arXiv:2601.19722v1 [stat.CO] 27 Jan 2026

零阶并行采样

Francesco Pozza*和 Giacomo Zanella†

摘要

如何有效利用并行计算加速马尔可夫链蒙特卡罗方法，是贝叶斯计算及相关学科中的一个重要问题。本文研究零阶设定：由于理论、实践或计算方面的原因，未能获得梯度，但可以计算未归一化的目标分布。我们还假设可以使用 m 个并行处理器来加速收敛。本文提出的方法受到现代零阶优化方法的启发：这类方法通过用零阶随机梯度估计量替代梯度，来模拟基于梯度的方案。我们的贡献有两方面。首先，我们证明，在马尔可夫链蒙特卡罗方法中直接应用常用的零阶随机梯度估计量，会导致算法对 m的依赖性较差，无论是未调整方案还是 Metropolis 调整方案。随后，我们提出一种简单的解决办法：采用随机切片视角，而非随机梯度视角，由此得到一类新的零阶采样器，并从理论上证明其可实现m多项式加速。数值研究支持了理论结果。

关键词：马尔可夫链蒙特卡罗；并行计算；对数凹分布；混合时间；谱隙；零阶优化

1 引言

人们对开发能够在其流程中利用并行计算以提高效率的马尔可夫链蒙特卡罗（马尔可夫链蒙特卡洛）方法抱有浓厚兴趣（Rosenthal，2000年；Liu 等人，2000年；Tjelmeland，2004年；Brockwell，2006 年；Calderhead，2014年；Gagnon 等人，2023年）。由于马尔可夫链蒙特卡罗算法本质上是顺序执行的，这项任务通常会带来不小的复杂性，因此，简单地并行运行不同的马尔可夫链（Rosenthal，2000年）可以说是实践中最常用的解决方案。除了实现相对简单之外，这种方法的优点是运行时不同处理器之间基本无需通信。此外，对于马尔可夫链的预热期（即每条链达到收敛所需的时间）相对于总运行时间较短的问题，它有助于提高模拟的有效样本量。相反，对于复杂的高维目标密度，马尔可夫链蒙特卡罗算法可能需要经过许多次迭代才能收敛至平稳分布。在这些情况下，与经典的单处理器方法相比，采用并行链马尔可夫链蒙特卡罗方法的优势往往较为有限，因此吸引力较低，尤其是在每条独立链的长度受限时。基于这一原因，过去二十年来，另一条研究路线专注于开发马尔可夫链蒙特卡罗

*博科尼数据科学与分析研究所，博科尼大学，意大利米兰。francesco.pozza2@unibocconi.it

†决策科学系及博科尼数据科学与分析研究所，博科尼大学，意大利米兰。giacomo.zanella@unibocconi.it

Both authors acknowledge support from the European Union (ERC), through the Starting grant 'PrSc-HDBayLe', project number 101076564.

方法，利用并行计算加速收敛。这些算法在单条马尔可夫链内部整合并行计算，其结构通常比并行链马尔可夫链蒙特卡罗方法更复杂。下文将这类技术称为步内并行化方法。

本文研究如何利用并行计算加速零阶框架下的收敛。具体而言，在这种设定中，只能评估目标密度的势函数，而无法计算其导数。尽管零阶算法通常不如一阶和二阶方法（即利用梯度和海森矩阵信息的方法）高效，但这一框架正重新受到关注，尤其是在优化文献中。这是因为，尽管自动微分技术的重大进展（Baydin 等人，2018年；Margossian，2019）使基于梯度的算法更易于实现，但在实际应用中，出于理论或计算方面的原因，应用这些技术仍并非总是简单可行。文献中讨论的例子（Nesterovand Spokoiny，2017年；Kozak 等人，2021年）包括：i）似然通过复杂的黑盒计算机程序获得，而这些程序不具备计算导数的功能（至少无法在不付出大量人力和时间的情况下实现）；ii）目标分布并非处处可微。此外，即使可以使用反向模式自动微分，以不超过四次目标函数评估的代价获得完整梯度（Nesterov and Spokoiny，2017年），这种方法仍需要存储多个中间步骤。在复杂模型中，这可能导致内存溢出，使得基于梯度的方法难以在这些情境中实现（Kozak 等人，2021年）。

本文采用一种受零阶优化领域近期进展启发的视角（例如，参见Duchi 等人，2015年；Nesterov and Spokoiny，2017年；Kozak 等人，2021年）。在这一领域，越来越多的证据表明，旨在模仿一阶方法的零阶算法，比完全不依赖导数的方法更高效。这些方法将经典的梯度下降推广为用零阶随机估计量替代精确梯度；该估计量通常通过沿随机方向对偏导数进行有限差分近似构造。

我们研究这种方法能否有效拓展至采样任务。我们的贡献有两方面。首先，我们表明，将优化领域中提出的策略直接用于采样，并不能得到高效的并行采样算法，详见第3节。其次，我们提出一种简单而有效的零阶采样方法，以克服这些低效问题，详见第4节。我们的方法针对高维场景设计，其中可用的并行处理器数量 m 严格小于目标分布的维度 d 。

1.1 论文的贡献

本文的组织结构如下。第 2 节介绍零阶框架，概述主要的技术假设（第2.1节），并回顾基于梯度零阶随机估计量的优化策略（第2.2节）。第 3 节探讨将第2 节所述方法直接扩展到采样的可能性，涵盖未校正方法和 Metropolis 校正方法。具体做法是在未校正朗之万算法（ULA）中用零阶估计替代精确梯度（Durmus 和 Moulines，2019年），并将其推广到相应的 Metropolis 校正方法（MALA）（Roberts 和 Tweedie，1996年）。研究表明，这些方法在依赖于 d 和 m 方面3.1和3.2节，分别表现欠佳。

基于上一节的见解，第4 节介绍了我们在方法上的主要贡献：零阶随机切片采样器。我们不再使用目标函数评估来构造全梯度的估计，而是提出在每次迭代时，将提议分布限制在一个随机的 m 维子空间上。在该子空间中，梯度可以通过

有限差分进行确定性近似。这一简单的改进解决了第3节指出的效率低下问题。在第4.2节中，我们将随机切片采样器与零阶版本的哈密顿蒙特卡罗（HMC）（Neal 等人，2011）相结合。第4.3 节研究了零阶随机切片采样器的理论性质，并证明它能随着并行工作器数量的增加，实现效率的多项式改进，从而显著优于现有的零阶并行 MCMC 方法。第5节通过两个贝叶斯回归模型展示了所提方法的实证表现。所有证明均见附录A。代码可在以下网址获取：
https://github.com/Francesco16p/ZO_并行_采样

1.2 MCMC 步内并行化的相关前期工作

当关注点是采样且只有一个计算单元可用时，随机游走梅特罗波利斯（RWM）马尔可夫链蒙特卡洛算法可谓是零阶采样器的黄金标准。若实现正确，该算法每次迭代只需评估一次势函数。若有多个并行处理器可用，情况则更加多样。许多已提出的步内并行化方法可视为随机游走 Metropolis 的推广，属于多提议族（Liu 等，2000；Tjelmeland，2004；Calderhead，2014；Craiu 和 Lemieux，2007；B´edard 等，2012；Casarin 等，2013；Martino 和 Louzada，2017；Fontaine 和 B´edard，2022；Gagnon 等，2023；Yang 和 Liu，2023）。这些算法具有共同的结构：每次迭代都会提出一组点，作为马尔可夫链可能的下一状态；随后，以一定概率从中选出一个点，以确保转移核相对于目标分布保持不变。重要的是，这种构造使算法的每一步都可以并行化。设计得当时，提出的点越多，性能就越好（Gagnon 等，2023）。然而，尽管这些技术在低维或中等维度问题中表现良好，近期的理论研究（Pozza and Zanella，2025；Caprio 等，2025）指出，在高维场景中，增加并行处理器数量 m 所带来的提升通常有限。事实上，在高维对数凹目标分布等重要情形下，多提议马尔可夫链蒙特卡洛的效率提升至多为 m 中的多对数级（Pozza and Zanella，2025）。

与多提议方法相比，另一种步内并行化形式是预取（Brockwell，2006 年）。该算法利用并行工作器，通过评估随机游走 Metropolis 马尔可夫链的所有可能轨迹，提前对其进行模拟。尽管这种方法与多提议马尔可夫链蒙特卡罗所采用的方法有很大不同，但目标函数评估次数会随马尔可夫链长度呈指数级增长。这意味着，预取带来的改进相对于并行工作器数量仅呈对数级增长（Brockwell，2006 年）。因此，利用步内并行化提高效率，并获得显著优于 RWM 算法的马尔可夫链蒙特卡罗方法，基本上仍是一个尚未解决的问题。

最近，Grazzi 和 Zanella (2025) 利用 Picard 映射开发了一种用于 RWM 的步内并行化方法，其性能随处理器数量呈多项式级提升 m。在此，我们从另一种视角出发，借鉴零阶优化领域的最新进展，也实现了 m的多项式级提升。

2 零阶并行采样

2.1 抽象框架

令 $\mathcal{B}(\mathbb{R}^d)$ 表示 $\mathbb{R}^d$ 上的 Borel σ-域，并用 $\mathcal{P}(\mathbb{R}^d)$ 表示 $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ 上概率测度的集合。令 π 为 $\mathcal{P}(\mathbb{R}^d)$ 中的一个目标概率测度，并稍微滥用一下记号，

算法 1. m 维有限差分方向导数

输入： $x \in \mathbb{R}^d$、$m \in \mathbb{N}$ 和
$\epsilon > 0$采样 $\mathbf{V} \sim \nu$计算

$$\frac{U(x + \epsilon v^{(i)}) - U(x)}{\epsilon}, \qquad \text{for} \quad i = 1, \ldots, m. \tag{1}$$

我们用 $\pi(x) = \exp(-U(x))$ 表示，$x \in \mathbb{R}^d$，其相对于勒贝格测度的概率密度函数。我们研究在以下假设下从 π 生成样本的问题：

(i) 可以使用针对 U 的零阶预言机，即可以计算 U，但无法计算其导数。请注意，该框架的独特之处不在于 U是否存在导数，而在于能否显式计算这些导数。

(ii) 可以使用 $m \geq 1$ 个处理器并行地在 m 个不同位置计算 U .

下文中，m 对 U 进行的并行计算称为并行轮次(Anari 等人，2024）。为简便起见，我们假设处理器之间的通信成本可以忽略不计，因此一次并行轮次的计算成本等于对 U进行一次顺序计算的成本。

在优化和采样中，只要适用，都可以利用梯度信息加速收敛。因此，推导出优秀零阶算法的一种可行策略是：利用零阶查询对 U 的导数进行数值估计，从而实现模仿基于导数的方法的方案。在串行应用中，与原生一阶算法相比，这些方法通常计算成本更高 Kozak 等人(2021年）。不过，如果有多个并行工作器可用，其算法中的部分计算可以并行完成，从而显著加快算法。

我们考虑使用算法 1 中给出的方向导数有限差分近似的算法。给定一点 $x \in \mathbb{R}^d$、一个矩阵 $(v^{(1)}, \ldots, v^{(m)}) = \mathbf{V} \in \mathbb{R}^{d \times m}$，其中包含 m 个感兴趣的方向，以及一个足够小的步长 $\epsilon > 0$，算法 1 依据导数定义为增量比的极限这一性质，估计 U 在 x 处相对于方向 V 的偏导数向量。令 $\epsilon \to 0$趋于极限，则式 (1) 中的每个增量比都收敛到 U 在x处沿方向 $v^{(i)}$的相应方向导数 $\partial_{v^{(i)}} U(x)$，即 $\partial_{v^{(i)}} U(x) = \lim_{\epsilon \to 0}(U(x + \epsilon v^{(i)}) - U(x))/\epsilon$。对于足够规则的优化问题，算法 1 中固定 ϵ 引入的误差通常可以忽略不计，相较于确定最优点所需的精度而言尤为如此（Nesterov and Spokoiny，2017）。在采样中，这个问题或许更无关紧要，因为导数信息并非用于确定某个特定的最优点，而是用于引导算法前往概率较高的区域。方向导数的另一个重要性质是，当梯度 $\nabla U(x)$存在时，有 $\partial_{v^{(i)}} U(x) = \nabla U(x)^\top v^{(i)}$.。也就是说，每个方向导数都等于 U 的梯度与所选方向的内积。这一恒等式极大地简化了依赖方向导数的方法的理论分析，因为它将涉及 ∇U性质的部分与取决于 V 生成方式的部分区分开来。下文为采用这一方法，将忽略算法 1 中的近似误差，并直接考虑 $\partial_{v^{(i)}} U(x)$，以替代其有限差分近似 (1)，例如 Nesterov and Spokoiny（2017）所采用的做法。

我们以简要讨论算法 1 是否适用于并行计算来结束本节。假设此前已获得 $U(x)$，算法 1 还需要 m 次额外

次目标函数求值。不过，如果有 m 个并行工作器，则可独立并行地执行式（1）中的每项操作，从而只需一个并行轮次。因此，若忽略通信和同步开销，使用 m 个处理器，算法 1 的执行时间基本上与调用一次零阶预言机相同。

2.2 与零阶优化的比较

文献中对零阶优化进行了广泛研究，尤其关注随机梯度算法（Nesterov and Spokoiny，2017年；Kozak 等人，2021年，2023年）。这些方法利用方向导数（可通过算法 1 进行估计）构造梯度的无偏估计 U，再将其纳入梯度下降类算法，以求得 $x_* = \arg\min_{x\in\mathbb{R}^d} U(x)$。鉴于采样与优化之间存在诸多相似之处，本节将简要总结后者的重要结果，旨在突出高效的零阶采样器应具备的一些基本性质，尤其是关于目标维度 d 和并行工作器数量 m方面的算法效率。

给定一个起始点 $x_0 \in \mathbb{R}^d$，通用零阶随机梯度下降算法的一次迭代 $t \in \mathbb{N}$ 可分为以下两个步骤：

$$\begin{aligned} &1.\ \text{Generate } \mathbf{V} \in \mathbb{R}^{d\times m} \text{ from a probability distribution } \nu, \\ &2.\ \text{Set } x_t = x_{t-1} - l_t \hat{\nabla}_{\mathbf{V}} U(x_{t-1}), \end{aligned} \tag{2}$$

其中，$\hat{\nabla}_{\mathbf{V}} U(x) = c_\nu \mathbf{V}\mathbf{V}^\top \nabla U(x)$，对于某个 $c_\nu \in \mathbb{R}$，使得对任意 $\mathbf{V} \sim \nu$，均有 $\mathbb{E}[\hat{\nabla}_{\mathbf{V}} U(x)] = \nabla U(x)$，它是梯度的零阶随机估计量，且 l_t 是学习率。注意到，根据等式 $\partial_{v^{(i)}} U(x) = \nabla U(x)^\top v^{(i)}$，通过将算法 1 与对 ν 和 c_ν 的适当选择相合，可以得到满足上述关于 $\hat{\nabla}_{\mathbf{V}} U(x)$ 条件的估计量（若忽略选择 $\epsilon > 0$所产生的近似误差）（Kozak 等人，2021，式 3）。

对于两个正函数 $f, g: \mathbb{R}^+ \to \mathbb{R}^+$，我们称 $f = O(g)$，如果存在 $x_0 \in \mathbb{R}^+$ 和一个正常数 C，使得对任意 $x \geq x_0$，均有 $f(x) \leq Cg(x)$。当 $m = 1$时，已知零阶随机梯度方法需要 $O(\kappa d \log(1/\epsilon))$ 次迭代，才能使强凸光滑函数达到 ϵ误差，其条件数为 κ；定义见假设 1。相比之下，在相同条件下，一阶方法只需 $O(\kappa \log(1/\epsilon))$ 次迭代（Bach，2024，第 5 章）。这意味着，用一个 $m = 1$ 零阶估计量替代精确梯度，会引入一个量级为 d的额外成本因子。当 $m > 1$ 且 $m \leq d$ 时，迭代次数会降至 $O(\kappa d/m \log(1/\epsilon))$（Kozak 等人，2021，推论 1）；这意味着，每轮增加偏导数评估次数，可以将收敛时间缩短 m倍。这一事实对零阶优化方法的串行实现和并行实现都有重要的实际意义。事实上，使用 $m^* > 1$个方向迭代 t 次所达到的精度，等同于使用 $m = 1$迭代 $m^* \times t$ 次所达到的精度。因此，在串行实现中，对于相同数量的目标函数评估，不同的 m 值会产生效率相当的算法。相反，在并行环境中，如果有 m个工作节点，则只需进行 t/m个并行轮次，即可达到在单个处理器上迭代 t 次所获得的相同精度。

这种线性加速远高于常见的步内并行马尔可夫链蒙特卡洛方法所呈现的对数加速（Brockwell，2006 年；Pozza and Zanella，2025）。本文旨在开发零阶并行 MCMC 方法，使其相对于 m 的改进更接近优化领域中观察到的情况，即至少随 m呈多项式增长。

我们通过更详细地讨论关于

V 生成的各种选择来结束本节。常见的选择包括相互独立的 d 维高斯分布，以及在 d 维单位球面上独立均匀分布的向量（在 d 维单位球面上均匀分布的随机变量，可以通过从 $Z \sim N_d(0, \mathbb{I}_d)$ 采样并取 $\mathrm{V} = Z/\|Z\|$ 得到（Vershynin，2018，第 68 页））。尽管计算上很有吸引力，但抽取 m 个独立方向会导致一个不理想的性质： $m \geq d$ 个方向导数无法精确重构梯度，因为一般而言，$\mathrm{VV}^\top \neq \mathbb{I}_d$。可以考虑不同的方案来解决这一问题。下文中，我们假设 $\mathrm{V} \in \mathbb{V}_m(\mathbb{R}^d)$，其中 $\mathbb{V}_m(\mathbb{R}^d) = \{\mathrm{V} \in \mathbb{R}^{d\times m}, \mathrm{V}^\top\mathrm{V} = \mathbb{I}_k\}$ 表示 m 维的 Stiefel 流形，定义在 $\mathbb{R}^d$ 上。通过对具有独立同分布 $N_d(0, \mathbb{I}_d)$ 列的 $d \times d$ 随机矩阵进行 QR分解，可以以 $O(m^2 d)$ 的计算成本，从均匀分布 $\nu = \nu_{\mathrm{unif}}$ 中生成 $\mathbb{V}_m(\mathbb{R}^d)$ 的样本（Mezzadri，2006 年，第 4 节）。请注意，除一个乘法修正因子外， $\nu = \nu_{\mathrm{unif}}$ 等价于例如 Kozak 等人（2021，2023）所作的假设。或者，一种计算量较小的方案是考虑 $\nu = \nu_{\mathrm{E}}$ ，它在每次迭代中从标准基 $\mathrm{E} = (e^{(1)}, \ldots, e^{(d)})$ 中随机选取一个m 维子集，其中 $e_i^{(i)} = 1$ 且 $e_j^{(i)} = 0$，对于 $i \neq j, j, i = 1, \ldots, d$。在 (2) 中，这种对 V 的选择类似于随机坐标下降（Aggarwal，2020，第 194 页）；而在采样中，它对应于 Gibbs型MCMC算法（Casella 和 George，1992；Gelfand，2000）。请注意，当 $m = d$, $c_\nu \mathrm{VV}^\top \nabla U(x) = \nabla U(x)$ 对 $\nu = \nu_{\mathrm{unif}}$ 和 $\nu = \nu_{\mathrm{E}}$ 都成立时，若 $c_\nu = d/m$，则忽略有限差分近似带来的误差后，算法 1 能够精确重构梯度。

3 朴素插件方法

3.1 未校正方法

一旦获得了 $\nabla U(x)$，一个自然的问题就是如何利用它来设计高效的采样算法^。一个可行的方法是在任意基于梯度的采样器中，用 $\nabla U(x)$ 替代 $\nabla_{\mathbf{V}} U(x$。通过在每次迭代中生成 $\mathbf{V}$ 的不同实现，这一策略原则上既适用于未校正的 Durmus 和 Moulines(2019年) 方法，也适用于经校正的（Roberts 和 Tweedie，1996年；Neal 等人，2011；Livingstone和Zanella，2022）基于梯度的马尔可夫链蒙特卡洛方法。

本节讨论未校正框架，其中考虑将 ~~V 取为~~ 朗之万算法（ULA）中（Roberts and Stramer，2002年）。在算法2中，在第 t 次迭代时，采样得到一组 m 个随机方向 $\mathbf{V}$ 。给定 $\mathbf{V}$ 后，执行一步 ULA，并用按（2）所述方法获得的随机估计替代精确梯度。

算法 2。零阶未校正朗之万算法

输入： $T \in \mathbb{N}$, $\gamma > 0$, $c_\nu \in \mathbb{R}$, $\mu \in \mathcal{P}(\mathbb{R}^d)$以及定义在 $\mathbb{V}_m(\mathbb{R}^d)$上的 ν 分布。

$X_0 \sim \mu$**对于** $t = 1, \ldots, T$**执行**采样$\mathbf{V} \sim \nu$ 和 $Z \sim N_d(0, \mathbb{I}_d)$，设置 $X_t = X_{t-1} - \gamma$^ $\nabla_{\mathbf{V}} U(X_{t-1}) + \sqrt{2\gamma} Z$，其中 ^ $\nabla_{\mathbf{V}} U(X_{t-1}) = c_\nu \mathbf{V}\mathbf{V}^\top \nabla U(X_{t-1})$

输出： 马尔可夫链轨迹（$X_0, X_1, \ldots, X_T$）。

未校正马尔可夫链蒙特卡罗算法的理论分析通常涉及两个不同方面（Durmus 和 Moulines，2019年）。首先，给定一个特定的度量指标（通常是二阶 Wasserstein 距离或全变差距离），需要确定该过程收敛到其平稳分布的速度，π_γ。其次，必须量化π_γ与 π 之间的距离。下文我们将论证，至少从理论层面来看，算法2 并不是一种特别理想的并行计算资源利用方式。这些

低效之处主要源于算法的不变分布对目标分布的近似方式，而非算法达到平稳状态的速度。

对于两个概率测度 μ，$\eta \in \mathcal{P}(\mathbb{R}^d)$，概率测度 ξ 在 $(\mathbb{R}^d \times \mathbb{R}^d,\ \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R}^d))$ 上是 μ 和 η 的一个耦合，当且仅当对于每个可测集 $A \subseteq \mathbb{R}^d$，$\xi(A \times \mathbb{R}^d) = \mu(A)$ 且 $\xi(\mathbb{R}^d \times A) = \eta(A)$。$\mu$ 和 η 的所有耦合构成的集合记为 $\Xi(\mu, \eta)$。最后，我们将二阶 Wasserstein 距离定义为

$$W_2(\mu, \eta) = \sqrt{\inf_{\xi \in \Xi(\mu,\eta)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \xi(dx, dy)}.$$

为了研究算法 2在二阶 Wasserstein 距离下朝其平稳分布收敛的收缩性质，我们需要两个额外的假设：第一个关于 π，第二个关于 ν。

假设 1。 $\pi \in \mathcal{P}(\mathbb{R}^d)$ 具有密度 $\exp(-U)$，其中 $U:\ \mathbb{R}^d \to \mathbb{R}$ 为

1. *L*-光滑，即 $\nabla U:\ \mathbb{R}^d \to \mathbb{R}^d$ 是 *L-Lipschitz* 连续的。
2. λ-凸，即 $U(x) - (\lambda/2)\|x\|^2$ 是凸的。
3. 二阶连续可微。

下文用 $\kappa = L/\lambda$ 表示 π 的条件数。

假设 2。 $c_\nu = d/m$ 且 ν 满足 $(d/m)\mathbb{E}[\mathbf{V}\mathbf{V}^\top] = \mathbb{I}_d$ 对于 $\mathbf{V} \sim \nu$.

假设 1 中的凸性和光滑性条件在马尔可夫链蒙特卡洛方法的理论分析中很常见（Chewi, 2024）。这些条件既允许得到明确的定量界，又涵盖了大量非平凡的分布。假设 2 则更针对当前问题，刻画了 $\mathbf{V}$ 的抽样机制。类似假设 2 的条件可见于 Ascolani 等人 (2024，第 8 节）以及随机优化文献（Kozak 等人，2021）。满足假设 2 的分布包括 $\mathbf{V}_m(\mathbb{R}^d)$ 上的均匀分布，以及通过从 m 个 $\mathbb{R}^d$的自然基向量中均匀无放回抽取得到的分布。

命题 1. 令 P^{SULA} 为算法 *2*的转移核。那么，在假设 *1*和 *2*成立的条件下，对于每个 x，$y \in \mathbb{R}^d$ 和 $\gamma \le m/(Ld)$，均有

$$W_2(\delta_x P^{SULA}, \delta_y P^{SULA}) \le (1 - \gamma\lambda)^{1/2} \|x - y\| \le \exp\left(-\frac{1}{2}\gamma\lambda\right) \|x - y\|, \tag{3}$$

其中，δ_x、δ_y 分别表示在 x 和 y上的狄拉克测度。

对于任意 $\epsilon > 0$，命题 1 表明，取 $\gamma = m/(Ld)$ 时，算法 2需要 $n_* = O(\kappa d/m \log(1/\epsilon))$ 个并行轮次，才能从分布 ϵ中生成一个与其自身平稳分布相差 π_γ的样本。这一速率与类似优化场景中获得的最佳速率相当（Kozak 等人，2021年，推论 1），这意味着在此设置下，就收敛速度而言，从优化转向采样并不会变慢。然而，与标准 ULA 的情况类似，如果步长 π_γ 未经过适当缩放，算法 2 的平稳分布可能会与目标分布π 存在显著差异。如果步长 γ 未经过适当缩放，算法所需的条件已在类似背景下由 Roy 等人 (2022) 进行了研究。他们的框架涵盖 π 仅能评估到带有随机误差的情形，但其中的结果也适用于我们的背景，因为在此处该误差

为零。他们的研究结果表明，类似于算法 2 的方案需要 $O(d^2/m)$ 个并行轮次，才能从目标分布 π 的精确近似中进行采样，详见其中的定理 2.1。鉴于命题 1，这一结果主要源于需要采用较小的步长，具体而言为 $\gamma = O(m/d^2)$，以确保平稳分布 π_γ具有较小偏差。在类似条件下，标准随机游走 Metropolis 只需 $O(\kappa d \log(1/\epsilon))$ 次序列零阶评估即可对 U 进行精确采样（Andrieu 等人，2024，定理 49），因此，至少从理论上看，算法 2 代表了在这一零阶背景下将并行计算用于采样的一种次优方式。

3.2 Metropolis 校正方法的朴素方法

第 3.1 节指出，算法 2 中发现的低效率主要归因于其平稳分布与目标分布之间的差异。因此，一种可能避免这一问题的方法是改用 Metropolis 调整算法。本节将进一步探究这种可能性。

ˆ 算法 3 说明了如何将 $\nabla_{\mathbf{V}} U(x)$ 纳入 Metropolis 校正朗之万算法（MALA）中。在第 t次迭代时，采样一组 m 随机方向 $\mathbf{V}$ 。在给定 $\mathbf{V}$的条件下，执行一步 MALA，将精确梯度替换为按照（2）获得的随机估计。

算法 3。朴素零阶 Metropolis 校正朗之万算法

输入： $T \in \mathbb{N}$、$\sigma^2 > 0$、$\mu \in \mathcal{P}(\mathbb{R}^d)$ 以及 ν 在 $\mathbb{V}_m(\mathbb{R}^d)$ 上的分布

$X_0 \sim \mu$

for $t = 1, \ldots, T$ **do**

Given $X_t = x$, sample:

$\mathbf{V} \sim \nu$,

$y \sim Q_{\mathbf{V}}(x, \cdot)$ where $Q_{\mathbf{V}}(x, \cdot) = N_d(x - (\sigma^2/2)\hat{\nabla}_{\mathbf{V}} U(x), \sigma^2 \mathbb{I}_d)$.

Set $X_{t+1} = y$ with probability

$$\min\left(1, \frac{\pi(dy) Q_{\mathbf{V}}(y, dx)}{\pi(dx) Q_{\mathbf{V}}(x, dy)}\right),$$

$X_{t+1} = X_t$ otherwise.

输出： 马尔可夫链轨迹（X_0，X_1，……，X_T）

等价地，与算法3 相关联的转移核可写为

$$P_\sigma^{SMALA}(x, A) = \int_{\mathbb{R}^d \times \mathbb{V}(\mathbb{R}^d)} \mathbb{1}_A(y) Q_{\mathbf{V}}(x, dy) \min\left(1, \frac{\pi(dy) Q_{\mathbf{V}}(y, dx)}{\pi(dx) Q_{\mathbf{V}}(x, dy)}\right) \nu(d\mathbf{V}), \tag{4}$$

可以很容易证明，如下文命题 π所示，具有 2-不变性。

命题 2。 P_σ^{SMALA}在 (4) 下 π具有 -不变性。

算法 3展示了一种相当直接的方法，可将（2）中的梯度估计纳入 Metropolis 调整的基于梯度的马尔可夫链蒙特卡洛。在本节中，我们将正式证明，与第 3.1 节讨论的未调整情形一样，除非 m 接近 d，否则这种方法效率低下。下文中，$\mathbf{V}^\perp$ 表示一个任意的 $d \times (d-m)$ 矩阵，其列构成 span(**V**) 的正交补的一组正交归一基。

引理 1。 设 $\mathbf{V} \in \mathbb{V}_m(\mathbb{R}^d)$， $x \in \mathbb{R}^d$ 且 $y \sim Q_{\mathbf{V}}(x, \cdot)$，其中 $Q_{\mathbf{V}}$ 如算法 3 中所述。那么， $s = \mathbf{V}^\top y$和 $s^\perp = (\mathbf{V}^\perp)^\top y$ 是相互独立的随机变量（对于固定的 x 和 $\mathbf{V}$），并且

$$s \sim N_m\big(\mathbf{V}^\top x - (\sigma^2/2)\mathbf{V}^\top \hat{\nabla}_{\mathbf{V}} U(x), \sigma^2 \mathbf{V}^\top \mathbf{V}\big), \tag{5}$$

$$s^\perp \sim N_{d-m}\big((\mathbf{V}^\perp)^\top x, \sigma^2 \mathbb{I}_{d-m}\big). \tag{6}$$

引理 $Q_{\mathbf{V}}(x, \cdot)$ 中的分解表明，算法 3的提议分布在由 $\mathbf{V}$ 张成的子空间及其正交补上的行为有所不同，这一区别可能会对其整体行为产生重要影响。随着 m 增大，算法3 预计将等价于预条件化版本的 MALA（Roberts and Stramer，2002年；Girolami and Calderhead，2011）。相反，若 $d - m$ 不够小，则（ 6 ）中的随机游走分量可能会主导算法的性能。因此，理解（ 5 ）与（ 6 ）之间的相互作用如何影响算法的整体效率至关重要。

下面的定理 1 通过研究 P_σ^{MALA}的右谱隙如何受 m影响，进一步探讨这一问题。回顾一下，对于一个 π-可逆转移核 P，右谱隙定义为

$$Gap_P = \inf_{f \in L^2_\pi} \frac{\int_{\mathbb{R}^{2d}} \{f(x) - f(y)\}^2 \pi(dx) P(x, dy)}{2\mathbb{V}_\pi(f)}, \tag{7}$$

其中 $\mathbb{V}_\pi(f) = \int_{\mathbb{R}^d} f(x)^2 \pi(dx) - \{\int_{\mathbb{R}^d} f(x)\pi(dx)\}^2$，而 L^2_π 是所有满足 $\mathbb{V}_\pi(f) < \infty$ 的函数 f: $\mathbb{R}^d \to \mathbb{R}$ 的集合。右谱隙量化了与 P 相关的两个最大正特征值之间的差异，对于正定核而言，它等价于谱隙。对于非正定核，右谱隙与谱隙通常不同，但任何转移核 P 都可以通过考虑其懒惰版本 $P_{\text{LAZY}} = (P + I)/2$与一个正定核联系起来。在马尔可夫链蒙特卡罗理论中，谱隙通常被视为衡量 π-可逆核效率的标准指标（Levin and Peres，2017年，第12节）；谱隙越大，性能通常越好。

定理 1。在假设 1-2成立的条件下，若 ν 满足其任意一列都是$\mathbf{V}$ 标准基 $\mathbf{E}$中的一个元素

$$\sup_{\sigma > 0} Gap_{P_\sigma^{SMALA}} = O\Big(\frac{\log(d-m)^2}{d-m}\Big) = O\Big(\frac{\log(d)^2}{d}\Big), \tag{8}$$

如 $d \to \infty$ 和 $m < c_0 d$，其中 $c_0 \in (0, 1)$。

该定理的证明基于引理 P_σ^{SMALA}中给出的提议分布分解1，并结合了 Andrieu 等人(2024) 以及 Pozza and Zanella (2025) 针对 RWM 开发的非渐近论证。定理 1 表明，当 P_σ^{SMALA}经过最优调优时，算法 3 的右谱隙至多为 $1/(d-m)$ 的阶（相差一个多对数项）。在类似条件下，Andrieu 等人 (2024) 给出了高斯分布随机游走 Metropolis 算法在 d维目标分布上的阶为 $1/d$。因此，当 m 相较于 d较小或适中时，算法3 的表现本质上类似于随机游走 Metropolis 算法，因而并未有效利用并行计算。

定理1 在图1中进行了说明。该图针对第5.1 节中的一个 200 维贝叶斯逻辑回归示例，并展示了不同 m取值下，算法3（NAIVE-MALA）相较于随机游走 Metropolis 所获得的增益，并将其与随机切片 MALA（RS-MALA）算法进行比较。后者将在下文第4 节中定义，旨在克服第3节所分析方法的局限性。对于 NAIVE-MALA，即使在 $m \simeq 100$时，相较于随机游走 Metropolis 的改进也很小（≤ 5）；而在相同的 m 取值下，RS-MALA 的效率已是随机游走 Metropolis 的 30–35 倍

随机游走 Metropolis。图 1 还表明，定理 $O(1/(d-m))$ 中的上界很好地描述了实践中观察到的行为。关于图 1 中所示改进幅度的正式定义，请参见下文的（12）。1。

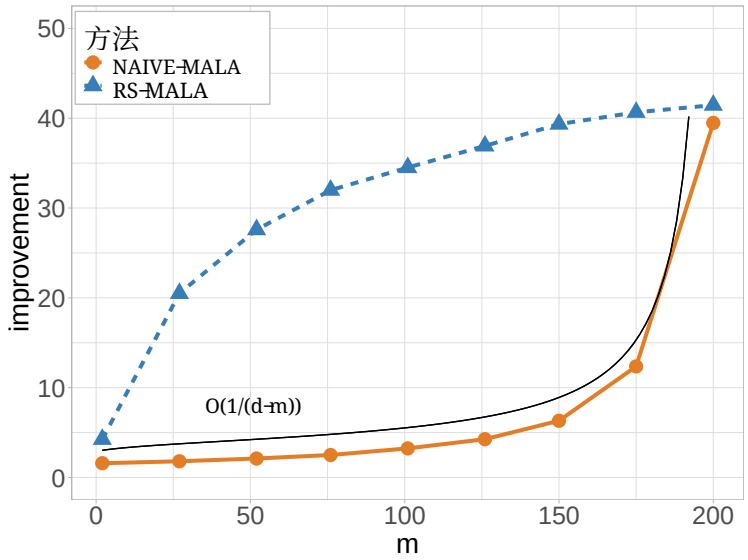


图 1：相对于随机游走 Metropolis 的改进幅度，以估计的平均期望平方跳跃距离（定义见（12））衡量；对象为 NAIVE-MALA 和 RS-MALA，采用第 5.1 节中的 200 维逻辑回归示例。

4 随机切片方法

4.1 一般方法

第3 节的核心结论是，尽管将零阶随机梯度估计量直接代入马尔可夫链蒙特卡洛在优化中效果良好，但随着并行工作器数量增加，这样做可能导致算法性能不佳。具体原因因算法而异，但都源于第3 节介绍的方法对随机误差 ˆ ˆ$\nabla_{\mathbf{V}} U(x)$缺乏稳健性。如第3.2节所示，在 Metropolis 调整方案中，这一问题的出现是因为 $\nabla_{\mathbf{V}} U$完全位于 m维子空间 span(V)中。因此，该提议无法提供关于 U 在正交补 span($\mathbf{V}^{\perp}$)上的行为的任何信息；在这个子空间上，它的表现就像简单的随机游走 Metropolis。

受第3节相关见解的启发，本节提出将提议分布限制为仅作用于 span(V)，从而完全移除随机游走分量。在实践中，这等价于将 Metropolis-Hastings 提议的噪声与随机梯度估计量的噪声相关联，使随机性仅注入梯度能够被可靠估计的方向。例如，为修正第3.2节所述的零阶 MALA，这归结为使用以下提议分布：

$$y = x - (\sigma^2/2)\mathbf{V}\mathbf{V}^{\top}\nabla U(x) + \sigma \mathbf{V} z, \qquad \text{with} \qquad z \sim N_m(0, \mathbb{I}_m),$$

其中 m维噪声 z 通过项 $\sigma \mathbf{V} z$被投影到 span($\mathbf{V}^{\perp}$) 中 .

更一般地，这种随机切片方法可以与除 MALA 之外的任何基于梯度的 Metropolis-Hastings 提议分布结合使用，如下文针对 HMC 所做的那样，也可以与任何其他高精度的基于梯度的采样器结合使用。一般方法在算法 4 中进行了描述，其中使用以下记号。给定 $x \sim \pi$，令 s和 $s^{\perp}$ 分别表示定义为 $s^{\perp} = (\mathrm{V}^{\perp})^{\top} x$ 和 $s = \mathrm{V}^{\top} x$的随机变量。将给定 $s^{\perp}$ 和 V 的 s 的条件分布定义为 $\pi_{\mathbf{V},s^{\perp}}(s) \propto \exp\{-U_{\mathbf{V},s^{\perp}}(s)\}$，其中对于 $s \in \mathbb{R}^m$，$U_{\mathbf{V},s^{\perp}}(s) = U(\mathrm{V}^{\perp} s^{\perp} + \mathrm{V}s)$。此外，令 $\eta_{\mathbf{V},x}$表示 $\pi_{\mathbf{V},s^{\perp}}$的某个通用近似的密度，其中在 $\eta_{\mathbf{V},x}$ 中的下标 x 用于强调提议可能依赖于马尔可夫链的当前状态。

算法 4。 m 维随机切片采样器

输入： $T \in \mathbb{N}$, μ, ν, $\{\eta_{\mathbf{V},x} \in \mathcal{P}(\mathbb{R}^m)\}_{\mathbf{V} \in \mathbb{V}_m(\mathbb{R}^d),\ x \in \mathbb{R}^d}$
$X_0 \sim \mu$**对于** $t = 1, .., T$**执行**给定 $X_{t-1} = x$，采样 $\mathbf{V} \sim \nu$ 并设置 $s^{\perp} = (\mathbf{V}^{\perp})^{\top} x$采样 $s' \sim \eta_{\mathbf{V},x}$设置 $X_t = \mathbf{V}^{\perp} s^{\perp} + \mathbf{V} s'$**输出：** 马尔可夫链轨迹（$X_0, X_1, \ldots, X_T$）。

给定当前状态 $X_t = x$，执行以下两个步骤：(i) 从 ν 中采样一组随机的 方向 $\mathbf{V}$，以及 (ii) 仅根据 $\eta_{\mathbf{V},\ \text{更新}\ x}$ 在 span(V) 上的投影x。等价地，算法 4 的转移核可表示为

$$P(x, A) = \int_{V_m(\mathbb{R}^d)} \int_{\mathbb{R}^m} \mathbb{1}_A(\mathbf{V}^{\perp}(\mathbf{V}^{\perp})^{\top} x + \mathbf{V} s') \eta_{\mathbf{V},x}(ds') \nu(d\mathbf{V}), \tag{9}$$

对于 $x \in \mathbb{R}^d$ 和 $A \subseteq \mathbb{R}^d \backslash \{x\}$。

算法4需要指定$\eta_{\mathbf{V},\ x}$的形式。实践中，我们建议将$\eta_{\mathbf{V},\ x}$ 作为针对$\pi_{\mathbf{V},\ s^{\perp}}$的基于梯度的采样器所进行的一次更新。采用这一选择后，算法4 可解释为一个 m维 hit-and-run 采样器中的基于梯度的条件更新（Chen 和 Schmeiser，1996年）。关键在于，由于$\pi_{\mathbf{V},\ s^{\perp}}$ 是 m维的，因此可以使用算法1 在每次迭代中通过一个并行轮次的 m 零阶评估来数值估计其梯度。下文将说明，与前几节讨论的方法相比，式（9）定义的马尔可夫核对m的依赖性有所改善。

4.2 P提出的实用方法：随机切片零阶 HM C

任何可表示为式（9）形式的算法，其性能都取决于条件提议分布$\eta_{\mathbf{V},\ x}$的质量。本节将讨论一种我们发现实践中尤为有效的方案，即将算法4 与Neal 等人（2011）提出的 HMC 零阶版本应用于 m 维切片。具体而言，给定 $X_t = x$，算法5 的每次迭代包括以下步骤：（i）从提议分布 ν中采样一组新的方向$\mathbf{V}$；（ii）以目标分布$\pi_{\mathbf{V},\ s^{\perp}}$ 执行一次 HMC 步骤，并使用$U_{\mathbf{V},\ s^{\perp}}$的梯度；该梯度按照算法1中的方法通过有限差分计算。

算法 5 的一些方面值得进一步讨论。的支持集是 $\mathbb{V}_m(\mathbb{R}^d)$。因此，原则上有许多不同的选择。我们认为，最具实践吸引力的选择是 $\mathbb{V}_m(\mathbb{R}^d)$上的均匀分布（记作 $\nu = \nu_{\text{unif}}$）以及 $\nu = \nu_{\text{E}}$，它会从 $\mathbb{R}^d$ 的标准基中随机选取 m个方向（另见第2节）。前者可能具有理论优势，例如在旋转重参数化下保持不变；而后者则允许

算法 5。随机切片零阶哈密顿蒙特卡罗

输入： $X_0 \in \mathbb{R}^d$、ϵ、$\gamma > 0$、m、L、$T \in \mathbb{N}$、U 在 $\mathbb{V}_m(\mathbb{R}^d)$上的目标分布的势函数和 ν 分布。

for $t = 1, \ldots, T$ **do**

Sample: $\mathbf{V} \sim \nu$ and $k \sim N_m(0, \mathbb{I}_m)$

Given $X_{t-1} = x$ set:

$k_0' = k - \frac{\gamma}{2} \sum_{i=1}^m \left[\frac{U(x+\epsilon v^{(i)}) - U(x)}{\epsilon}\right] v^{(i)}$

$s_0 = 0 \in \mathbb{R}^m$

for $l = 1, \ldots, L$ **do**

$s_l = s_{l-1} + \gamma k_{l-1}'$

if $l < L$ **then**

$k_l' = k_{l-1} - \gamma \sum_{i=1}^m \left[\frac{U(x+\mathbf{V}s_l+\epsilon v^{(i)}) - U(x+\mathbf{V}s_l)}{\epsilon}\right] v^{(i)}$

$k_L' = -k_{L-1} + \frac{\gamma}{2} \sum_{i=1}^m \left[\frac{U(x+\mathbf{V}s_L+\epsilon v^{(i)}) - U(x+\mathbf{V}s_L)}{\epsilon}\right] v^{(i)}$

Set $X_t = x + \mathbf{V}s_L$ with probability

$$\min\left(1, \exp\left(U(x) - U(x + \mathbf{V}s_L) + 0.5(\|k\|^2 - \|k_L'\|^2)\right)\right),$$

or $X_t = X_{t-1}$ otherwise.

Output 和 of ow c ain 轨迹 g、1、...(X、X_T。 X)

以显著更低的成本进行采样。因此，ν_{E} 是我们在实践中的首选。当$\nu = \nu_{\mathrm{E}}$时，算法 5 可以被理解为吉布斯采样中的哈密顿蒙特卡洛采样器的零阶变体。另一个重要之处在于，算法 5 的每次迭代都不会对$\pi_{\mathbf{V}}$执行精确的哈密顿蒙特卡洛更新，$s^{\perp}$因为$U_{\mathbf{V}}$的梯度$s^{\perp}$ 仅进行了数值近似。下面的引理 3 表明，该算法仍保持 π不变性。

Prop命题 3。 算法 $P_{m\,L}^{\text{RS-HMC}}$的马尔可夫转移核 $P_{m\,L}^{\text{RS-HMC}}$，算法 5 的 π不变性成立 t.

命题 3 的证明源于 Tierney (1998) 中的确定性提议示例，并结合以下事实：算法 5 中使用的映射（s_L，k_L'）$= T_L$（s_0，k）是一个对合，且其雅可比矩阵的行列式等于一。

与标准 HMC 算法一样，算法 5 的效率也可能受益于适当的预条件。给定预条件矩阵 $A \in \mathbb{R}^{d\times d}$，一种简单而有效的方法是重新参数化 $y_t = Ax_t$，并在变换空间中执行一次算法 5 的迭代以计算 y_{t+1}。然后，通过取 $x_{t+1} = A^{-1}y_{t+1}$. 将所得点映射回原始空间。第 5 节通过两项模拟研究考察了算法 5的实证性能。

4.3 算法 4 的复杂度理论界

我们通过说明以下内容来结束本节：当与合适的提议分布结合时，算法 4 的性能可随着并行处理器数量的增加而呈多项式级提升。这表明，正如上文图

所示，该方法在实践中可作为一种效率显著更高的替代方案，取代文献中先前讨论的方法以及第 3 节分析的方法。1。

在我们的分析中，必须考虑两个不同的方面：

1. 与任何击中与运行型算法一样，算法 4 的每次迭代仅作用于 m 维子空间 $\mathbb{R}^d$。对此类算法的理论分析需要采用不同于全空间运行的方法所使用的证明技术。在假设 1 下，具有吉布斯型和击中与运行型更新的马尔可夫链蒙特卡洛方法的 Kullback-Leibler 收缩意义下的收敛性质已由Ascolani 等人 (2024) 进行了研究。我们在此基础上，将这些结果调整到本文所考虑的零阶框架中。

2. 该算法受到提议密度 $\eta_{\mathbf{V}}$ 的强烈影响，x既体现在相对于 $\pi_{\mathbf{V}}$ 的$s^{\perp}$近似质量方面，也体现在计算成本方面。因此，区分结构（9）本身所固有的影响与由 $\eta_{\mathbf{V}}$ 的具体形式x所产生的影响非常重要。对此，我们作出如下假设：对于每个 $\delta > 0$、 $x \in \mathbb{R}^d$ 和 $\mathbf{V} \in \mathbb{V}_m(\mathbb{R}^d)$，提议 $\eta_{\mathbf{V}}$ 满足x $\mathrm{KL}(\eta_{\mathbf{V},x} \mid \pi_{\mathbf{V},s^{\perp}}) \leq \delta$，其中

$$\mathrm{KL}(\eta_{\mathbf{V},x} \mid \pi_{\mathbf{V},s^{\perp}}) = \int \log\left(\frac{\pi_{\mathbf{V},s^{\perp}}}{\eta_{\mathbf{V},x}}(s)\right) \eta_{\mathbf{V},x}(ds),$$

是 $\pi_{\mathbf{V},s^{\perp}}$ 与 $\eta_{\mathbf{V},x}$之间的 Kullback–Leibler 散度。对于定义在 $\mathbb{R}^m$ 上的强对数凹且对数光滑目标分布，这类 $\eta_{\mathbf{V},x}$ 是否存在及其计算成本，将在本节最后部分讨论。

在假设 1-2下，定理 4研究了算法 π 在从 P^t 进行 Kullback-Leibler 收缩方面的效率。下文中，P 表示将 9 中定义的 t 应用 $\mu \in \mathcal{P}$次后得到的转移核；对于一般的概率分布 $\mathbb{R}^d$，$\mu P^t\, A = \int \mu\, dy\, P^t\, y,\ A$，其中 $A \subseteq \mathbb{R}^d$。

定理 2. 假设对于每个 $\delta > 0$，$\mathbf{V} \in \mathbb{V}_m(\mathbb{R}^d)$ 以及 $x \in \mathbb{R}^d$，

$$\mathrm{KL}(\eta_{\mathbf{V},x} \mid \pi_{\mathbf{V},s^{\perp}}) \leq \delta, \tag{10}$$

并且假设 *1-2* 成立。那么，转移核 P （定义见 (9) ）满足

$$\mathrm{KL}(\mu P^t \mid \pi) \leq \left(1 - \frac{1}{\kappa}\frac{m}{d}\right)^t \mathrm{KL}(\mu \mid \pi) + \frac{\delta\kappa d}{m}, \tag{11}$$

对于所有 $t \in \mathbb{N}$。因此，如果 $\delta \leq \epsilon m/(2\kappa d)$ 且 $t \geq (\kappa d/m)\log(2\mathrm{KL}(\mu \mid \pi)/\epsilon)$，则有

$$\mathrm{KL}(\mu P^t \mid \pi) \leq \epsilon\,.$$

对于每个固定的 ϵ，定理 2 表明，算法 4 需要 $O(\kappa d/m)$ 次迭代，才能从一个与 π 的 Kullback–Leibler 散度相差 ϵ 的分布中获得样本。因此，改进幅度关于 m 呈线性。然而，这并未计入$\eta_{\mathbf{V},x}$ 的计算成本。

在假设 1 下，在一阶框架中（即假设存在一个预言机，同时提供有关目标分布及其梯度的信息），可以构造一个分布$\eta_{\mathbf{V},x}$ 使得 $\mathrm{KL}(\eta_{\mathbf{V},x} \mid \pi_{\mathbf{V},s^{\perp}}) \leq \delta$，并使用 $O\left(\kappa^{3/2} m^{1/2} \log^3(1/\delta)\right)$ 次顺序预言机调用即可实现。这一结果例如可以通过采用Altschuler 和 Chewi (2024，定理 D.1) 的方法得到。一阶假设不符合我们所考虑的零阶设定。然而，如果假设算法1 中的近似误差可以忽略，并且有 m 个并行处理器可用，则可以在单个并行轮次中重构梯度。基于这些考虑，我们引入假设 3。

假设 3。 对于每个 $\delta > 0$、$\mathbf{V} \in \mathbb{V}_m(\mathbb{R}^d)$ 和 $x \in \mathbb{R}^d$，存在一个算法，使用

$$N(m, \kappa, \delta) = O\left(\kappa^{3/2} m^{1/2} \log^3(1/\delta)\right),$$

并行轮次，每轮包含 m 次对 U 的并行评估，以从分布$\eta_{\mathbf{V},\, x} \in \mathcal{P}(\mathbb{R}^m)$ 中采样，使得 $\mathrm{KL}(\eta_{\mathbf{V},\, x} \mid \pi_{\mathbf{V},\, s^\perp}) \leq \delta$。

在假设 1-2-3下，推论1 表明，由算法4 所描述类别中的方法能够在 m内实现多项式加速。

推论 1。 在假设 *1-2-3*下，算法*4* 需要

$$n_* = O\Big(\frac{\kappa^{5/2} d}{m^{1/2}} \log^3\Big(\frac{2\kappa d}{\epsilon m}\Big) \log\Big(\frac{2\mathrm{KL}(\mu \mid \pi)}{\epsilon}\Big)\Big),$$

使用零阶预言机进行并行轮次，以生成一个在 ***Kullback-Leibler*** 散度意义下与 π 相差不超过 ϵ 的样本。

固定 ϵ、κ 和 $\mathrm{KL}(\mu \mid \pi)$ 后，推论 1 表明，可以从π 近似采样，在 $O(d/m^{1/2})$ 个并行轮次内达到误差 ϵ，忽略多对数项。相比之下，RWM 算法需要 $O(d)$ 次迭代才能达到相同精度（Andrieu 等，2024）。假设评估 π 构成主要计算瓶颈，则随机切片方法可带来约为 $m^{1/2}$ 的多项式加速。这一提升远高于标准步内并行 MCMC 方法所实现的多对数级增益（Brockwell，2006 年；Pozza 和 Zanella，2025）。此外，与谱隙为 $O(1/(d-m))$ 的算法 3 相比，算法 4 对 m 的依赖更有利，尤其是在 m 远小于 d 时。图 1 所示结果也从经验上验证了这一点。

备注 *1*（m的最优选择）。在算法 *4* 中，我们必须选择每次迭代所使用的随机方向数量m。对于 m_0 个并行处理器，自然的选择是至少取 $m = m_0$（根据推论 *1*，选择 $m < m_0$ 会增加达到相同精度所需的并行轮次数）。不过，也可以考虑 $m = c \times m_0$ 大于一的策略。在这种情况下，按照上述相同的推理，需要来自一阶预言机的$O\left(\kappa^{3/2}\ (cm_0)^{\ 1/2}\, \log^3\ (1/\delta)\ \right)$ 次顺序调用，才能得到 $KL\ (\eta_{\mathbf{V},\ x} \mid \pi_{\mathbf{V},\ s^\perp}) \leq \delta$。在零阶框架下，使用 m_0 个并行处理器，可以通过 c 个并行轮次重构单次一阶预言机调用所获得的相同信息（忽略算法 *1* 中的有限差分误差）。这会改变推论 *1* 给出的并行轮次总数；在这种情况下，其阶为

$$O\left(\frac{\kappa^{5/2} d c^{3/2} m_0^{1/2}}{c m_0} \log^3\left(\frac{2\kappa d}{\epsilon m}\right) \log\left(\frac{2\mathrm{KL}(\mu \mid \pi)}{\epsilon}\right)\right) = O\left(\frac{\kappa^{5/2} d c^{1/2}}{m_0^{1/2}} \log^3\left(\frac{2\kappa d}{\epsilon m}\right) \log\left(\frac{2\mathrm{KL}(\mu \mid \pi)}{\epsilon}\right)\right),$$

因此，对于 $c = 1$，其值达到最小。因此，从理论上讲，选择 $m = m_0$ 应当是最优的，这一点也得到了图 *4* 和图 *5*中所报告结果的实证验证。

5 数值计算

5.1 逻辑回归

我们首先考察算法5 在两个贝叶斯逻辑回归上的表现，其维度分别为 $d = 25$，200，样本量分别为 $n = 25$，200。遵循贝叶斯领域的标准

记号，我们用 $\beta \in \mathbb{R}^d$ 表示模型参数。假设先验分布为 $\beta \sim N_d(0,\ \sigma^2 I_d)$，其中$\sigma^2 = 25/d$ 。采用这一似然和先验后，势函数是m-凸的、 L-光滑的，且等于

$$U(\beta) = \sum_{i=1}^{n} \Big[\log\big(1 + e^{z_i^\top \beta}\big) - y_i\, z_i^\top \beta \Big] + \frac{1}{2\sigma^2}\, \beta^\top \beta\,,$$

其中， $y_i \in \{0,1\}$ 是响应变量， $z_i \in \mathbb{R}^d$ 是$i = 1,\ \ldots\ldots,\ n$的解释变量向量。在模拟中，真实参数 β_0 生成自 $\beta_0 \sim N_d(0,\ (1/8)\mathbb{I}_d)$，每个$z_i \overset{\text{i.i.d.}}{\sim} N_d(0,\ \mathbb{I}_d)$，并且给定$(\beta_0,\ z_i)$后， $y_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}\big(1/(1 + \exp(-z_i^\top \beta_0))\big)$。

考虑算法 5 的两个替代版本。在这两个版本中，每次迭代都会通过均匀且无放回地抽取 m 个标准基元素来生成矩阵 $\mathbf{V}$。在第一个版本中，蛙跳迭代次数 L 设为 1，从而得到 MALA 的随机切片变体（记为 RS-MALA）。在第二个版本中， L经过优化以提升性能。为与其他零阶步内并行化方法进行比较，我们考虑一种具有局部平衡接受概率的多重尝试马尔可夫链蒙特卡罗算法（Liu 等，2000年）（Gagnon 等，2023年）。如前文所述，Pozzaand Zanella（2025）证明，对于所考虑的这类模型，MTM 的加速 m 至多是对数多项式级别的。因此，值得观察的是，在实际设置中，第 4 节对随机切片方法所预测的多项式加速，是否能够相较于 MTM 的对数级加速带来显著收益。在这三种算法中，提议方差均采用自适应方式进行优化（Andrieuand Thoms，2008，算法 4）。对于每个 m 值，马尔可夫链运行 5×10^4 次迭代。

对于一般的马尔可夫链 $X_0, \ldots, X_T$，若其 $X_i = (X_{i,1}, \ldots, X_{i,d}) \in \mathbb{R}^d$，我们将（经验）期望平方跳跃距离（ESJD）定义为

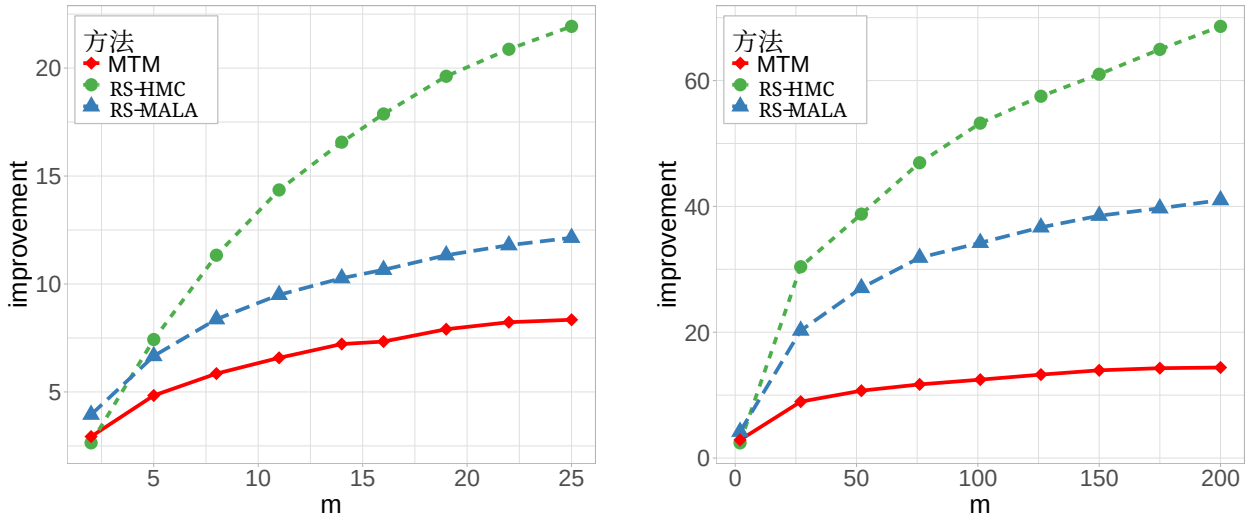
$$\text{ESJD} = \frac{1}{d(T-1)} \sum_{j=1}^{d} \sum_{i=2}^{T} (X_{i+1,j} - X_{i,j})^2. \tag{12}$$

图 2 展示了所考虑的两种模型中三种方法的 ESJD，并以随机游走 Metropolis–Hastings 马尔可夫链蒙特卡罗算法的 ESJD 进行归一化。对于 RS-HMC，ESJD 还除以 leapfrog 步数 L，因此对于 m的任意取值，三种方法的每次迭代计算成本大致相同。RS-MALA 和 RS-HMC 的表现均显著优于 MTM：即使并行处理器数量适中，相较于 RWM–MCMC 所获得的加速也大于 MTM 在 $m = d$时达到的加速，尤其是在 $d = 200$ 情形下。还要注意，当$m = d$时，除算法1中的离散化误差外，RS-MALA 和 RS-HMC 分别等价于全梯度 MALA 和 HMC 算法。鉴于此，值得考察这两种方法如何随着 m 增大而趋近其 $m = d$ 版本，并将其与 MTM 相应的表现进行比较。对于 MTM，改进很快趋于稳定：当 $d = 200$ 时，$m = 75$ 的结果与 $m = 200$时相近。相比之下，RS-MALA 和 RS-HMC 在 $m < 75$ 时提升显著，并且当 m 超过 75 后仍持续明显改善。RS-HMC 尤其如此。

5.2 随机波动率模型

现在，我们比较 RS-MALA、RS-HMC 和 MTM 在时间序列数据随机波动率模型上的表现（Kim 等人，1998）。该模型包含 $n + 3$ 潜在参数 μ， ϕ、对数 $\sigma \in \mathbb{R}$ 以及 $\eta \in \mathbb{R}^n$，它们用于对时间序列的观测波动率进行参数化 $\{y_i\}_{i=1}^n$，假设

$$y_i \sim N(0, \exp(h_i)), \qquad i = 1, \ldots, n,$$



a) d $= 25$ b) d $= 200$

图 2：相对于随机游走 Metropolis，RS-MALA、RS-HMC 和 MTM 在第5.1节所讨论的两个逻辑回归问题中，按估计的平均期望平方跳跃距离衡量的相对改进。结果已根据算法每次迭代中的 m 并行目标函数评估总数进行标准化。

其中 $h_1 = \mu + \exp(\log\sigma)/\{1-\tanh(\phi)^2\}\eta_1$ 以及

$$h_t = \tanh(\phi)(h_{t-1}-\mu) + \exp(\log\sigma)\eta_t, \qquad t = 2,\ldots,n.$$

对于模型参数的先验分布，我们取

$$\mu \sim N(0,10), \quad \phi \sim N(0,1), \quad \log\sigma \sim N(0,1), \quad \eta_i \overset{\text{iid}}{\sim} N(0,1) \quad i = 1,\ldots,n.$$

随后通过设置 $n = 200$、$\mu_0 = 1$、$\tanh(\phi_0) = 0.5$、$\log\sigma_0 = 0$ 并采样 η_0、$_i$ 独立同分布$\sim N(0,1)$来生成数据，由此得到一个包含 $d = 203$ 个参数的模型。对于每种算法以及所考虑的每个m 值，均获得长度为 5×10^4 的马尔可夫链。图 5.2 展示了 RS-MALA、RS-HMC 和 MTM 算法的性能随 m 变化的情况。与第 5.1 节相比，这三种方法相对于 RWM–MCMC 的改进幅度均略大。尽管如此，RS-MALA 和 RS-HMC 仍优于 MTM，得出了与第 5.1 节相同的定性结论。RS-MALA 和 RS-HMC 的性能均优于 MTM，且 $m = d$，所需的目标函数评估次数显著更少。此外，当 m 趋近于 d 时，它们的性能并未趋于稳定，反而在 m取较大值时仍显著提升。

5.3 m 的最优选择

算法 5 的实现需要指定每次迭代中所使用的随机方向数量 m 。对于 m_0 个并行处理器，备注 1 指出，至少从理论上讲，设置 $m = m_0$ 可使算法效率达到最大。本节通过将所得核的收敛速度与每次迭代成本之比作为算法效率的衡量指标，对此问题进行实证研究。我们使用期望平方跳跃距离（ESJD）来估计收敛速度，具体定义见

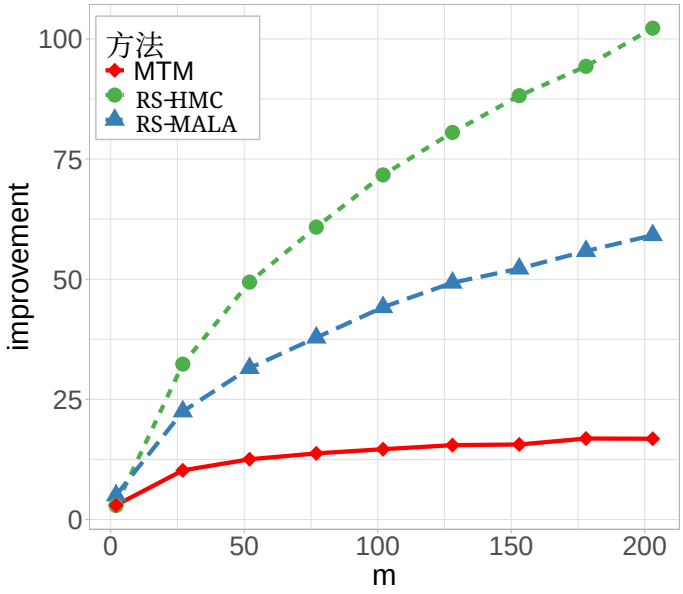


图 3：对于第 5.2 节所讨论的随机波动率模型，以估计的期望平方跳跃距离（ESJD）衡量 RS-MALA、RS-HMC 和 MTM 相对于随机游走 Metropolis 的相对改进。结果已针对算法每次迭代的总 m 次并行目标函数评估进行了标准化。

式（12），由此得到算法5 的效率度量，形式为

$$Eff(m) = \frac{ESJD_{P^{\text{RS-HMC}}_{m,L}}}{(m/m_0)L}, \tag{13}$$

其中 $ESJD_{P^{\text{RS-HMC}}_{m\,L}}$ 表示转移核为 $P^{\text{RS-HMC}}_{m\,L}$ 的马尔可夫链的期望平方跳跃距离（*ESJD*）。图4和图 5分别展示了在第 $Eff\,(m)\,/Eff\,(m_0)$ 中，对于不同 m_0 和 m取值时，200 维逻辑回归模型和第 5.1 节及第 5.2 节所讨论的随机波动率模型中的相对效率。在所考虑的所有情形下，当$m = m_0$时可获得最高效率。选择 $m > m_0$ 会导致算法性能迅速下降，尤其是当 m 显著小于 d时。这表明，简单的选择 $m = m_0$ 不仅具有理论依据，而且在实践中也很有效。

致谢 作者感谢 Martin Chak 提供有益的讨论和建议，尤其是关于定理 2。

参考文献

Charu C A ggarwal。 机器学习的线性代数与优化。施普林格，2020 .

Jason M Altschuler 和 Sinho Chewi。通过算法热启动实现更快的高精度对数凹采样。*ACM* 期刊，71(3):1–55，2024。

Nima Anari、Sinho Chewi 和 Thuy-Duong Vuong。等周不等式条件下的快速并行采样。载于第三十七届学习理论年会，第 161–185 页。PMLR，2024。

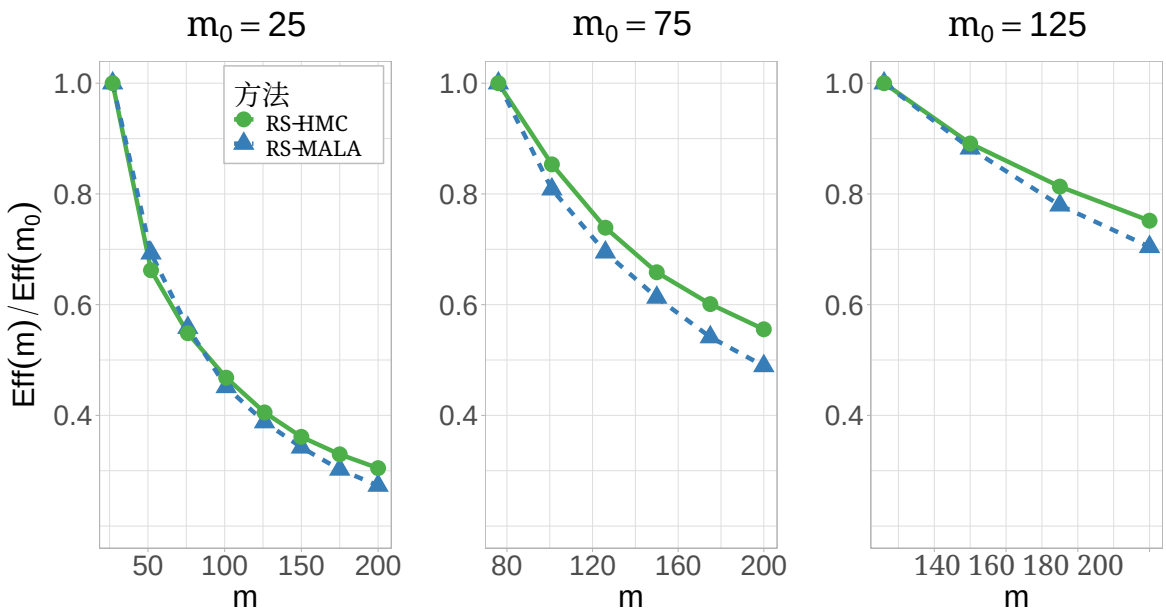


图 4：不同取值下的相对效率 m_0 在第 5.1节所述的 200 维逻辑回归中

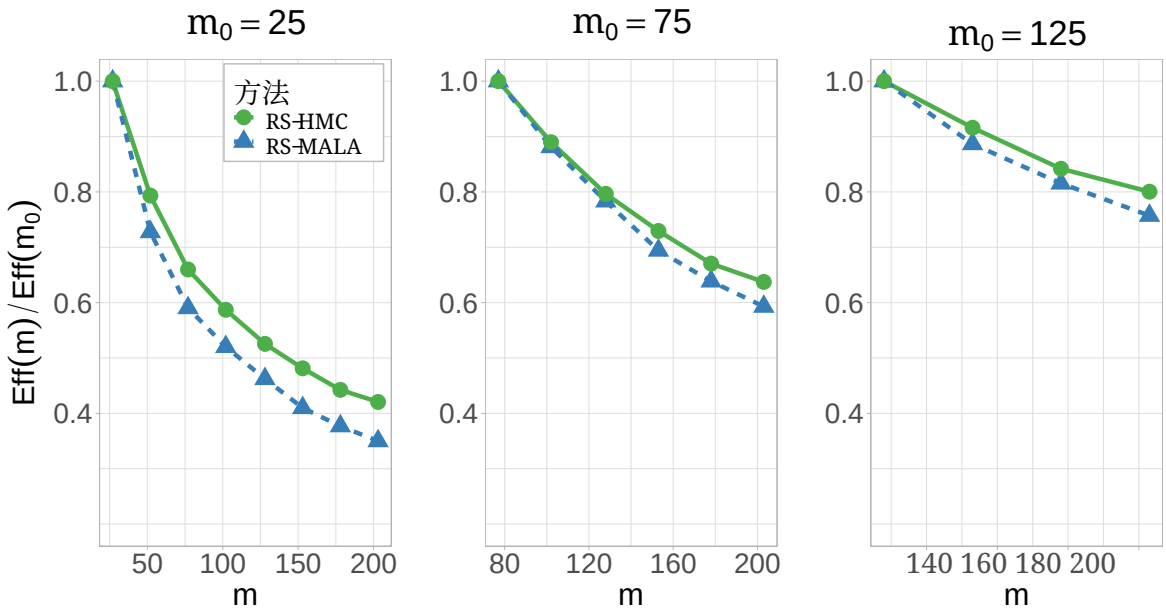


图 5：不同取值下的相对效率 m_0 在第5.1节所述的 203 维随机波动模型中

Christophe Andrieu 和 Johannes Thoms。自适应马尔可夫链蒙特卡罗教程。*Statistics and Computing*，18:343–373，2008年。Christophe Andrieu、Anthony Lee、Sam Power 和 Andi Q Wang。Metropolis 马尔可夫链的显式收敛界：等周不等式、谱隙与剖面。《应用概率年刊》，34(4):4022–4071，2024。Filippo Ascolani、Antonio Lijoi、Giovanni Rebaudo 和 Giacomo Zanella。狄利克雷过程混合模型的聚类一致性。*Biometrika*，110(2):551–558，2023。Filippo Ascolani、Hugo Lavenant 和 Giacomo Zanella。对数凹性下吉布斯采样器的熵收缩。*arXiv* 预印本 *arXiv:2410.00858*，2024。Francis Bach。从第一性原理出发的学习理论。麻省理工学院出版社，2024。Atilim Gunes Baydin、Barak A Pearlmutter、Alexey Andreyevich Radul 和 Jeffrey Mark Siskind。机器学习中的自动微分：综述。《机器学习研究杂志》，18(153):1–43，2018。Myl`ene B´edard、Randal Douc 和 Eric Moulines。多重尝试马尔可夫链蒙特卡罗方法的尺度分析。《随机过程及其应用》，122(3):758–786，2012。Anthony E Brockwell。通过预取实现并行马尔可夫链蒙特卡罗模拟。《计算与图形统计学杂志》，15(1):246–261，2006 年。Ben Calderhead。并行化 Metropolis-Hastings 算法的一般构造。*Proceed-ings of the National Academy of Sciences*，111(49):17408–17413，2014年。Rocco Caprio、Sam Power 和 Andi Wang。通过 poincar\'e 不等式分析多重尝试 Metropolis。*arXiv* 预印本 *arXiv:2504.18409*，2025。Roberto Casarin、Radu Craiu 和 Fabrizio Leisen。具有不同提议分布的交互式多重尝试算法。*Statistics and Computing*，23:185–200，2013。George Casella 和 Edward I George。解释吉布斯采样器。《美国统计学家》，46(3):167–174，1992。Ming-Hui Chen 和 Bruce W Schmeiser。用于评估多维积分的一般击中并运行蒙特卡罗抽样。《运筹学快报》，19(4):161–169，1996年。Sinho Chewi。对数凹抽样。可在 *https://chewisinho. github. io* 获取的书稿，2024。Radu V Craiu 和 Christiane Lemieux。使用反变量抽样和分层抽样加速多重尝试 Metropolis 算法。*Statistics and Computing*，17:109–120，2007。John C Duchi、Michael I Jordan、Martin J Wainwright 和 Andre Wibisono。零阶凸优化的最优速率：两次函数评估的力量。《*IEEE*信息论汇刊》，61(5):2788–2806，2015年。Daniele Durante、Francesco Pozza 和 Botond Szabo。偏斜伯恩斯坦–冯·米泽斯定理与偏斜模态近似。《统计年刊》，52(6):2714–2737，2024。´Alain Durmus 和 Eric Moulines。通过未校正朗之万算法进行高维贝叶斯推断。*Bernoulli*，25(4A):2854–2882，2019。

Simon Fontaine 和 Myl`ene B´edard。自适应多重尝试 Metropolis 算法。*Bernoulli*，28 (3):1986–2011，2022。Philippe Gagnon、Florian Maire 和 Giacomo Zanella。通过局部平衡改进多重尝试 Metropolis。《机器学习研究杂志》，24(248):1–59，2023。Alan E Gelfand。吉布斯抽样。美国统计学会会刊，95(452):1300– 1304，2000年。Mark Girolami 和 Ben Calderhead。黎曼流形朗之万方法与哈密顿蒙特卡罗方法。英国皇家统计学会杂志 *B* 辑：统计方法学，73(2):123– 214，2011。Sebastiano Grazzi 和 Giacomo Zanella。利用 Picard 映射进行 Metropolis 马尔可夫链的并行计算。*arXiv* 预印本 *arXiv:*，2025。Sangjoon Kim、Neil Shephard 和 Siddhartha Chib。随机波动率：似然推断及其与 ARCH 模型的比较。经济研究评论，65(3):361–393，1998。David Kozak、Stephen Becker、Alireza Doostan 和 Luis Tenorio。高维无梯度优化的随机子空间方法。计算优化与应用，79(2):339–368，2021。David Kozak、Cesare Molinari、Lorenzo Rosasco、Luis Tenorio 和 Silvia Villa。利用正交随机方向进行零阶优化。数学规划，199(1):1179–1219，2023。David A Levin 和 Yuval Peres。马尔可夫链与混合时间，第107卷。美国数学学会，2017年。Jun S Liu、Faming Liang 和 Wing Hung Wong。Metropolis 抽样中的多重尝试方法与局部优化。美国统计学会会刊，95(449):121–134，2000年。Samuel Livingstone 和 Giacomo Zanella。Barker 提议：结合稳健性与基于梯度的马尔可夫链蒙特卡罗的效率。英国皇家统计学会杂志 *B* 辑：统计方法学，84(2):496–523，2022。Charles C Margossian。自动微分及其高效实现综述。*Wiley* 跨学科评论：数据挖掘与知识发现，9(4):e1305，2019。Luca Martino 和 Francisco Louzada。多重尝试 Metropolis 混合中的问题。计算统计学，32:239–252，2017。Francesco Mezzadri。如何从经典紧群生成随机矩阵。*arXiv* 预印本 *math-ph/0609050*，2006 年。Radford M Neal 等人。使用哈密顿动力学进行马尔可夫链蒙特卡罗。马尔可夫链蒙特卡洛手册，2(11):2，2011。Yurii Nesterov 和 Vladimir Spokoiny。凸函数的随机无梯度最小化。计算数学基础，17(2):527–566，2017。Yurii Nesterov 等人。凸优化讲义，第137卷。施普林格，2018。

Francesco Pozza 和 Giacomo Zanella. 关于多提议马尔可夫链蒙特卡罗算法的根本局限。*Biometrika*，112(2):asaf019，2025。Gareth O Roberts 和 Osnat Stramer. 朗之万扩散与Metropolis-Hastings算法。*Methodology and computing in applied probability*，4:337–357，2002。Gareth O Roberts 和 Richard L Tweedie. 朗之万分布及其离散近似的指数收敛。*Bernoulli*，2:341–363，1996。Jeffrey S Rosenthal. 并行计算与蒙特卡洛算法。*Far East Journal of Theo-retical Statistics*，4(2):207–236，2000。Abhishek Roy、Lingqing Shen、Krishnakumar Balasubramanian 和 Saeed Ghadimi. 用于贝叶斯推断的朗之万扩散随机零阶离散化。*Bernoulli*，28(3):1810–1834，2022。Adrien Saumard 和 Jon A Wellner. 对数凹性与强对数凹性：综述。*Statisticssurveys*，8:45，2014。Luke Tierney. 关于一般状态空间的Metropolis-Hastings核的说明。*Annals of appliedprobability*，第1–9页，1998。Hakon Tjelmeland. 使用所有 Metropolis-Hastings 提议估计均值。技术报告，数学科学系，挪威科技大学，2004。Roman Vershynin. 高维概率：数据科学应用导论，第47卷。剑桥大学出版社，2018。Martin J Wainwright. 高维统计：非渐近视角，第48卷。剑桥大学出版社，2019。Xiaodong Yang 和 Jun S Liu. 多重尝试Metropolis独立采样器的收敛速率。*Statistics and Computing*，33(4):79，2023。

A 证明

A.1 命题的证明1

证明。 令 $\gamma > 0$和 $x,\ y \in \mathbb{R}^d$。我们使用 $X \sim \delta_x P^{SULA}$和 $Y \sim \delta_y P^{SULA}$的同步耦合，定义如下

$$X = x - \gamma \frac{d}{m} \mathbf{V}\mathbf{V}^\top \nabla U(x) + \sqrt{2\gamma} Z,$$
$$Y = y - \gamma \frac{d}{m} \mathbf{V}\mathbf{V}^\top \nabla U(y) + \sqrt{2\gamma} Z.$$

满足 $Z \sim N_d(0,\ \mathbb{I}_d)$、$\mathbf{V} \sim \nu$，且 Z 与 $\mathbf{V}$ 彼此独立。上述耦合给出上界

$$\begin{aligned} W_2^2(\delta_x P^{SULA}, \delta_y P^{SULA}) &\leq \mathbb{E}[\|Y - X\|^2] \\ &= \int \|y - x - \gamma \frac{d}{m} \mathbf{V}\mathbf{V}^\top \left(\nabla U(x) - \nabla U(y)\right)\|^2 \nu(d\mathbf{V}) \,. \end{aligned}$$

由假设 2 和引理8 可得

$$\begin{aligned} &\int \|y - x - \gamma \frac{d}{m} \mathbf{V}\mathbf{V}^\top \left(\nabla U(x) - \nabla U(y)\right)\|^2 \nu(d\mathbf{V}) \\ &= \frac{(d-m)}{d} \|y - x\|^2 + \frac{m}{d} \|y - x - \gamma \frac{d}{m} (\nabla U(x) - \nabla U(y))\|^2. \end{aligned} \tag{14}$$

By A假设 1 以及 Nesterov等人 (2018年，定理 2.1.15）可得，对于 $\gamma \leq m/(dL)$

$$\|y - x - \gamma \frac{d}{m} (\nabla U(x) - \nabla U(y))\|^2 \leq (1 - \gamma \frac{d}{m} \lambda) \|y - x\|^2 \,. \tag{15}$$

(14) 和 (15) 结合可得

$$\begin{aligned} W_2^2(\delta_x P^{SULA}, \delta_y P^{SULA}) \leq & \frac{(d-m)}{d} \|y - x\|^2 + \frac{m}{d} \left(1 - \gamma \frac{d}{m} \lambda\right) \|y - x\|^2 \\ = & \left(1 - \gamma\lambda\right) \|y - x\|^2, \end{aligned}$$

正如所期望的。 □

A.2 命题的证明2

证明。 令 $A, B \subset \mathbb{R}^d$满足 $A \cap B = \emptyset$。直接应用 Fubini定理可得

$$\begin{aligned} \int \mathbb{1}_B(x) P_\sigma^{SMALA}(x, A) \pi(dx) &= \int_{\mathbb{V}(\mathbb{R}^d)} \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d} \mathbb{1}_B(x) \mathbb{1}_A(y) \min\left(\pi(dx) Q_{\mathbf{V}}(x, dy), \pi(dy) Q_{\mathbf{V}}(y, dx)\right) \Big) \nu(d\mathbf{V}) \\ &= \int \mathbb{1}_A(y) P_\sigma^{SMALA}(y, B) \pi(dy), \end{aligned}$$

这就证明了该命题。 □

A.3 引理 1 的证明

证明。 我们从 s的边缘分布出发。根据 $Q_{\mathbf{V}}(x,\ dy)$ 如算法 3所示，由于 y 具有高斯性，因此 s服从均值为

$$\mathbb{E}(s) = \mathbf{V}^\top \mathbb{E}(y) = \mathbf{V}^\top x - \frac{\sigma^2}{2}\mathbf{V}^\top \hat{\nabla}_{\mathbf{V}} U(x),$$

以及协方差矩阵 $\mathbb{V}(s) = \mathbf{V}^\top \mathbb{V}(y)\mathbf{V} = \sigma^2 \mathbf{V}^\top \mathbf{V}$。这证明了式（5）。类似地，由于 $\mathbf{V}^\perp$ 和 $\mathbf{V}$的列正交， $s^\perp$ 是一个均值为 $d-m$ 的高斯分布，维数为

$$\begin{aligned}
\mathbb{E}(s^\perp) =& (\mathbf{V}^\perp)^\top \mathbb{E}(y) \\
=& (\mathbf{V}^\perp)^\top x - \frac{\sigma^2}{2}(\mathbf{V}^\perp)^\top \hat{\nabla}_{\mathbf{V}} U(x) \\
=& (\mathbf{V}^\perp)^\top x - \frac{\sigma^2 c_\nu}{2}(\mathbf{V}^\perp)^\top \mathbf{V}\mathbf{V}^\top \nabla U(x) \\
=& (\mathbf{V}^\perp)^\top x,
\end{aligned}$$

以及协方差矩阵 $\mathbb{V}(s^\perp) = (\mathbf{V}^\perp)^\top \mathbb{V}(y)(\mathbf{V}^\perp) = \sigma^2 \mathbb{I}_{d-m}$，由此证明了式（6）。最后，注意到

$$\begin{aligned}
\mathbb{E}\Big((s - \mathbb{E}(s))(s^\perp - \mathbb{E}(s^\perp))^\top\Big) =& \mathbf{V}^\top (\mathbb{E}(yy^\top) - \mathbb{E}(y)\mathbb{E}(y^\top))\mathbf{V}^\perp \\
=& \sigma^2 \mathbf{V}^\top \mathbf{V}^\perp = \mathbf{0}_{m\times(d-m)},
\end{aligned}$$

最后一个等号同样是由$\mathbf{V}^\perp$ 和$\mathbf{V}$ 的列正交这一事实推出的，其中 $\mathbf{0}_{m\times(d-m)}$ 表示零矩阵 $m\times$（$d-m$）。至此，引理得证。 □

A.4 定理 1 的证明

证明。回顾一下，对于 $x \sim \pi$， $s^\perp = (\mathrm{V}^\perp)^\top x$ 和 $s = \mathrm{V}^\top x$。此外，$\pi_{\mathbf{V}^\perp}(ds^\perp)$ 表示 $s^\perp$ 的边缘分布，而 $\pi_{\mathbf{V},\ s^\perp}(ds)$ 表示在 $s^\perp$ 给定条件下 s 的条件分布，其密度为 $\pi_{\mathbf{V},\ s^\perp}(s) \propto \exp(-U(\mathrm{V}^\perp s^\perp + \mathrm{V}s))$。利用引理 1， P_σ^{SMALA} 可改写为

$$\begin{aligned}
P_\sigma^{SMALA}(x, A) = \int_{\mathbb{R}^{d-m}\times\mathbb{R}^m} & \mathbb{1}_A(\mathbf{V}^\perp r^\perp + \mathbf{V}r) Q_{\mathbf{V}^\perp}(s^\perp, dr^\perp) Q_{\mathbf{V}}((s, s^\perp), dr) \\
& \times \min\left(1, \frac{\pi_{\mathbf{V}^\perp}(dr^\perp)\pi_{\mathbf{V},r^\perp}(dr) Q_{\mathbf{V}}((r, r^\perp), ds)}{\pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds) Q_{\mathbf{V}}((s, s^\perp), dr)}\right) \nu(d\mathbf{V}),
\end{aligned} \tag{16}$$

对于每个 $A \subseteq \mathbb{R}^d \backslash \{x\}$，满足 $Q_{\mathbf{V}^\perp}(s^\perp, dr^\perp) = \phi_{d-m}(dr^\perp; s^\perp, \sigma^2 \mathbb{I}_{d-m})$，并且

$$Q_{\mathbf{V}}((s, s^\perp), dr) = \phi_m\big(dr; s + (\sigma^2/2)\mathbf{V}^\top \hat{\nabla}_{\mathbf{V}} U(\mathbf{V}s + \mathbf{V}^\perp s^\perp), \sigma^2 \mathbb{I}_m\big),$$

其中， $Q_{\mathbf{V}}$ 的协方差为 $\sigma^2 \mathbb{I}_m$ 因为根据假设， $\mathbf{V}$ 的每一列都是标准基 $\mathbf{E}$中的一个元素，因此 $\sigma^2 \mathbf{V}^\top \mathbf{V} = \sigma^2 \mathbb{I}_m$。回顾 $Gap_{P_\sigma^{SMALA}} \leq 2\Phi_{P_\sigma^{SMALA}}$(Andrieu 等人， 2024， 引理 5)，其中

$$\Phi_{P_\sigma^{SMALA}} = \inf\left\{\frac{\int_A P_\sigma^{SMALA}(x, A^c)\pi(dx)}{\pi(A)},\ A \subset \mathbb{R}^d,\ 0 < \pi(A) \leq 1/2\right\}, \tag{17}$$

是 P_σ^{SMALA}的导通率。令 $\mathcal{A}_{1/4}=\{A\subset\mathbb{R}^d\colon\ 1/4\le\pi(A)<1/2\}$。根据（17），$P_\sigma^{SMALA}$属于（33）所述的类别，由引理 2可得

$$\begin{aligned}\frac{1}{2}Gap_{P_\sigma^{SMALA}} &\le \inf_{A\in\mathcal{A}_{1/4}}\frac{\int_A P_\sigma^{SMALA}(x,A^c)\pi(dx)}{\pi(A)}\\ &\le \inf_{A\in\mathcal{A}_{1/4}} 4\int\int \mathbb{1}_{A^\perp}(s^\perp)\pi_{\mathbf{V}^\perp}(ds^\perp)P^{RW}_{\mathbf{V}^\perp}(s^\perp,\mathbb{R}^{d-m}\backslash\{s^\perp\})\nu(d\mathbf{V}),\end{aligned}\tag{18}$$

其中 $A^\perp=\{s^\perp=(\mathbf{V}^\perp)^\top x\colon\ x\in A\}$ 和

$$P^{RW}_{\mathbf{V}^\perp}(s^\perp,\mathbb{R}^{d-m}\backslash\{s^\perp\})=\int \mathbb{1}_{\{\mathbb{R}^{d-m}\backslash\{s^\perp\}\}}(r^\perp)Q_{\mathbf{V}^\perp}(s^\perp,dr^\perp)\min\left(1,\frac{\pi_{\mathbf{V}^\perp}(dr^\perp)}{\pi_{\mathbf{V}^\perp}(ds^\perp)}\right),$$

是一个以 $\pi_{\mathbf{V}^\perp}$为目标分布的高斯随机游走 Metropolis-Hastings 核。另请注意，根据假设1 和引理5，$\pi_{\mathbf{V}^\perp}$ 的势函数 m-凹、L-光滑，且对每个$\mathbf{V}^\perp$二阶连续可微。因此，可以将引理3 应用于（18），得到

$$Gap_{P_\sigma^{SMALA}}\le 24\exp\left(\frac{-\lambda^3\sigma^2(d-m)}{128L^2}\right)+32\exp\left(-\frac{(d-m)}{32}\right).\tag{19}$$

现在令 $g_a(x)=a^\top(x-\mathbb{E}_\pi(x))$，其中 $a\in\mathbb{R}^d$。由（7）中谱隙的定义可得

$$\begin{aligned}Gap_{P_\sigma^{SMALA}} &\le \inf_{a\in\mathbb{R}^d,\,\|a\|=1}\frac{\int(g_a(x)-g_a(y))^2\pi(dx)P_\sigma^{SMALA}(x,dy)}{2\mathbb{V}_\pi(g_a)}\\ &= \inf_{a\in\mathbb{R}^d,\,\|a\|=1}\frac{\int\{a^\top(x-y)\}^2\pi(dx)P_\sigma^{SMALA}(x,dy)}{2\mathbb{V}_\pi(g_a)}.\end{aligned}$$

为给前一公式右侧的量提供上界，我们选择固定 a为$a=(1/\sqrt{d})(\mathbf{1},\ldots,\mathbf{1})$。利用（16）以及$\mathbf{V}\mathbf{V}^\top+\mathbf{V}^\perp(\mathbf{V}^\perp)^\top=\mathbb{I}_d$，前一公式右侧的分子可以表示为

$$\begin{aligned}&\int\{a^\top(x-y)\}^2\pi(dx)P_\sigma^{SMALA}(x,dy)\\ &\le\int\left\{a^\top\mathbf{V}(s-r)\right\}^2\pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V})\\ &+\int\left\{a^\top\mathbf{V}^\perp(s^\perp-r^\perp)\right\}^2\pi_{\mathbf{V}^\perp}(ds^\perp)Q_{\mathbf{V}^\perp}(s^\perp,dr^\perp)\nu(d\mathbf{V})\\ &+2\int\left[a^\top\mathbf{V}(s-r)\right]\left[a^\top\mathbf{V}^\perp(s^\perp-r^\perp)\right]\pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}^\perp}(s^\perp,dr^\perp)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}).\end{aligned}\tag{20}$$

我们分别处理上一公式右侧的三个项。根据 $Q_{\mathbf{V}^\perp}$ 的定义，给定 $s^\perp$和 $\mathbf{V}^\perp$，$a^\top\mathbf{V}^\perp(s^\perp-r^\perp)\sim N(0,\ \sigma^2a^\top\mathbf{V}^\perp(\mathbf{V}^\perp)^\top a)$ 满足$a^\top\mathbf{V}^\perp(\mathbf{V}^\perp)^\top a\le\|a\|^2=1$。因此，式（20）右侧的第二项上界为

$$\int\left\{a^\top\mathbf{V}^\perp(s^\perp-r^\perp)\right\}^2\pi_{\mathbf{V}^\perp}(s^\perp)Q_{\mathbf{V}^\perp}(s^\perp,dr^\perp)\nu(d\mathbf{V})\le\sigma^2.\tag{21}$$

其次，根据引理 1，给定 s，$s^\perp$ 和 $\mathbf{V}$，随机变量 $\left[a^\top\mathbf{V}(s-r)\right]$ 和 $\left[a^\top\mathbf{V}^\perp(s^\perp-r^\perp)\right]$相互独立，这意味着等式（20）右侧的第三项为零。

等式（20）右侧的第一项还可以进一步分解为

$$
\begin{aligned}
&\int \{a^\top \mathbf{V}(s-r)\}^2 \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}) \\
=&\int \{a^\top \mathbf{V}\big(s-\frac{\sigma^2}{2}\frac{d}{m}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp)-r\big)\}^2 \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}) \\
+&\int \{\frac{\sigma^2}{2}\frac{d}{m}a^\top \mathbf{V}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp)\}^2 \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}) \\
+2&\int \{a^\top \mathbf{V}(\frac{\sigma^2}{2}\frac{d}{m}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp))\}\{a^\top \mathbf{V}\big(s-\frac{\sigma^2}{2}\frac{d}{m}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp)-r\big)\} \\
&\quad \times \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}). \qquad (22)
\end{aligned}
$$

在给定（s， $s^\perp$）和 $\mathbf{V}$的条件下，引理 1 可得

$$
a^\top \mathbf{V}\Big(s-\frac{\sigma^2}{2}\frac{d}{m}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp)-r\Big) \sim N(0,\sigma^2 a^\top \mathbf{V}\mathbf{V}^\top a). \qquad (23)
$$

此外，根据假设 2 以及 $\|a\|=1$

$$
\int (a^\top \mathbf{V}\mathbf{V}^\top a)\nu(d\mathbf{V}) = \frac{m}{d}\|a\|^2 = \frac{m}{d}. \qquad (24)
$$

由（23）和（24）可得

$$
\begin{aligned}
&\int \{a^\top \mathbf{V}\big(s-\frac{\sigma^2}{2}\frac{d}{m}\mathbf{V}^\top U(\mathbf{V}s+\mathbf{V}^\perp s^\perp)-r\big)\}^2 \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}) \\
=&\sigma^2\int (a^\top \mathbf{V}\mathbf{V}^\top a)\nu(d\mathbf{V}) \le \sigma^2\frac{m}{d}. \qquad (25)
\end{aligned}
$$

同理，

$$
\begin{aligned}
&\int \{a^\top \mathbf{V}(\frac{\sigma^2}{2}\frac{d}{m}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp))\}\{a^\top \mathbf{V}\big(s-\frac{\sigma^2}{2}\frac{d}{m}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp)-r\big)\} \\
&\quad \times \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}) = 0. \qquad (26)
\end{aligned}
$$

最后，注意到根据假设，$\mathbf{V}$ 的每一列都属于 $\mathbb{R}^d$ 的标准基，结合 $x=\mathbf{V}s+\mathbf{V}^\perp s^\perp$， $a=$ $(1/\sqrt{d})$（1，..，1），由柯西–施瓦茨不等式和假设 2 可得

$$
\begin{aligned}
&\int \{\frac{\sigma^2}{2}\frac{d}{m}a^\top \mathbf{V}\mathbf{V}^\top \nabla U(\mathbf{V}s+\mathbf{V}^\perp s^\perp)\}^2 \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}) \\
\le&\frac{\sigma^4}{4}\frac{d^2}{m^2}\int (a^\top \mathbf{V}\mathbf{V}^\top a)(\nabla U(x)^\top \mathbf{V}\mathbf{V}^\top \nabla U(x))\nu(\mathbf{V})\pi(dx) \\
=&\frac{\sigma^4}{4}\int \|\nabla U(x)\|^2\pi(dx) \le \sigma^4\frac{L^2}{\lambda}\frac{d}{4}, \qquad (27)
\end{aligned}
$$

最后一个不等式由以下事实得出：根据假设 1 和引理 7， $\|\nabla U(x)\|^2 \le L^2\|x-x_*\|^2$ 且 $\|x-x_*\|^2 \overset{s}{\preccurlyeq} \lambda^{-1}\chi_d^2$ 其中 χ_d^2 是自由度为 d 的卡方随机变量， $x_*\in\mathbb{R}^d$ 是 $U(x)$ 的极小化点，且 $\overset{s}{\preccurlyeq}$ 表示随机序。因此，结合（25）-（26）-（27），我们得到

$$
\int \{a^\top \mathbf{V}(s-r)\}^2 \pi_{\mathbf{V}^\perp}(ds^\perp)\pi_{\mathbf{V},s^\perp}(ds)Q_{\mathbf{V}}((s,s^\perp),dr)\nu(d\mathbf{V}) \le \sigma^2\frac{m}{d}+\sigma^4\frac{L^2}{\lambda}\frac{d}{4}. \qquad (28)
$$

最后，根据假设 $a^\top X) \geq \|a\|^2/L = 1/L$，方差 1 以及 Saumard 和 Wellner（Saumard andWellner(2014年）中的公式 10.25。将其与（19）、（21）和（28）结合可得

$$Gap_{P_\sigma^{SMALA}} \leq \min\Big(f_1(\sigma), f_2(\sigma)\Big), \tag{29}$$

其中

$$\begin{aligned} f_1(\sigma) &= 24\exp\Big(\frac{-\lambda^3\sigma^2(d-m)}{128L^2}\Big) + 32\exp\Big(-\frac{(d-m)}{32}\Big), \\ f_2(\sigma) &= \sigma^2\frac{L}{2}\frac{m}{d} + \frac{\sigma^4}{8}\frac{dL^3}{\lambda}. \end{aligned}$$

注意，$f_1(\sigma)$ 关于 σ 为 $f_2(\sigma)$ 关于 σ。此外，$\lim_{\sigma\to 0^+} f_1(\sigma)/f_2(\sigma) = +\infty$ 且 $\lim_{\sigma\to 0} f_1(\sigma)/f_2(\sigma) = 0$。由此以及式（29）可得

$$\sup_{\sigma>0} Gap_{P_\sigma^{SMALA}} \leq \sup_{\sigma>0} \min\Big(f_1(\sigma), f_2(\sigma)\Big) \leq \max\Big(f_1(\sigma), f_2(\sigma)\Big), \tag{30}$$

对每个 $\sigma > 0$。为证明式（8），令 $\sigma_*^2 = 128L^2\log(d-m)/\{\lambda^3(d-m)\}$。由 $m < c_0 d$可知 $m/(d-m) < c_1$（当 $c_1 = c_0/(1-c_0)$ 时）。结合式（30）得到

$$\begin{aligned} \sup_\sigma Gap_{P_\sigma^{SMALA}} \leq& \max\Big(f_1(\sigma^*), f_2(\sigma^*)\Big) \\ =& \max\Big(\frac{24}{d-m} + 32\exp\Big(-\frac{(d-m)}{32}\Big), \frac{64c_0L^3\log(d-m)}{(d-m)} + \frac{2048L^7\log(d-m)^2}{\lambda^7(d-m)^2}d\Big) \\ =& O\Big(\frac{\log(d-m)^2}{(d-m)}\Big), \end{aligned}$$

由此完成了定理的证明。 □

A.5 定理 2 的证明

证明。给定任意概率测度 $\mu \in \mathcal{P}(\mathbb{R}^d)$和 $\mathrm{V} \in \mathbb{V}_m(\mathbb{R}^d)$，我们用 $\mu_{\mathbf{V}^\perp}$表示 $s^\perp = \mathrm{V}^\perp x$ 关于 $x \sim \mu$ 的边缘分布，并用 $\mu_{\mathbf{V},s^\perp}$ 表示给定 $s^\perp$时$s = \mathrm{V}^\top x$ 的条件分布。我们首先考虑情形 $t = 1$。利用式（9）中给出的 $P(x,\cdot)$ 的结构，以及 Kullback-Leibler 散度的凸性和链式法则性质，再结合 $(\mu P_{\mathbf{V}})_{\mathbf{V}^\perp} = \mu_{\mathbf{V}^\perp}$ 和 $(\mu P)_{\mathbf{V},s^\perp} = \eta_{\mathbf{V},x}$（其中 $\eta_{\mathbf{V},x}$ 按式（10）定义），可得

$$\begin{aligned} \mathrm{KL}(\mu P \mid \pi) \leq& \int \mathrm{KL}(\mu P_{\mathbf{V}} \mid \pi)\nu(d\mathbf{V}) \\ =& \int \mathrm{KL}(\mu_{\mathbf{V}^\perp} \mid \pi_{\mathbf{V}^\perp})\nu(d\mathbf{V}) + \int \mathbb{E}_{s^\perp\sim\mu_{\mathbf{V}^\perp}}\left[\mathrm{KL}(\eta_{\mathbf{V},x} \mid \pi_{\mathbf{V},s^\perp})\right]\nu(d\mathbf{V}). \end{aligned}$$

在假设 1-2下，根据 Ascolani 等人 (2024) 的推论 8.2 证明中给出的结果可知

$$\int \mathrm{KL}(\mu_{\mathbf{V},s^\perp} \mid \pi_{\mathbf{V},s^\perp})\nu(d\mathbf{V}) \leq \Big(1 - \frac{1}{\kappa}\frac{m}{d}\Big)\mathrm{KL}(\mu \mid \pi).$$

由（10）可得，$\mathrm{KL}(\eta_{\mathbf{V},\,x} \mid \pi_{\mathbf{V},\,s^\perp}) < \delta$，这意味着

$$\mathrm{KL}(\mu P \mid \pi) \leq \Big(1 - \frac{1}{\kappa}\frac{m}{d}\Big)\mathrm{KL}(\mu \mid \pi) + \delta.$$

对 $t>1$ 重复同样的推理（将 μ 替换为 μP^{t-1}）可得

$$\begin{aligned}\mathrm{KL}(\mu P^t \mid \pi) \leq & \left(1-\frac{1}{\kappa}\frac{m}{d}\right)^t \mathrm{KL}(\mu \mid \pi)+\delta \sum_{i=1}^{t-1}\left(1-\frac{1}{\kappa}\frac{m}{d}\right)^i \\ \leq & \left(1-\frac{1}{\kappa}\frac{m}{d}\right)^t \mathrm{KL}(\mu \mid \pi)+\frac{\delta \kappa d}{m}, \end{aligned} \tag{31}$$

最后一个不等式由 $1-m/(\kappa d)<1$ 得出。由此证明了 (11)。

最后，注意到由 $(1-x) \leq \exp(-x)$。因此，(11) 右侧的第一项上界为 $\exp(-tm/(\kappa d))$。于是，

$$\mathrm{KL}(\mu P^t \mid \pi) \leq \frac{\epsilon}{2}, \quad \text{if} \quad t \geq (\kappa d)/m \log(2\mathrm{KL}(\mu \mid \pi)/\epsilon), \tag{32}$$

对任意 $\epsilon>0$。结合 (31) 和 (32)，并取 $\delta \leq \epsilon m/(2\kappa d)$，即可证明该定理。 □

A.6 推论 1 的证明

证明。 结论直接由定理 2 和假设 (3) 得出。 □

A.7 命题 3 的证明

证明。为了证明该结论，我们说明：给定 $\mathbf{V} \in \mathbb{V}_m(\mathbb{R}^d)$ 和 $x \in \mathbb{R}^d$，算法 5 的每次迭代都是 $\pi_{\mathbf{V},s^\perp}$-不变的。为此，首先注意，若我们将 T_L 定义为满足 $(s_L,k_L')=T_L(s_0,k)$ 的映射（其中 (s_0,k) 和 (s_L,k_L') 如算法 5 所述），则 $T_L^{-1}=T_L$。此外，T_L 是由函数 $F_1(s,k)=(s+k,k)$ 和 $F_{2,\gamma_*}(s,k)=(s,k+\gamma_* \hat{\nabla}_{\mathbf{V}} U(x+\mathbf{V}s))$ 的多次复合得到的，$\gamma_*>0$。F_1 和 F_{2,γ_*}的雅可比矩阵的行列式满足 $|J_{F_1}(s,k)|=1$ 以及 $|J_{F_{2,\gamma_*}}(s,k)|=1$，对每个 $(s,k) \in \mathbb{R}^{2m}$ 均成立；这意味着（鉴于雅可比矩阵的链式法则）$|J_{T_L}(s_0,k)|=1$ 对每个 $(s_0,k) \in \mathbb{R}^{2m}$均成立。因此，算法 5 的转移规则属于 Tierney（1998）讨论的确定性提议族。于是，选择

$$\min\left(1, \exp\left(U(x+\mathbf{V}s_0)-U(x+\mathbf{V}s_L)+0.5(\|k\|^2-\|k_L'\|^2)\right)\right),$$

由于接受概率使转移核相对于联合分布$\pi_{\mathbf{V},\,s^\perp} \otimes N_m(\mathbf{0},\,\mathbb{I}_d)$可逆，因此也是$\pi_{\mathbf{V},\,s^\perp}$-可逆的。命题得证。 □

B 技术结果

B.1 可分解核的效率

设 $(X_1,\,X_2)=X \sim \pi$, $\pi \in \mathcal{P}(\mathbb{R}^d)$。我们用 $\pi_1 \in \mathcal{P}(\mathbb{R}^{d-p})$ 表示X_1 的边缘分布，用 $\pi_{2,\,x_1}$ 表示给定 $X_1=x_1$ 时 X_2 的条件分布。接下来，我们考虑如下形式的核

$$\begin{aligned} P(x,A)=&\int_{\mathbb{R}^{m\times(d-m)}} \mathbb{1}_A((y_1,y_2))Q_1(x_1,dy_1)Q_{2,y_1}(x,dy_2) \\ &\times \min\left(1,\frac{\pi_1(dy_1)\pi_{2,y_1}(dy_2)Q_1(y_1,dx_1)Q_{2,x_1}(y,dx_2)}{\pi_1(dx_1)\pi_{2,x_1}(dx_2)Q_1(x_1,dy_1)Q_{2,y_1}(x,dy_2)}\right), \end{aligned} \tag{33}$$

其中，$Q_1:\ \mathbb{R}^{d-m} \to \mathcal{P}\left(\mathbb{R}^{d-m}\right)$ 和 $Q_{2,\,y_1}:\ \mathbb{R}^m \times \mathbb{R}^{d-m} \to \mathcal{P}\left(\mathbb{R}^m\right)$ 相对于勒贝格测度绝对连续。注意，Q_1 仅取决于 x_1 的前一个值，而$Q_{2,\,y_1}$ 则可以取决于 x_1、x_2 和 y_1。

引理 2. 设 $\pi \in \mathcal{P}(\mathbb{R}^d)$ 相对于勒贝格测度绝对连续，且 P 如 (33)所示。那么，对每个 $x_1 \in \mathbb{R}^{d-m}$,

$$\int_{\mathbb{R}^m} \pi_{2,x_1}(dx_2) P((x_1,x_2), \mathbb{R}^d \backslash \{x\}) \leq P_1(x_1, \mathbb{R}^{d-m} \backslash \{x_1\}), \tag{34}$$

其中

$$P_1(x_1, \mathbb{R}^{d-m} \backslash \{x_1\}) = \int \mathbb{1}_{\{\mathbb{R}^{d-m} \backslash \{x_1\}\}}(x_1) Q_1(x_1, dy_1) \min\left(1, \frac{\pi_1(dy_1) Q_1(y_1, dx_1)}{\pi_1(dx_1) Q_1(x_1, dy_1)}\right).$$

证明。 根据 P 在（33）中的定义，以及 π 相对于勒贝格测度绝对连续，结合富比尼定理，并对 $\min(1, t)$ 应用詹森不等式，可得

$$\begin{aligned}
\int_{\mathbb{R}^m} \pi_{2,x_1}(dx_2) P(x, \mathbb{R}^d \backslash \{x\}) &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^d \backslash \{x\}} \pi_{2,x_1}(dx_2) Q_1(x_1, dy_1) Q_{2,y_1}(x, dy_2) \\
&\quad \times \min\left(1, \frac{\pi_1(dy_1)\pi_{2,y_1}(dy_2) Q_1(y_1, dx_1) Q_{2,x_1}(y, dx_2)}{\pi_1(dx_1)\pi_{2,x_1}(dx_2) Q_1(x_1, dy_1) Q_{2,y_1}(x, dy_2)}\right) \\
&\leq \int_{\mathbb{R}^{d-m} \backslash \{x_1\}} Q_1(x_1, dy_1) \min\left(1, \frac{\pi_1(dy_1) Q_1(y_1, dx_1)}{\pi_1(dx_1) Q_1(x_1, dy_1)}\right. \\
&\quad \left. \times \int_{\mathbb{R}^m \times \mathbb{R}^m} \pi_{2,y_1}(dy_2) Q_{2,x_1}(y, dx_2)\right) \\
&\leq \int_{\mathbb{R}^{d-m} \backslash \{x_1\}} Q_1(x_1, dy_1) \min\left(1, \frac{\pi_1(dy_1) Q_1(y_1, dx_1)}{\pi_1(dx_1) Q_1(x_1, dy_1)}\right),
\end{aligned}$$

最后一个不等式由 $\int_{\mathbb{R}^m \times \mathbb{R}^m} \pi_{2,\ y_1}(dy_2) Q_{2,\ x_1}(y,\ dx_2) = 1$ 得出。证毕。

□

B.2 补充技术结果

引理 3 (随机游走梅特罗波利斯算法的平均接受率)。 设 π 满足假设 *1* ，且 P^{RW} 是一个随机游走 *Metropolis–Hastings* 转移核

$$P^{\mathrm{RW}}(x, A) = \int_{\mathbb{R}^d} \mathbb{1}_A(x) \phi_d(y; x, \sigma^2 \mathbb{I}_d) \min\Big(1, \exp(U(x) - U(y))\Big) dy, \tag{35}$$

对每个 $A \subset \mathbb{R}^d$ 均成立。则

$$P^{\mathrm{RW}}(x, \mathbb{R}^d \backslash \{x\}) \leq \exp\left(-\frac{\lambda \sigma^2 d}{4}\right) + 2\left\{\exp\left(-\frac{(\lambda \sigma d)^2}{64 \|\nabla U(x)\|^2}\right) + \exp\left(-\frac{d}{32}\right)\right\}, \tag{36}$$

and

$$\int_{\mathbb{R}^d} \pi(dx) P^{\mathrm{RW}}(x, \mathbb{R}^d \backslash \{x\}) \leq 3 \exp\left(\frac{-\lambda^3 \sigma^2 d}{128 L^2}\right) + 4 \exp\Big(-\frac{d}{32}\Big). \tag{37}$$

证明。 我们先将 y 在（35）中分解为 $y = x + \sigma z$ 其中 $z \sim N_d(0,\ \mathbb{I}_d)$ 。由 U 是m凸可知

$$\begin{aligned}
U(x) - U(x + \sigma z) &\leq -\sigma \nabla U(x)^\top z - \frac{\lambda \sigma^2}{2} \|z\|^2 \\
&= -\sigma \nabla U(x)^\top z - \frac{\lambda \sigma^2 d}{2} - \frac{\lambda \sigma^2}{2} \sum_{i=1}^{d} (|z_i|^2 - 1).
\end{aligned}$$

结合并集界可得

$$\begin{aligned}\Pr\left(\exp\big(U(x)-U(x+\sigma z)\big)\geq\exp\Big(-\frac{\lambda\sigma^2 d}{4}\Big)\right)&\leq\Pr\left(\sigma\nabla U(x)^\top z+\frac{\lambda\sigma^2 d}{2}+\frac{\lambda\sigma^2}{2}\sum_{i=1}^d(|z_i|^2-1)<\frac{\lambda\sigma^2 d}{4}\right)\\&=\Pr\left(\sigma\nabla U(x)^\top z+\frac{\lambda\sigma^2}{2}\sum_{i=1}^d(|z_i|^2-1)<-\frac{\lambda\sigma^2 d}{4}\right)\\&\leq\Pr\left(\nabla U(x)^\top z<-\frac{\lambda\sigma d}{8}\right)+\Pr\left(\frac{1}{d}\sum_{i=1}^d(|z_i|^2-1)<-\frac{1}{4}\right)\\&\leq 2\left\{\exp\left(-\frac{(\sigma\lambda d)^2}{64\|\nabla U(x)\|^2}\right)+\exp\left(-\frac{d}{32}\right)\right\},\end{aligned}\tag{38}$$

最后一行由伯恩斯坦不等式推出。利用（38），并注意到对于任意$t>0$，$\mathbb{E}(\min(1,\ x)\ \leq t+\Pr($ $x>t)$，我们得到

$$P^{\mathrm{RW}}\big(x,\mathbb{R}^d\backslash\{x\}\big)\leq\exp\Big(-\frac{\lambda\sigma^2 d}{4}\Big)+2\Big\{\exp\Big(\frac{-(\sigma\lambda d)^2}{64\|\nabla U(x)\|^2}\Big)+\exp\Big(-\frac{d}{32}\Big)\Big\},$$

这证明了（36）。关于（37），令 x_* 为 U（x）的唯一极小化点，并回忆L-光滑势函数 $\|\nabla U$（x）$\|^2\leq L^2\|x-x_*\|^2$。此外，由假设1和引理7可知，若 $X\sim\exp(-U)$，则 $\|X-x_*\|^2\overset{s}{\preccurlyeq}\lambda^{-1}\chi^2_d$。结合这些观察结果与（38），并利用 $\exp(-1/k)$ 关于 k 在 $k>0$时单调递增这一事实，以及引理4，我们得到

$$\begin{aligned}\int_{\mathbb{R}^d}\pi(dx)P^{\mathrm{RW}}(x,\mathbb{R}^d\backslash\{x\})&\leq\exp\Big(-\frac{\lambda\sigma^2 d}{4}\Big)+2\exp\Big(-\frac{d}{32}\Big)+2\int\pi(dx)\exp\Big(\frac{-(\sigma\lambda d)^2}{64\|\nabla U(x)\|^2}\Big)\\&\leq\exp\Big(-\frac{\lambda\sigma^2 d}{4}\Big)+2\exp\Big(-\frac{d}{32}\Big)+2\int\pi(dx)\exp\Big(\frac{-\lambda^3(\sigma d)^2}{64L^2\|x-x_*\|^2}\Big)\\&\leq\exp\Big(-\frac{\lambda\sigma^2 d}{4}\Big)+2\exp\Big(-\frac{d}{32}\Big)+2\exp\Big(\frac{-\lambda^3\sigma^2 d}{128L^2}\Big)+2\exp\Big(-\frac{d}{8}\Big)\\&\leq 3\exp\Big(\frac{-\lambda^3\sigma^2 d}{128L^2}\Big)+4\exp\Big(-\frac{d}{32}\Big).\end{aligned}$$

证明完毕。 □

引理4。令 $W\sim\chi^2_d$ 以及 $c>0$

$$\mathbb{E}_W\exp\Big(-\frac{c}{W}\Big)\leq\exp\Big(-\frac{c}{2d}\Big)+\exp\Big(-\frac{d}{8}\Big),$$

证明。 由于 $\exp(-c/W)$ 为正且小于 1，我们将期望表示为生存函数的形式：

$$\begin{aligned}\mathbb{E}_W\exp\Big(-\frac{c}{W}\Big)&=\int_0^1\Pr\Big(W>-\frac{c}{\log u}\Big)du\\&=\int_0^1\Pr\Big(\frac{1}{d}\sum_{i=1}^n Z_i^2-1>-\frac{c}{d\log u}-1\Big)du,\end{aligned}$$

其中 Z_i，$i=1$，……，n，是独立同分布标准正态随机变量。注意，$-c/(d$对数 $u)-1>1$若 $u>u^*=\exp(-c/(2d)$。因此，根据 Wainwright (2019年) 中的式 (2.18)，可得

$$\mathbb{E}_W \exp\left(-\frac{c}{W}\right) \le \exp\left(-\frac{c}{2d}\right) + \int_{u^*}^1 \Pr\left(\frac{1}{d}\sum_{i=1}^n Z_i^2 - 1 > -\frac{c}{d\log u} - 1\right) \le \exp\left(-\frac{c}{2d}\right) + \exp\left(-\frac{d}{8}\right),$$

至此，引理的证明完成。 □

引理 5. 设 $x\sim\pi$ 满足 π，并满足假设 1。设 $V\in\mathbb{V}_m(\mathbb{R}^d)$，并考虑线性变换 $g_{\mathbf{V}}$：$\mathbb{R}^d\to\mathbb{R}^m\times\mathbb{R}^{d-m}$，使得 $g_{\mathbf{V}}(x)=(s^\perp,s)$，其中 $s^\perp=(V^\perp)^\top x$，$V^\perp$ 的定义如引理 1 所示，且 $s=V^\top x$。那么，$s^\perp$，$\pi_{\mathbf{V}}$ 的边缘分布满足相同常数 λ 和 L下的假设 1。

证明。 L-光滑性和 λ-凸性在正交线性变换下保持不变。因此，只需对 x 的任意划分 (x_1,x_2) 证明该引理，其中 $x_1\in\mathbb{R}^{d-m}$ 且 $x_2\in\mathbb{R}^m$。记 π_1 为 x_1 的边缘分布，并记 $U_1(x_1)=-\log\left(\int_{\mathbb{X}_2}\pi(x_1,dx_2)\right)$ 为 π_1的势函数。在假设 1 下，Saumard 和 Wellner（2014年，定理 3.8）已证明 $U_1(x)$ 的 λ-凸性。为证明 L-光滑性，令

$$\pi(dx_2\mid x_1)=\pi(x_1,dx_2)/\int_{\mathbb{X}_2}\pi(x_1,dx_2).$$

应用莱布尼茨积分法则并进行简单的代数运算可得，$U_1(x)$的一阶和二阶导数等于

$$\frac{\partial}{\partial x_1}U_1(x_1)=\int_{\mathbb{X}_2}\left(\frac{\partial}{\partial x_1}U(x_1,x_2)\right)\pi(dx_2\mid x_1),$$

and

$$\frac{\partial^2}{\partial x_1\partial x_1^\top}U_1(x_1)=\int_{\mathbb{X}_2}\left(\frac{\partial^2}{\partial x_1\partial x_1^\top}U(x_1,x_2)\right)\pi(dx_2\mid x_1)-\mathrm{Cov}\left(\frac{\partial}{\partial x_1}U(x_1,x_2)\right), \tag{39}$$

其中，对于给定的 $x_1\in\mathbb{R}^m$，(39) 右侧最后一项表示在条件分布 $\pi(dx_2\mid x_1)$ 下，$(\partial/\partial x_1)U(x_1,x_2)$ 的协方差矩阵。根据假设 1 以及补充材料中 Durante 等（2024）的引理 C.3，可得 $(\partial^2/(\partial x_1\partial x_1^\top))U(x_1,x_2)\preccurlyeq L\mathbb{I}_m$，而 $\mathrm{Cov}\big((\partial/\partial x_1)U(x_1,x_2)\big)$ 根据构造为半正定。这意味着

$$\frac{\partial^2}{\partial x_1\partial x_1^\top}U_1(x_1)\preccurlyeq L\mathbb{I}_m,$$

由此完成该引理的证明。 □

引理 6。 *(由 l2 范数诱导的球面分布)* 令 $p(x)$为一个支撑集为 $\mathbb{R}^d$ 的概率密度函数，并对每个 $x\in\mathbb{R}^d$ 定义映射

$$(x_1,\dots x_d)\to(y_1,\dots,y_{d-1},r),$$

其中 $y_i=x_i/\|x\|_2$，$i=1,\dots,d-1$ 和 $r=\mathrm{sign}(x_d)\|x\|_2$，其中 $\|x\|_2=(\sum_{i=1}^d|x_i|^2)^{1/2}$。

那么，向量 (y,r) 的概率密度函数形式为

$$p(y,r)=p(x(y,r))|r|^{d-1}(1-\|y\|_2^2)^{-1/2},$$

其中

$$x(y,r)=(|r|y_1,|r|y_2,\dots,r(1-\|y\|_2^2)^{1/2}). \tag{40}$$

证明。结论可直接由概率密度函数的变量替换公式得出。

□

引理 7. 设 U 是以 $x_* \in \mathbb{R}^d$为最小值的 λ-凸函数。则 $\|X - x^*\|^2 \overset{s}{\preccurlyeq} \lambda^{-1}\chi_d^2$，其中 $\overset{s}{\preccurlyeq}$表示随机序。

证明。无需失去一般性，假设 $x^* = 0$。令 $X \sim \pi$的密度为 $\pi(x) \propto \exp(-U(x))$。令 $Z \sim N_d(0, \lambda^{-1}\mathbb{I}_d)$。为证明该结论，考虑随机变量 $R_X = \text{sign}(X_d)\|X\|_2$、$R_Z = \text{sign}(Z_d)\|Z\|_2$，并注意到 $|R_Z| \sim \sqrt{\lambda^{-1}\chi_d^2}$。令 $Y_{X,i} = X_i/\|X\|_2$, $i = 1, \ldots, d-1$。根据引理 6，在给定 $Y_X = (Y_X, 1, \ldots, Y_{X,d})$ 的条件下，R_X的条件概率密度函数为 $p_{R_X|Y_X=y}(r) \propto \exp(-U(x(y,r)) + (d-1)\log(|r|))$，其中 $x(y,r)$ 在式 (40) 中定义。类似地，R_Z的边际概率密度函数满足 $p_{R_Z}(r) \propto \exp(-\lambda r^2/2 + (d-1)\log(|r|))$。$_s$通过证明 $R_X|\{R_X \geq 0\} \preccurlyeq R_Z|\{R_Z \geq 0\}$且$_s R_X|\{R_X < 0\} \succcurlyeq R_Z|\{R_Z < 0\}$，即可证明该引理的结论。为此，我们先研究比值的变化

$$\frac{p_{R_X}(r)\mathbb{1}_{r>0}}{p_{R_Z}(r)} = \int \frac{p_{R_X|Y_X=y}(r)}{p_{R_Z}(r)} p_{Y_X}(dy) = \int \exp\left\{\frac{\lambda r^2}{2} - U(x(r,y))\right\} p_{Y_X}(dy).$$

对上一个公式中所报告的量求导后，可应用莱布尼茨积分法则将其重新整理如下

$$\frac{\partial}{\partial r}\int \exp\left\{\frac{\lambda r^2}{2} - U(x(r,y))\right\} p_{Y_X}(dy) = \int \exp\left\{\frac{\lambda r^2}{2} - U(x(r,y))\right\}\left(\lambda r - \frac{\partial}{\partial r}U(x(r,y)\right) p_{Y_X}(dy).$$

此外，令 $\tilde{y} = (y_1, \ldots\ldots, y_{d-1}, \sqrt{1 - \|y\|_2^2})$，并注意到

$$\frac{\partial}{\partial r}U(x(r,y)) = \left\{\frac{\partial}{\partial x}U(x(r,y))\right\}^\top \tilde{y},$$

由于 $U(x)$ 是 λ-凸的，

$$\frac{\partial^2}{\partial^2 r}U(x(r,y)) = \tilde{y}^\top\left\{\frac{\partial^2}{\partial x \partial x^\top}U(x(r,y))\right\}\tilde{y} \geq \lambda. \tag{41}$$

因此，利用（41）并且对于 $r \geq 0$,

$$\lambda r - \frac{\partial}{\partial r}U(x(r,y)) = \lambda r - \int_0^r \tilde{y}^\top\left\{\frac{\partial^2}{\partial x \partial x^\top}U(x(r^*,y))\right\}\tilde{y}dr^* \leq \lambda r - \lambda\int_0^r dr^* = 0.$$

这意味着比值 $p_{R_X}(r)/p_{R_Z}(r)$ 随 r递减（即，$p_{R_Z}(r)/p_{R_X}(r)$ 递增）。$_s$因此，根据 Ascolani 等人 （2023年）的引理 5，$R_X|\{R_X \geq 0\} \preccurlyeq \{R_Z|R_Z \geq 0\}$。沿用$_s$相同的步骤，可以证明 $R_X|\{R_X < 0\} \succcurlyeq R_Z|\{R_Z < 0\}$。这两个结果结合起来便证明了该引理的陈述。

□

引理 8. 设 $a, b \in \mathbb{R}^d$ 以及 $\mathbf{V} \sim \nu$ 满足 ν 假设 2。则

$$\mathbb{E}\left(\|a + \mathbf{V}\mathbf{V}^\top b\|^2\right) = \frac{d-m}{d}\|a\|^2 + \frac{m}{d}\|a+b\|^2.$$

证明。回顾一下，对于任意$\mathbf{V} \in \mathbb{V}_m(\mathbb{R}^d)$斯蒂费尔德，都有 $\mathbb{I}_d = \mathbf{V}\mathbf{V}^\top + \mathbf{V}^\perp(\mathbf{V}^\perp)^\top$。这一点以及$\mathbb{E}(\mathbf{V}\mathbf{V}^\top) = (m/d)\mathbb{I}_d$这一事实，意味着 $\mathbb{E}(\mathbf{V}^\perp(\mathbf{V}^\perp)^\top) = \{(d-m)/d\}\mathbb{I}_d$。因此，利用恒等式$a = \mathbf{V}\mathbf{V}^\top a + \mathbf{V}^\perp(\mathbf{V}^\perp)^\top a$，通过简单的代数变换可得

$$\begin{aligned}\mathbb{E}\big(\|a + \mathbf{V}\mathbf{V}^\top b\|^2\big) =& a^\top \left\{\mathbb{E}(\mathbf{V}\mathbf{V}^\top) + \mathbb{E}(\mathbf{V}^\perp(\mathbf{V}^\perp)^\top)\right\} a + 2a^\top \mathbb{E}(\mathbf{V}\mathbf{V}^\top)b + b^\top \mathbb{E}(\mathbf{V}\mathbf{V}^\top)b \\ =& \frac{(d-m)}{d}\|a\|^2 + \frac{m}{d}\|a+b\|^2.\end{aligned}$$

引理的证明至此结束。 □