\varepsilon^{3/4}[\log(1/\varepsilon)]^c. \tag{16}$$

设 $e_i(x) = h(x) - h_{Q_i}$，根据定义，$\int_{Q_i} e_i \mathrm{d}\lambda = 0$ 和 $\int_{Q_i} e_i^2 \mathrm{d}\lambda \le M_2\varepsilon^{3/4}[\log(1/\varepsilon)]^c$.

在引理3中，令 $S = Q_i$，$e = e_i$，$M = M_2\varepsilon^{3/4}[\log(1/\varepsilon)]^c$，以及 $c = \varepsilon^{1/4}$，根据假设1(a)，我们得到$\inf_{Q_i}(f+e_i) \leq \varepsilon^{1/4}$。由于 $f+e_i$ 是连续的，且 Q_i 是紧致的，因此必存在一点 $x^* \in Q_i$ 使得

$$f(x^*) + e_i(x^*) = \inf_{Q_i}(f+e_i) \leq \varepsilon^{1/4}.$$

于是我们得到

$$h_{Q_i} = h(x^*) - e_i(x^*) = g(x^*) - f(x^*) - e_i(x^*) \geq -[f(x^*) + e_i(x^*)] \geq -\varepsilon^{1/4}. \tag{17}$$

另一方面，$\int_Q g\mathrm{d}\lambda \leq 1$ 且 $\int_Q f\mathrm{d}\lambda \geq 1-\varepsilon^{1/4}$ 根据假设1(a)，因此

$$\varepsilon^{1/4} \geq \int_Q (g-f)\mathrm{d}\lambda = \sum_{i=1}^K \int_{Q_i}(g-f)\mathrm{d}\lambda = \sum_{i=1}^K \int_{Q_i}(e_i + h_{Q_i})\mathrm{d}\lambda.$$

注意，$\int_{Q_i} e_i\mathrm{d}\lambda = 0$ 和 $\int_{Q_i} h_{Q_i}\mathrm{d}\lambda = h_{Q_i}|Q_i|$，因此我们有

$$h_{Q_i}|Q_i| \leq \sum_{i=1}^K h_{Q_i}|Q_i| \leq \varepsilon^{1/4}. \tag{18}$$

结合（17）和（18），然后得到 $|h_{Q_i}| \leq \max(1, |Q_i|^{-1}) \cdot \varepsilon^{1/4}$.

最后，（16）表明$\|h - h_{Q_i}\|_{L^2(Q_i)} \leq M_2^{1/2}\varepsilon^{3/8}[\log(1/\varepsilon)]^{c/2}$，因此

$$\begin{aligned}\|h\|_{L^2(Q_i)} &\leq \|h - h_{Q_i}\|_{L^2(Q_i)} + \|h_{Q_i}\|_{L^2(Q_i)} \leq M_2^{1/2}\varepsilon^{3/8}[\log(1/\varepsilon)]^{c/2} + h_{Q_i}|Q_i|^{1/2} \\ &\leq M_2^{1/2}\varepsilon^{3/8}[\log(1/\varepsilon)]^{c/2} + \max(|Q_i|^{1/2}, |Q_i|^{-1/2}) \cdot \varepsilon^{1/4}.\end{aligned}$$

参考文献

Ambrosio, L., Gigli, N.,andSavare, G.（2008年）。梯度流：度量空间与概率测度空间。Birkhäuser。

Bhattacharya, R.（1978）。多维扩散过程的常返性与不变测度存在性判据。《概率年刊》，第541–553页。

Bond-Taylor, S., Leach, A., Long, Y. 和 Willcocks, C. G.（2021）。深度生成建模：对变分自编码器、生成对抗网络、归一化流、基于能量的模型和自回归模型的比较综述。*arXiv* 预印本，*arXiv:2103.04922*。

Brooks, S., Gelman, A., Jones, G. 和 Meng, X.-L.（2011）。马尔可夫链蒙特卡罗手册。Chapman & Hall/CRC。Che, T., Zhang, R., Sohl-Dickstein, J., Larochelle, H., Paull, L., Cao, Y. 和 Bengio, Y.（2020）。你的生成对抗网络实际上是一个基于能量的模型，因此你应该使用判别器驱动的潜变量采样。载于神经信息处理系统进展 *33*。Cheeger, J.（1970）。拉普拉斯算子最小特征值的一个下界。载于分析中的问题，第195–199页。普林斯顿大学出版社。Chen, R. T., Rubanova, Y., Bettencourt, J. 和 Duvenaud, D. K.（2018）。神经常微分方程。载于神经信息处理系统进展第*31*卷。Chen, Y.（2021）。KLS猜想中等周系数的一个几乎为常数的下界。几何与泛函分析，31(1):34–61。Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L. 和 Jordan, M. I.（2018）。非凸情形下朗之万动力学的精确收敛速率。*arXiv* 预印本 *arXiv:1805.01648*。Chewi, S., Gouic, T. L., Lu, C., Maunu, T., Rigollet, P. 和 Stromme, A. J.（2020）。镜像朗之万扩散的指数遍历性。*arXiv* 预印本 *arXiv:2005.09669*。Chua, S.-K.（1993）。满足链条件的区域上的加权索o6列夫不等式。美国数学会会刊，117(2):449–457。Chwialkowski, K., Strathmann, H. 和 Gretton, A.（2016）。一种拟合优度核检验。载于国际机器学习会议，第2606–2615页。PMLR。Dinh, L., Krueger, D. 和 Bengio, Y.（2014）。NICE：非线性独立成分估计。*arXiv* 预印本 *arXiv:1410.8516*。Dinh, L., Sohl-Dickstein, J. 和 Bengio, S.（2016）。使用Real NVP进行密度估计。*arXiv* 预印本 *arXiv:1605.08803*。Dolatabadi, H. M., Erfani, S. 和 Leckie, C.（2020）。使用线性有理样条进行可逆生成建模。载于国际人工智能与统计会议，第4236–4246页。Dongarra, J. 和 Sullivan, F.（2000）。客座编辑引言：十大算法。计算科学与工程，2(01):22–23。Dunson, D. B. 和 Johndrow, J.（2020）。Hastings算法问世五十周年。*Biometrika*，107(1):1–23。Durkan, C., Bekasov, A., Murray, I. 和 Papamakarios, G.（2019）。神经样条流。载于神经信息处理系统进展第*32*卷。Earl, D. J. 和 Deem, M. W.（2005）。并行回火：理论、应用与新视角。物理化学化学物理，7(23):3910–3916。

Falcioni, M. and Deem, M. W. (1999). 一种用于沸石结构解析的有偏蒙特卡罗方案。《化学物理学杂志》, 110(3):1754–1766. Gao, Y., Huang, J., Jiao, Y., Liu, J., Lu, X., and Yang, Z. (2022). 通过欧拉粒子输运进行深度生成学习。载于数学与科学机器学习, 页码 336–368. Gao, Y., Jiao, Y., Wang, Y., Wang, Y., Yang, C., and Zhang, S. (2019). 通过变分梯度流进行深度生成学习。载于国际机器学习会议, 页码 2093–2101. PMLR. Ge, R., Lee, H., and Risteski, A. (2018). 模拟退火朗之万蒙特卡洛 II：使用软马尔可夫链分解改进的证明。*arXiv* 预印本 *arXiv:1812.00793*. Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). 贝叶斯数据分析. Chapman & Hall/CRC. Geyer, C. J. (1991). 马尔可夫链蒙特卡洛最大似然。载于第*23*届界面专题研讨会. Geyer, C. J. and Thompson, E. A. (1995). 退火马尔可夫链蒙特卡洛及其在祖先推断中的应用。美国统计学会会刊, 90(431):909–920. Gilks, W., Richardson, S., and Spiegelhalter, D. (1995). 《实践中的马尔可夫链蒙特卡洛》. Chapman & Hall/CRC. Goodfellow, I., Bengio, Y., and Courville, A. (2016). 深度学习. 麻省理工学院出版社. Grathwohl, W., Chen, R. T., Bettencourt, J., and Duvenaud, D. (2019). 具有自由形式连续动力学的可扩展可逆生成模型。载于国际学习表征会议. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). 一种用于双样本问题的核方法。载于神经信息处理系统进展 *19*. Hastings, W. K. (1970). 使用马尔可夫链的蒙特卡洛采样方法及其应用。*Biometrika*, 57(1):97–109. Hoffman, M., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. (2019). 使用神经输运改善哈密顿蒙特卡洛中的不良几何结构。*arXiv* 预印本，*arXiv:1903.03704*. Horn, R. A. and Johnson, C. R. (1991). 矩阵分析专题. 剑桥大学出版社. Huang, J., Jiao, Y., Kang, L., Liao, X., Liu, J., and Liu, Y. (2021). 薛定谔-费尔默采样器：无需遍历性的采样。*arXiv* 预印本 *arXiv:2106.10880*.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). 福克–普朗克方程的变分表述。*SIAM* 数学分析杂志, 29(1):1–17. Kingma, D. P. and Ba, J. (2015). Adam：一种随机优化方法。载于国际学习表征会议, 页码 1–13. Kingma, D. P., Salimans, T., Jozefo wicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). 使用逆自回归流改进变分推断。载于神经信息处理系统进展 *29*. Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). 通过模拟退火进行优化。《科学》, 220(4598):671–680. Kofke, D. A. (2002). 关于副本交换蒙特卡罗试验的接受概率。《化学物理学杂志》, 117(15):6911–6914. Kone, A. and Kofke, D. A. (2005). 并行回火模拟的温度间隔选择。《化学物理学杂志》, 122(20):206101. Levine, R. A. and Casella, G. (2001). 蒙特卡罗期望最大化算法的实现。《计算与图形统计学杂志》, 10(3):422–439. Ley, C. and Swan, Y. (2013). Stein 密度方法与信息不等式。《概率电子通讯》, 18:1–14. Liu, Q., Xu, J., Jiang, R., and Wong, W. H. (2021). 使用深度生成神经网络进行密度估计。《美国国家科学院院刊》, 118(15). Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). 野外人脸属性的深度学习。载于*IEEE* 国际计算机视觉会议论文集, 页码 3730–3738. Marinari, E. and Parisi, G. (1992). 模拟退火：一种新的蒙特卡罗方案。《欧洲物理快报》, 19(6):451. Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. (2016). 通过测度传输进行采样：简介。载于 Ghanem, R., Higdon, D., and Owhadi, H., 编者, 《不确定性量化手册》. Springer. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). 使用高速计算机进行状态方程计算。《化学物理学杂志》, 21(6):1087–1092. Neal, R. M. (1996). 使用回火跃迁从多峰分布中采样。《统计与计算》, 6(4):353–366. Nelsen, R. B. (2006). *Copula* 导论 . 施普林格科学与商业媒体 . Owen, A. B. (2013). 蒙特卡罗理论、方法与实例。Pang, B., Han, T., Nijkamp, E., Zhu, S.-C., and Wu, Y. N. (2020). 学习潜在空间中的基于能量的先验模型。载于神经信息处理系统进展 *33*.

Papamakarios, G., Pavlakou,T.和Murray, I.（2017年）。用于密度估计的掩码自回归流。载于神经信息处理系统进展第*30*卷。Qiu,Y.和Wang,X.（2021年）。ALMOND：通过神经网络和朗之万扩散实现的自适应潜变量建模与优化。美国统计学会会刊，116(535)：1224–1236。Raginsky, M.、Rakhlin,A.和 Telgarsky,M.（2017年）。通过随机梯度朗之万动力学进行非凸学习：非渐近分析。载于学习理论会议，第1674–1703页。PMLR。Rezende,D.和Mohamed, S.（2015）。使用归一化流进行变分推断。载于国际机器学习会议，第1530–1538页。PMLR。Robbins, H. 和 Monro, S.（1951）。一种随机逼近方法。数学统计年刊，22(3)：400–407。Romano, Y.、Sesia, M.和 Candès, E.（2020）。深度knockoff。美国统计学会会刊，115(532)：1861–1872。Salakhutdinov, R.（2015）。学习深度生成模型。统计学及其应用年度评论，2：361–385。Santambrogio, F.（2017年）。 {欧几里得、度量和 Wasserstein}梯度流：概述。数学科学通报，7(1)：87–154。Sullivan,T.J.（2015）。不确定性量化导论。S pringer。Sun, Y.、Song, Q. 和Liang, F.（2021年）。一致稀疏深度学习：理论与计算。已被美国统计学会会刊接收。Swendsen,R.H. 和Wang, J.-S.（1986）。自旋玻璃的副本蒙特卡罗模拟。物理评论快报，57(21)：2607。Tabak, E.G. 和Turner, C. V.（2013）。一类非参数密度估计算法。纯粹与应用数学通讯，66(2)：145–164。Tabak, E. G.和 Vanden-Eijnden, E.（2010）。通过对数似然的对偶上升进行密度估计。数学科学通讯，8(1)：217–233。Vempala, S. 和Wibisono,A.（2019）。未校正朗之万算法的快速收敛：等周不等式已足够。载于神经信息处理系统进展第*32*卷。Villani,C.（2009年）。最优传输：旧与新。Springer。Vousden,W.、Farr,W. M. 和Mandel, I.（2016）。马尔可夫链蒙特卡罗模拟中的并行回火动态温度选择。皇家天文学会月报，455(2)：1919–1937。

Wei, G. C. 和 Tanner, M. A.（1990年）。EM算法的蒙特卡洛实现与简易数据增广算法。美国统计学会会刊，85(411):699–704。Woodard, D. B.、Schmidler, S. C. 和 Huber, M.（2009年）。多峰分布上并行退火和模拟退火快速混合的条件。应用概率年刊，19(2):617–640。Xiao, H.、Rasul, K. 和 Vollgraf, R.（2017年）。Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集。*arXiv*预印本 *arXiv:1708.07747*。Yuan, Y.、Deng, Y.、Zhang, Y. 和 Qu, A.（2020）。统计视角下的深度学习。*Stat*，9(1):e294。Zheng, Z.（2003年）。论交换算法和模拟退火算法。随机过程及其应用，104(1):131–154。Zhou, X.、Jiao, Y.、Liu, J. 和 Huang, J.（2021年）。一种用于条件采样的深度生成方法。已被美国统计学会会刊接收。