arXiv:2301.01898v1 [math.NA] 5 Jan 2023

快速副本交换随机梯度朗之万动力学

Guanxun Li*、Guang Lin †、Zecheng Zhang ‡、QuanZhou §

摘要

将副本交换（即并行回火）技术应用于朗之万蒙特卡洛算法，尤其是随机梯度朗之万动力学（SGLD），在非凸学习问题中取得了巨大成功，但一个潜在的局限是运行多条链所带来的计算成本。我们观察到，SGLD 中梯度估计器的高方差实质上会提高平稳分布的温度。基于此，我们提出通过直接估计随机梯度估计器所造成的偏差，加速 SGLD 的回火方案。这一简单思路使我们能够以可忽略不计的计算成本（相较于低温链的计算成本）模拟高温链，同时保持向目标分布收敛。我们的方法与近期提出的 m-reSGLD（多方差副本交换随机梯度朗之万动力学）方法有本质区别：后者会因梯度估计器准确性较低而受到影响（例如，链可能无法收敛到目标分布），而我们的方法则从中受益。此外，我们还推导出一种易于计算的交换率，相较于 m-reSGLD 实现了另一项重大改进。为从理论上展示我们方法的优势，我们给出了 Wasserstein距离下的收敛界。我们还提供了高斯混合和逆偏微分方程模型的数值示例，结果表明，我们的方法比传统的多方差副本交换方法收敛更快。

1 引言

给定概率分布 $\pi(\theta) \propto \exp(-U(\theta))$，假设能量函数 U 已知，可以使用马尔可夫链蒙特卡罗（MCMC）方法从π中生成样本。当 π 是对数凹的（即 U 是凸的）时，已知大多数实际应用中的 MCMC 算法（例如 Metropolis-Hastings 和哈密顿蒙特卡洛算法）都能快速混合（Dwivedi等，2018；Mangoubi和Smith，2021）。然而，现实中我们经常面临更具挑战性的问题，此时 π 往往具有严重的多峰性。克服多峰性最广泛使用的技术之一是并行回火，也称为副本交换蒙特卡洛（Machta，2009）。其主要思路是，不针对 π，而是设计以 $\pi_\tau(\theta) \propto \exp(-U(\theta)/\tau)$为目标的 MCMC 算法，其中参数 $\tau > 0$ 通常称为温度。当 τ 足够大时，π_τ 的形状会变得平坦，使链能够轻松地在局部模态之间移动。最简单形式的副本交换会并行运行两个 MCMC 算法，其中一个以某个较小的 π_{τ_1} 为目标

*德州农工大学统计系，美国德克萨斯州大学城，邮编 77843。（电子邮箱：guanxun@stat.tamu.edu）

†普渡大学数学系，美国印第安纳州西拉斐特，邮编 47907。（电子邮箱：guan- glin@purdue.edu）

‡卡内基梅隆大学数学系，美国宾夕法尼亚州匹兹堡，邮编 15213。（电子邮箱：zecheng.zhang.math@gmail.com）§德州农工大学统计系，美国德克萨斯州大学城，邮编 77843。（电子邮箱：quan@stat.tamu.edu）

τ_1 （即“低温”链），另一个以某个较大的 π_{τ_2} 为目标 τ_2 （即“高温”链）。低温链可以快速找到附近的局部模态，而高温链则持续探索整个状态空间。通过允许两条链交换状态，低温链便能够高效地在局部模态之间跳转。要恢复分布 π，只需设置 $\tau_1 = 1$ 并收集低温链的样本，或者使用重要性温度调整方法（Gramacy 等人，2010年）。与单链马尔可夫链蒙特卡罗方法相比，这种方案的一个潜在局限是，每次迭代的计算成本都会翻倍。

1.1 背景与动机

本文考虑使用朗之万蒙特卡洛方法模拟每条链。在对 U 作出一些正则性假设后，我们可以构造一个平稳分布为 π_τ 的朗之万扩散（LD）（Roberts 和 Stramer，2002年；Nguyen 等人，2019年）；参见式（1）。对于大型数据集，精确计算 U 或 U 的梯度（模拟 LD 时需要用到）可能非常耗时。采样和优化领域中一种广泛采用的方法是，利用随机批次样本估计 U 和 ∇U（Welling 和 Teh，2011年；Dalalyan，2017年；Simsekli 等人，2020年）；这种方法称为随机梯度朗之万动力学（SGLD）。我们注意到，也可以采用其他方法进行梯度估计；例如，对于反问题偏微分方程（Efendiev 等人，2006年；Stuart，2010年；Chung 等人，2020年），U 通过正向偏微分方程求解器计算。为了节省计算时间，可以使用仅提供 U 近似值的快速求解器。对于非凸优化和多模态采样问题，传统的朗之万蒙特卡洛算法可能会陷入局部模态。为此，人们提出了副本交换朗之万扩散（reLD）和副本交换随机梯度朗之万动力学（reSGLD）算法，将副本交换与朗之万动力学结合起来（Chen 等人，2019年；Deng 等人，2020年）。这两种方法并行运行两条链，理论研究表明，与单链方法相比，reLD 和 reSGLD 收敛所需的迭代次数更少（Zhang 等人，2017年；Raginsky 等人，2017年；Deng 等人，2020年；Chen 等人，2019年）。

然而，由于 reSGLD 使用两条链，其每次迭代的计算成本翻了一倍，总计算成本也可能高于单链 SGLD（当然，也可以将 reSGLD 推广为并行运行两条以上的链，此时总计算成本可能会高得多）。为解决这一问题，提出了多方差 reSGLD（Lin等人，2022年，2021年；Na 等人，2022年）（m-reSGLD），旨在降低高温链的计算成本。与 reSGLD 不同，m-reSGLD 对两条链采用不同的梯度估计方案，且高温链所用的估计量在计算上比低温链所用的估计量更高效（但均方误差也更大）。由于高温链用于探索，低温链则用于利用，人们预期高温链或许能够容忍更大程度的梯度估计误差。m-reSGLD 在许多具有挑战性的任务中取得的经验成功（Lin 等人，2022年，2021年）支持了这一直觉。尽管如此，m-reSGLD 相较于 reSGLD 的优势尚未得到理论论证。更重要的是，目前尚不清楚高温链中的梯度估计误差会如何影响算法的整体性能，也不清楚该误差允许达到多大，才能确保高温链仍依据温度调整后的目标分布探索整个空间。

1.2 本文的主要贡献

我们提出了一种简单但极其有效的方法，并为其提供了理论保证，能够以可忽略不计的计算成本模拟高温随机梯度朗之万动力学链。其核心思想是学习

SGLD链的“有效温度”，方法是估计梯度估计器的方差。令 $v(\theta)$表示梯度估计器 $\nabla U(\theta)$ 的协方差矩阵，该矩阵可能取决于状态 θ。尽管从估计的角度来看，较大的 $v(\theta)$并不理想，但对于副本交换方案中的高温链而言，它能带来促进探索所需的随机性。事实上，在数值模拟朗之万扩散的离散时间近似时，我们需要在每个时间步注入随机噪声。将 $\nabla U(\theta)$的噪声与注入噪声合并后可以看出， $\nabla U(\theta)$的高变异性实质上提高了链的温度。换个角度来看，这也意味着，为了达到固定温度，我们除了注入较大的随机噪声，也可以通过使用“更粗糙”的梯度估计器来增大 $v(\theta)$。推导的详细过程见第3节。我们注意到，对于单链SGLD方法，梯度估计器的偏差校正已得到充分研究（Teh等人，2016年；Vollmer等人，2016年）。我们方法的新颖之处在于，我们有意利用这种偏差（甚至增大偏差），以低得多的成本运行高温随机梯度朗之万动力学链。我们将该方法称为f-reSGLD（快速副本交换SGLD）。

对于实践中遇到的大多数目标分布， $v(\theta)$ 是未知的，但通常不难构造一个“良好”的估计量 $\hat{v}(\theta)$（见公式 (14)）。在这种情况下，我们的理论分析表明，该方法带来的计算增益可能非常显著。此外，我们推导了 f-reSGLD 在 2-瓦瑟斯坦距离下收敛至目标分布的收敛速率。与 Deng 等人（2020年，定理 1）的结果相比，作为主导项的梯度估计误差有所降低。尽管副本交换主要用于求解具有多个局部最优点的非凸问题，但它也可能为凸问题提供高效的解决方案，尤其是在缺乏良好初始化的情况下。当能量函数 $U(\theta)$ 为强凸时，我们推导了一个收敛结果。与非凸情形下的收敛估计相比，学习率趋于零时，梯度估计带来的误差项将消失。

我们将主要贡献总结如下。

1. 我们提出通过将梯度估计器的方差视为随机性的主要来源来模拟高温随机梯度朗之万动力学链，这可以显著降低计算成本。2. 我们为 f-reSGLD 算法提出了两条链之间的交换率，其计算效率高于 m-reSGLD 中采用的规则（Lin等人，2022年）。3. 我们证明了凸和非凸目标分布在 2-瓦瑟斯坦距离下的收敛界，说明了 f-reSGLD 算法的优势。

本文其余部分安排如下。第 2 节回顾 SGLD、reSGLD 和 m-reSGLD 算法。第 3 节正式介绍我们的方法，第 4 节给出理论收敛结果。第 5 节展示数值实验。最后，我们在第 6 节讨论并总结本文。所有证明均置于附录中。

2 预备知识

2.1 随机梯度朗之万动力学

假设我们希望从某个分布中进行采样

$$\pi_\tau(\theta) \propto \exp(-U(\theta)/\tau),$$

其中 $\theta \in \mathbb{R}^p$, $\tau > 0$ 是温度，能量函数 U 已知。从 $\pi_\tau(\theta)$ 生成样本的一种方法，是模拟一个平稳分布为 π_τ 的随机过程。在满足某些正则性条件时（例如参见 Bhattacharya (1978)；Roberts and Tweedie (1996)），以下随机微分方程（SDE），即所谓的朗之万扩散（LD），具有平稳分布 π_τ:

$$d\theta_t = -\nabla U(\theta_t)\, dt + \sqrt{2\tau} dW_t, \tag{1}$$

其中，W_t 是 p维标准布朗运动。为了在实践中进行采样，我们将 (1) 离散化为

$$\theta_{k+1,\eta} = \theta_{k,\eta} - \eta_k \nabla U(\theta_{k,\eta}) + \sqrt{2\eta_k \tau}\xi_k, \tag{2}$$

其中，$\xi_k \sim \mathcal{N}(0, I_p)$，$\eta_k$是第 k步所使用的步长。我们将 (2) 称为“精确更新”。在许多问题中，$U(\theta)$还可以进一步写成 $U(\theta) = \sum_{i=1}^N U_i(\theta)$，其中 $U_1, U_2, \ldots, U_N$是一些函数，并且

$$\nabla U(\theta) = \sum_{i=1}^N \nabla U_i(\theta). \tag{3}$$

例 2.1. 考虑包含 N 个独立同分布观测值的贝叶斯推断。令 $\mathcal{D} = \{X_i\}_{i=1}^N$ 为我们观测到的数据，$f(\cdot|\theta)$为给定参数θ时观测值的概率密度函数。令 $\Pi(\theta)$为我们设定的 θ的先验分布。后验分布记为 $\Pi(\theta|\mathcal{D})$，可通过以下方式计算：

$$\Pi(\theta|\mathcal{D}) \propto \Pi(\theta) \prod_{i=1}^N f(X_i|\theta).$$

令 $\pi(\theta) = \Pi(\theta|\mathcal{D})$，我们可以将能量函数表示为

$$U(\theta) = -\log \Pi(\theta) - \sum_{i=1}^N \log f(X_i|\theta) = \sum_{i=1}^N U_i(\theta),$$

其中 $U_i(\theta) = -\frac{1}{N}\log\ \Pi(\theta)\ - \log\ f(X_i|\theta)$。

当样本量 N 较大时，由于需要计算 ∇U，实现 (2) 可能会耗费较长时间。此时可以使用随机梯度朗之万扩散（随机梯度朗之万动力学，SGLD）来加快 LD 模拟。我们不再通过 (3) 精确计算 ∇U，而是从 $\mathcal{D}$ 中抽取 n 个观测值，并估计 $\nabla U(\theta)$，得到

$$\widehat{\nabla U}(\theta) \coloneqq \frac{N}{n} \sum_{i \in \mathcal{I}_n} \nabla U_i(\theta), \tag{4}$$

其中，$\mathcal{I}_n$ 是从 $\{1, \ldots, N\}$ 中不放回抽取的一个随机样本，样本量为 n。以 π_τ 为目标的 SGLD 更新式为

$$\widehat{\theta}_{k+1,\eta} = \widehat{\theta}_{k,\eta} - \eta_k \widehat{\nabla U}(\widehat{\theta}_{k,\eta}) + \sqrt{2\eta_k \tau}\xi_k.$$

在随机梯度朗之万动力学中，步长 η_k 的选择使得 $\eta_k \to 0$ 当 $k \to \infty$。我们注意到，如果目标是从 π_τ中生成样本，也可以通过选择恒定步长并使用 Metropolis-Hastings规则来校正离散化偏差，从而采用随机梯度朗之万动力学；这种方法称为 Metropolis调整的朗之万算法（Besag，1994；Roberts and Rosenthal，1998）。特别地，只要有 $e^{-U/\tau}$的无偏估计量，就可以使用伪边际MCMC技术来计算接受率（Andrieu and Roberts，2009）。

2.2 副本交换随机梯度朗之万动力学

副本交换朗之万扩散（reLD）是一种算法，旨在当目标分布为非凸或多峰时加速随机微分方程的收敛（Chen 等人，2019）。令τ_1，τ_2 表示两个温度，且满足 $\tau_2 > \tau_1$，定义两个并行的朗之万扩散（LD）如下:

$$\begin{aligned} d\theta_t^{(1)} &= -\nabla U(\theta_t^{(1)})\,dt + \sqrt{2\tau_1}dW_t^{(1)}, \\ d\theta_t^{(2)} &= -\nabla U(\theta_t^{(2)})\,dt + \sqrt{2\tau_2}dW_t^{(2)}, \end{aligned} \tag{5}$$

其中，$W_t^{(1)}$和 $W_{(2)\,t}$是两个相互独立的布朗运动。reLD 算法还允许两个 LD 交换其状态，即从

$$(\theta_t^{(1)} = y_1, \theta_t^{(2)} = y_2) \text{ to } (\theta_{t+dt}^{(1)} = y_2, \theta_{t+dt}^{(2)} = y_1)$$

以概率 $a\min\{1, S(\theta_t^{(1)}, \theta_t^{(2)})\}\,dt$进行，其中 $a > 0$是一个常数，

$$S(\theta^{(1)}, \theta^{(2)}) \coloneqq \exp\left\{\tau_\delta\left(U(\theta^{(1)}) - U(\theta^{(2)})\right)\right\}, \tag{6}$$

且 $\tau_\delta = 1/\tau_1 - 1/\tau_2$。众所周知，reLD 的平稳分布为（Chen 等人，2019年）

$$\pi_{\mathrm{re}}(\theta_t^{(1)}, \theta_t^{(2)}) \propto \exp\left\{-\frac{U(\theta_t^{(1)})}{\tau_1} - \frac{U(\theta_t^{(2)})}{\tau_2}\right\}. \tag{7}$$

Deng 等人（2020年）首次提出了副本交换随机梯度朗之万扩散（reSGLD）。假设我们有一个能量函数估计器 $\widehat{U}(\theta)$和一个梯度估计器 $\widehat{\nabla U}(\theta)$。我们通过以下方式模拟两个温度化 LD 的离散时间近似:

$$\begin{aligned} \widehat{\theta}_{k+1,\eta}^{(1)} &= \widehat{\theta}_{k,\eta}^{(1)} - \eta_k\widehat{\nabla U}(\widehat{\theta}_{k,\eta}^{(1)}) + \sqrt{2\eta_k\tau_1}\xi_k^{(1)} \\ \widehat{\theta}_{k+1,\eta}^{(2)} &= \widehat{\theta}_{k,\eta}^{(2)} - \eta_k\widehat{\nabla U}(\widehat{\theta}_{k,\eta}^{(2)}) + \sqrt{2\eta_k\tau_2}\xi_k^{(2)}, \end{aligned} \tag{8}$$

其中，$\xi_k^{(1)}$ 和 $\xi_k^{(2)}$ 独立地服从标准正态分布。假设 $\widehat{U}(\theta) \sim \mathcal{N}(U(\theta)\ \sigma^2)$，Deng 等人（2020年）提出采用交换率 $a\eta_k\min\{1, \widehat{S}(\widehat{\theta}_{k,\eta}^{(1)}, \widehat{\theta}_{k,\eta}^{(2)})\}$，其中

$$\widehat{S}(\theta^{(1)}, \theta^{(2)}) \coloneqq e^{\tau_\delta\left(\widehat{U}(\theta^{(1)}) - \widehat{U}(\theta^{(2)}) - \tau_\delta\sigma^2\right)}. \tag{9}$$

由于对任意 $b \in \mathbb{R}$ 都有 $\mathbb{E}[e^{b\widehat{U}(\theta)}] = e^{bU(\theta)+b^2\sigma^2/2}$，因此 $\mathbb{E}[\widehat{S}(\theta^{(1)}, \theta^{(2)})] = S(\theta^{(1)}, \theta^{(2)})$。这种交换率并非完全“无偏”，因为根据詹森不等式，

$$\mathbb{E}[\min\{1,\ \widehat{S}(\theta^{(1)}, \theta^{(2)})\}] \le \min\{1, S(\theta^{(1)}, \theta^{(2)})\},$$

当 $\sigma > 0$ 时，严格不等式成立。Deng 等人（2020）通过实证发现，这条规则对大多数问题都有效。

Lin 等（2022）通过对两个链使用不同的估计量，将 reSGLD 推广了。更具体地说，假设我们有

$$\widehat{U}_1(\theta) \sim \mathcal{N}(U(\theta), \sigma_1^2) \text{ and } \widehat{U}_2(\theta) \sim \mathcal{N}(U(\theta), \sigma_2^2),$$

其中，U_1 是低温链中使用的能量函数估计器，$U_2(\theta)$用于高温链。m-reSGLD 算法通过以下方式更新两个离散时间过程 $\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}$

$$\begin{aligned} \widehat{\theta}_{k+1,\eta}^{(1)} &= \widehat{\theta}_{k,\eta}^{(1)} - \eta_k\widehat{\nabla U}_1(\widehat{\theta}_{k,\eta}^{(1)}) + \sqrt{2\eta_k\tau_1}\xi_k^{(1)} \\ \widehat{\theta}_{k+1,\eta}^{(2)} &= \widehat{\theta}_{k,\eta}^{(2)} - \eta_k\widehat{\nabla U}_2(\widehat{\theta}_{k,\eta}^{(2)}) + \sqrt{2\eta_k\tau_2}\xi_k^{(2)} \end{aligned} \tag{10}$$

交换率为 $a\eta_k \min\{1, \widehat{S}(\widehat{\theta}_k^{(1)},_\eta, \widehat{\theta}_k^{(2)},_\eta)\}$。令 a_1, a_2 为两个非负常数，使得 $a_1 + a_2 = 1$，我们可以通过以下方式计算函数 $\widehat{S}(\theta^{(1)}, \theta^{(2)})$

$$\begin{aligned}\widehat{S}(\theta^{(1)}, \theta^{(2)}) = \exp\Big\{\tau_\delta\Big[a_1\left(\widehat{U}_1(\theta^{(1)}) - \widehat{U}_1(\theta^{(2)})\right) + a_2\left(\widehat{U}_2(\theta^{(1)}) - \widehat{U}_2(\theta^{(2)})\right) \\ - (a_1^2\sigma_1^2 + a_2^2\sigma_2^2)\tau_\delta\Big]\Big\}.\end{aligned} \tag{11}$$

通过直接计算可以验证，

$$\mathbb{E}[\widehat{S}(\theta^{(1)}, \theta^{(2)})] = S(\theta^{(1)}, \theta^{(2)}).$$

3 基于偏差校正的 SGLD 快速退火

现在，我们提出方法 f-reSGLD。与 reSGLD 相比，该方法在准确性和效率方面均有所提升。为简化讨论，与（Deng 等人，2020年；Lin等人，2022年）一样，我们假设梯度估计量服从正态分布，即 $\widehat{\nabla U}(\theta) \sim \mathcal{N}(\nabla U(\theta)\ s(\theta)s(\theta)^\top)$，其中某个正定的 $s(\theta) \in \mathbb{R}^{p\times p}$。请注意，我们允许估计量的协方差矩阵依赖于 θ。

首先，考虑一个以 $\pi_\tau \propto e^{-U/\tau}$ 为目标分布的 SGLD 链。我们提出模拟具有以下动力学的离散-$\widehat{}$时间过程 θ

$$\widetilde{\theta}_{k+1,\eta} = \widetilde{\theta}_{k,\eta} - \eta_k\widehat{\nabla U}(\widetilde{\theta}_{k,\eta}) + \sqrt{2}c_k(\widetilde{\theta}_{k,\eta})\xi_k, \tag{12}$$

其中 $\xi_k \sim \mathcal{N}(0,\ I_p)$，并假设矩阵 $c_k(\theta)$ 为正定矩阵，使得

$$c_k(\theta)c_k(\theta)^\top \coloneqq \tau\eta_k I_p - \frac{1}{2}\eta_k^2 s(\theta)s(\theta)^\top. \tag{13}$$

只要令 η_k 足够小，这一点总是可以满足的。为了解释式 (13) 背后的推理，注意到我们可以将式 (12) 改写为

$$\begin{aligned}\widetilde{\theta}_{k+1,\eta} &= \widetilde{\theta}_{k,\eta} - \eta_k\nabla U(\widetilde{\theta}_{k,\eta}) + \eta_k(\nabla U(\widetilde{\theta}_{k,\eta}) - \widehat{\nabla U}(\widetilde{\theta}_{k,\eta})) + \sqrt{2}c_k(\widetilde{\theta}_{k,\eta})\xi_k \\ &= \widetilde{\theta}_{k,\eta} - \eta_k\nabla U(\widetilde{\theta}_{k,\eta}) - \eta_k s(\widetilde{\theta}_{k,\eta})\zeta_k + \sqrt{2}c_k(\widetilde{\theta}_{k,\eta})\xi_k,\end{aligned}$$

其中 $\zeta_k \sim \mathcal{N}(0,\ I_p)$，且与 ξ_k 无关。由式 (13) 可知，式 (12) 与精确更新式 (2) 相同。这称为随机梯度朗之万动力学方法的偏差校正（Teh等人，2016年；Vollmer等人，2016年）。

在实践中，我们通常不知道真实的协方差矩阵 $s(\theta)s(\theta)^\top$，但可以对其进行估计。假设我们有一个估计量 $\widehat{s}(\theta)$，满足

$$s(\theta)s(\theta)^\top = \widehat{s}(\theta)\widehat{s}(\theta)^\top + \varphi(\theta)\varphi(\theta)^\top,$$

其中 $\varphi(\theta)$ 满足

$$\mathrm{trace}(\varphi(\theta)\varphi(\theta)^\top) = o(\mathrm{trace}(s(\theta)s(\theta)^\top)) \tag{14}$$

对于任意 θ。在式 (14) 中。注意，

$$\mathrm{trace}(\varphi(\theta)\varphi(\theta)^\top) = \mathbb{E}[\|\varphi(\theta)\zeta\|_2^2]$$

其中 $\zeta \sim \mathcal{N}(0,\ I_p)$。然后，我们提出如下更新

$$\widetilde{\theta}_{k+1,\eta} = \widetilde{\theta}_{k,\eta} - \eta_k\widehat{\nabla U}(\widetilde{\theta}_{k,\eta}) + \sqrt{2}\widehat{c}_k(\widetilde{\theta}_{k,\eta})\xi_k, \tag{15}$$

其中假设 $\widehat{c}_k(\theta)$ 为正定矩阵，使得

$$\eta_k^2 \widehat{s}(\theta)\widehat{s}(\theta)^\top + 2\widehat{c}_k(\theta)\widehat{c}_k(\theta)^\top = 2\tau\eta_k I_p.$$

类似的计算表明，(15) 也可以表示为

$$\begin{aligned}\widetilde{\theta}_{k+1,\eta} = \widetilde{\theta}_{k,\eta} - \eta_k \nabla U(\widetilde{\theta}_{k,\eta}) \\ - \eta_k \varphi(\widetilde{\theta}_{k,\eta})\zeta_k + \sqrt{2\tau\eta_k}\xi_k.\end{aligned} \tag{16}$$

为将此方法扩展到多方差副本交换随机梯度朗之万动力学，我们假设可以获得以下独立估计量，

$$\begin{aligned}\widehat{U}_1(\theta) &\sim \mathcal{N}(U(\theta), \sigma_1^2(\theta)), & \widehat{U}_2(\theta) &\sim \mathcal{N}(U(\theta), \sigma_2^2(\theta)), \\ \widehat{\nabla U}_1(\theta) &\sim \mathcal{N}(\nabla U(\theta), s_1(\theta)s_1(\theta)^\top), & \widehat{\nabla U}_2(\theta) &\sim \mathcal{N}(\nabla U(\theta), s_2(\theta)s_2(\theta)^\top).\end{aligned}$$

此外，假设我们有估计量 $\hat{s}_1, \hat{s}_2$ 和正定矩阵 $\widehat{c}_{i,k}(\theta)$，使得

$$\eta_k^2 \widehat{s}_i(\theta)\widehat{s}_i(\theta)^\top + 2\widehat{c}_{i,k}(\theta)\widehat{c}_{i,k}(\theta)^\top = \tau\eta_k I_p$$

对于 $i = 1,2$。我们通过以下方式更新这两条链：

$$\begin{aligned}\widetilde{\theta}^{(1)}_{k+1,\eta} &= \widetilde{\theta}^{(1)}_{k,\eta} - \eta_k \widehat{\nabla U}_1(\widetilde{\theta}^{(1)}_{k,\eta}) + \sqrt{2}\widehat{c}_{1,k}(\widetilde{\theta}^{(1)}_{k,\eta})\xi_k^{(1)} \\ \widetilde{\theta}^{(2)}_{k+1,\eta} &= \widetilde{\theta}^{(2)}_{k,\eta} - \eta_k \widehat{\nabla U}_2(\widetilde{\theta}^{(2)}_{k,\eta}) + \sqrt{2}\widehat{c}_{2,k}(\widetilde{\theta}^{(2)}_{k,\eta})\xi_k^{(2)},\end{aligned} \tag{17}$$

此外，我们不再使用 (11)，而是定义新的交换率$a\eta_k \min\{1, \widetilde{S}(\widetilde{\theta}^{(1)}_k, \widetilde{\theta}^{(2)}_k)\}$，其中

$$\begin{aligned}\widetilde{S}(\theta^{(1)}, \theta^{(2)}) := \exp\Big\{\tau_\delta\Big(\widehat{U}_1(\theta^{(1)}) - \widehat{U}_2(\theta^{(2)}) \\ - \tau_\delta\left(\sigma_1^2(\theta^{(1)}) + \sigma_2^2(\theta^{(2)})\right)/2\Big)\Big\}.\end{aligned} \tag{18}$$

注意，当 $\sigma^2 = \sigma_1^2(\theta^{(1)}) = \sigma_2^2(\theta^{(2)})$ 时，(18) 就简化为等方差情形 (9)。与 (11) 相比，新的交换率的主要优点是，在 (18) 中我们只需计算一次 U_1 和 U_2。对于计算能量函数 U 极其耗时的问题（例如第5.2节中介绍的反问题偏微分方程），这可以显著节省计算时间。

引理3.1. 估计量 (18) 满足 $\mathbb{E}[\widetilde{S}(\theta^{(1)}, \theta^{(2)})] = S(\theta^{(1)}, \theta^{(2)})$，其中 $S(\theta^{(1)}, \theta^{(2)})$ 由 (9) 给出。

证明. 对 $\mathbb{E}[e^{bZ}] = e^{b^2/2}$ 在 $Z \sim \mathcal{N}(0,1)$ 上进行常规计算，并利用 $\widehat{U}_1(\theta^{(1)})$ 与 $\widehat{U}_2(\theta^{(2)})$ 的独立性，即可得到结论。 □

4 收敛性分析

本节将介绍我们的理论结果。为证明收敛结果，我们需要先给出更多定义和假设，以供后续使用。对于两个概率测度 μ 和ν，μ 和 ν 之间的 2-瓦瑟斯坦距离定义为

$$W_2(\mu,\nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int \|\theta_\mu - \theta_\nu\|_2^2 \, d\gamma(\theta_\mu, \theta_\nu)\right)^{1/2},$$

其中， $\Gamma(\mu,\ \nu)$是耦合空间，包含所有以 μ 和 ν 为边缘分布的联合分布 γ 。对于任意连续可微函数 f ，我们将狄利克雷形式定义为

$$\mathbb{D}(U) \coloneqq \int \tau_1 \|\nabla_{\theta^{(1)}} f\|_2^2 + \tau_2 \|\nabla_{\theta^{(2)}} f\|_2^2 \, d\pi_{\mathrm{re}}(\theta^{(1)}, \theta^{(2)}), \tag{19}$$

其中 π_{re} 是式 (7) 中定义的平稳分布。此外，定义

$$\begin{aligned} &\mathbb{D}_S(U) \coloneqq \mathbb{D}(U) \\ &+ \frac{a}{2} \int S(\theta^{(1)}, \theta^{(2)}) \left(f(\theta^{(2)}, \theta^{(1)}) - f(\theta^{(1)}, \theta^{(2)}) \right)^2 d\pi_{\mathrm{re}}(\theta^{(1)}, \theta^{(2)}). \end{aligned} \tag{20}$$

对于任意两个概率测度 μ 和 ν，我们将相对熵记为

$$\mathcal{E}(\mu|\nu) = \int \log \frac{d\mu}{d\nu} \, d\mu. \tag{21}$$

下面给出能量函数及其导数所需的一些假设。

1. 耗散性。函数 U 对于 $\alpha > 0$ 和 $\beta \geq 0$ 是 (α, β)-耗散的，即

$$\langle \theta, \nabla U(\theta) \rangle \geq \alpha \|\theta\|_2^2 - \beta$$

对所有 $\theta \in \mathbb{R}^p$ 均成立。

2. 光滑性。函数 U 的梯度是 M-利普希茨连续的，即

$$\|\nabla U(\theta) - \nabla U(\theta')\|_2 \leq M \|\theta - \theta'\|_2$$

对所有 $\theta,\ \theta' \in \mathbb{R}^p$。

这两个条件广泛用于随机梯度朗之万动力学的理论分析中（Raginsky 等人，2017年；Chen 等人，2019年；Deng 等人，2020年）。现在我们可以陈述第一个收敛结果。

定理 4.1。 设 $\widetilde{\beta}_{k,\ \eta} = (\widetilde{\theta}_k^{(1)}\ ,\ \eta,\ \ \widetilde{\theta}_k^{(2)}\ ,\ \eta)$ 由式 (17) 得到，且 μ_k 为其分布。记$\varphi_k^{(1)} \coloneqq \varphi(\widetilde{\theta}_k^{(1)}\ ,\ \eta)$ 和 $\varphi_k^{(2)} \coloneqq \varphi(\widetilde{\theta}_k^{(2)}\ ,\ \eta)$，阻塞矩阵 $\varphi_k \coloneqq \mathrm{diag}(\varphi_k^{(1)}\ ,\ \ \varphi_k^{(2)}\ )$。令

$$\psi_k = \widetilde{S}(\widetilde{\theta}_{k,\eta}^{(1)}, \widetilde{\theta}_{k,\eta}^{(2)}) - S(\widetilde{\theta}_{k,\eta}^{(1)}, \widetilde{\theta}_{k,\eta}^{(2)})$$

表示由交换率估计造成的误差。此外，设 π_{re} 为式 (7) 中定义的 reLD 的平稳分布。在满足耗散性 1 和光滑性 2 假设的条件下，我们有

$$\begin{aligned} W_2(\mu_k, \pi_{\mathrm{re}}) \leq \widetilde{C}_0 \exp\left\{ -\frac{k\eta(1 + C_{\mathbb{D}})}{C_{\mathcal{E}}} \right\} \\ + \widetilde{C}_{\tau_1,\tau_2,a,M,p} \left(\eta^{1/2} + \max_i \mathrm{trace}(\varphi_i \varphi_i^\top)^{1/2} + \max_i (\mathbb{E}[|\psi_i|^2])^{1/2} \right), \end{aligned}$$

其中 $C_{\mathcal{E}} > 0$ 是常数， $\widetilde{C}_0 = \sqrt{2C_{\mathcal{E}}\mathcal{E}(\mu_0|\pi)}$,

$$C_{\mathbb{D}} = \inf_{t>0} \frac{\mathbb{D}\left(\frac{d\nu_k}{d\pi}\right)}{\mathbb{D}_S\left(\frac{d\nu_k}{d\pi}\right)} - 1$$

$\widetilde{\ }$是一个取决于交换率 S 的非负常数，且 $C_{\tau_1,\ \tau_2,\ a,\ M,\ p}$ 是一个取决于 τ_1、 τ_2、 a、 M 和维度 p的常数。

证明。 思路是将 Deng 等人（2020）定理 1 中的 ϕ_k 替换为 $\varphi_k\zeta_k$，其中 $\zeta_k \sim \mathcal{N}(0,\ I_{2p})$。证明细节见附录 B。 □

将我们的定理 4.1 与 Deng 等人（2020）的定理 1 进行比较后发现，由梯度估计带来的误差显著降低，而这类误差是主要误差（Teh等人，2016年）。更具体地，记 $s_k^{(1)} := s(\widetilde{\theta}_k^{(1)},\ \eta)$ 和 $s_k^{(2)} := s(\widetilde{\theta}_k^{(2)},\ \eta)$，阻塞矩阵为 $s_k = \mathrm{diag}(s_k^{(1)},\ s_k^{(2)})$。在我们的设定 $\widehat{\nabla U}(\beta) \sim \mathcal{N}(\nabla U(\beta)\ s(\beta)s(\beta)^\top)$ 下，Deng 等人（2020年）中由梯度估计带来的误差为

$$\max_i \operatorname{trace}\left(s_i s_i^\top\right)^{1/2}$$

而我们的方法中的误差为

$$\max_i \operatorname{trace}\left(\varphi_i \varphi_i^\top\right)^{1/2}.$$

根据式（14），我们的方法显著降低了由梯度估计带来的误差。

副本交换方法通常用于求解具有多个局部模态的非凸问题。然而，它也可用于求解凸问题，尤其是在无法提供适当初始化的情况下。我们还给出了能量函数 $U(\theta)$ 为强凸时的收敛结果。

3. 强凸假设。能量函数 U 是 m-强凸的，即

$$U(\theta) - U(\theta') - \nabla U(\theta)^{\mathrm{T}}(\theta - \theta') \geq \frac{m}{2}\|\theta - \theta'\|_2^2$$

对所有 $\theta,\ \theta' \in \mathbb{R}^p$。

当能量函数为强凸时，我们有如下估计。

定理 4.2. 设 $\beta_k,\ \eta, \varphi_k, \psi_k$ 和 π_{re} 与定理 4.1 中相同。在强凸性 3 和光滑性 2 假设下，若 $\eta < 1/(m + M)$，则有

$$\begin{aligned} W_2(\mu_k, \pi_{\mathrm{re}}) \leq\ & (1 - m\eta)^k W_2(\mu_0, \pi) \\ & + C_{\tau_1,\tau_2,a,m,M}\left\{\eta \max_i \operatorname{trace}(\varphi_i\varphi_i^\top) + \max_i p\mathbb{E}[|\psi_i|] + \eta p\right\}^{1/2}, \end{aligned}$$

其中 $C_{\tau_1,\ \tau_2,\ a,\ m,\ M}$ 是一个依赖于 τ_1, τ_2, a, m 和 M的常数。

证明。 证明过程见附录 A。 □

与非凸情形下的定理 4.1 相比，得益于强凸性，在定理 4.2 中，随着步长 η 趋于 0，由梯度估计产生的偏差将消失，只剩下由交换率估计产生的偏差。证明的主要思路与 Dalalyan 和 Karagulyan（2019）中的定理 4 类似，但我们将梯度估计所导致的误差从 $\max_i$ 迹 $\left(s_i s_i^\top\right)^{1/2}$ 降低到 $\max_i$ 迹 $\left(\varphi_i\varphi_i^\top\right)^{1/2}$。

5 实验

我们将通过几个数值示例展示该方法的优势。

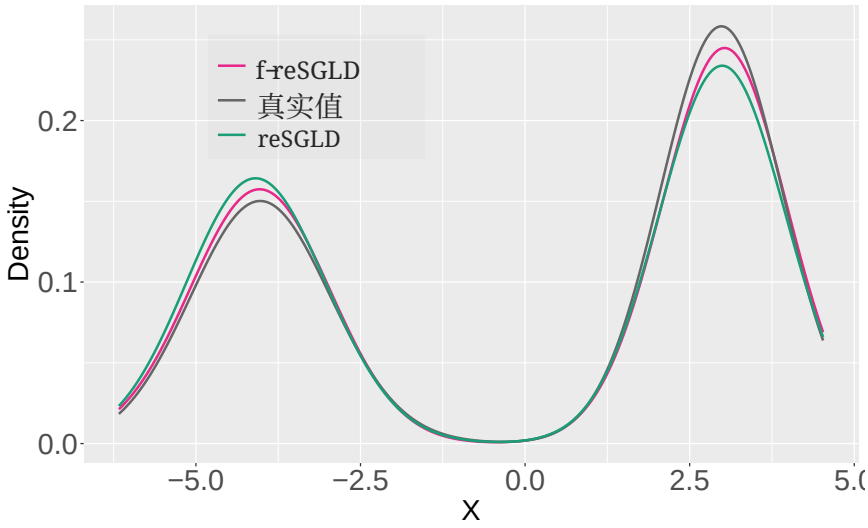


图 1：观测噪声方差固定的高斯混合示例 .

5.1 高斯混合分布模拟

本节中，我们评估该方法在高斯混合分布上的表现。设高温$\tau_2 = 10$ 和低温 $\tau_1 = 1$，则从低温下采样得到的样本服从我们所需的目标分布。考虑高斯混合分布

$$\exp(-U(\theta)) = 0.4\mathcal{N}(\theta; -4, 0.7^2) + 0.6\mathcal{N}(\theta; 3, 0.5^2). \tag{22}$$

我们假设，在低温链中只能获得无偏估计量 $\widehat{U}_1(\theta) \sim \mathcal{N}(U(\theta), 1^2)$，而在高温链中只能获得 $\widehat{U}_2(\theta) \sim \mathcal{N}(U(\theta), 3^2)$。对于梯度，我们也只能获得加噪后的版本，其中低温下为 $\widehat{\nabla U}_1(\theta) \sim \mathcal{N}(\nabla U(\theta), 2^2)$，高温下为 $\widehat{\nabla U}_2(\theta) \sim \mathcal{N}(\nabla U(\theta), 5^2)$。在更新步骤中，我们将步长固定为 0.03。由于实际中不知道方差值，为了实现我们的方法，我们采用 Deng 等人（2020）中的相同方法来估计方差。具体而言，在每个状态 $\widetilde{\theta}_k^{(l)}$，下，我们计算样本方差 $\widehat{s}^2(\widetilde{\theta}_k^{(l)})$，并通过以下方式更新方差估计量

$$\widehat{s}_k^2 = (1 - 1/k)\widehat{s}_{k-1}^2 + (1/k)\widehat{s}^2(\widetilde{\theta}_{k,\eta}^{(l)}),$$

其中，$\widehat{s}_k^2$ 是第 k 步的方差估计量，l 在低温样本中取 1，在高温样本中取 2。我们使用相同的方法估计了 σ_1^2 和 σ_2^2，即能量函数估计量的方差。

图1展示了在此设置下 1000 个样本的密度图，其中黑线表示真实值，绿线表示基本 reSGLD方法的结果，红线表示我们方法的结果。显然，我们的方法拟合得更好，相较之下，reSGLD方法的结果更接近真实值。

在式（22）给出的同一高斯混合分布下，接下来我们考虑一种更一般的情况：误差方差 $s^2(\theta)$ 取决于状态 θ。在我们的模拟中，假设

$$s(\theta) = \frac{5\exp(\theta)}{1 + \exp(\theta)}, \text{and } \sigma(\theta) = \frac{3\exp(U(x))}{2(1 + \exp(U(x)))},$$

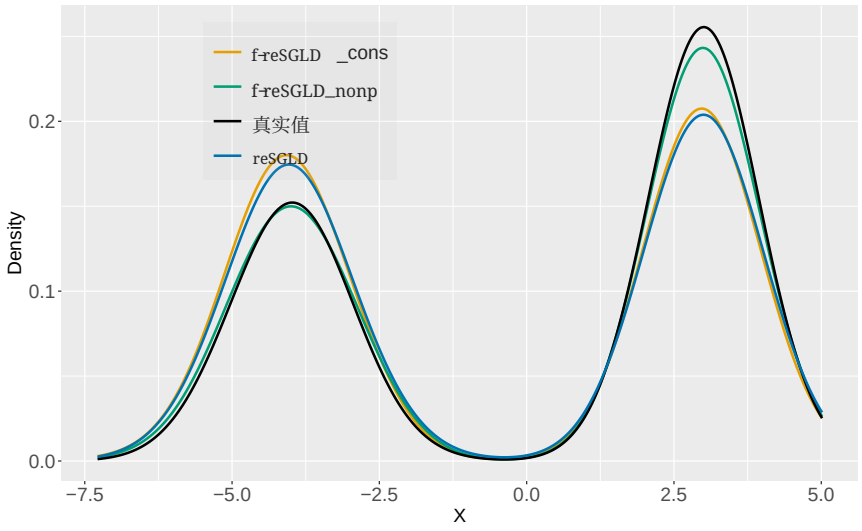


图 2：观测噪声方差固定的高斯混合示例 .

其中 $\widehat{U}(\theta) \sim \mathcal{N}(U(\theta)\ \sigma^2(\theta))$，且 $\widehat{\nabla U\ \theta}(\ ) \sim \mathcal{N}(\nabla U(\theta)\ s^2(\theta))$。我们尝试了两种不同的方法来估计方差。第一种方法假设方差为常数，并采用上文所述 Deng 等人（2020）的方法。我们将此方法称为“f-reSGLD cons”。第二种方法采用非参数方法获得方差估计量，记为“f-reSGLD nonp”。具体而言，我们收集了高温链中的前 100 个样本$\{\widetilde{\theta}_i^{(2)}\}_{i=1}^{100}$ 及其样本方差 $\{\widehat{s}^2(\widetilde{\theta}_i^{(2)})\}_{i=1}^{100}$。接下来，我们拟合一个核岭回归（KRR）Wainwright（2019），得到方差估计量 $\mathcal{F}$。然后，在后续更新中，对于每个状态 θ，我们通过 $\mathcal{F}(\theta)$ 计算其方差。

图2展示了非恒定方差情形下的结果，其中黑线表示真实值，蓝线表示标准 reSGLD，黄线表示我们的方法（将方差估计为常数），绿线表示通过 KRR 估计方差的 f-reSGLD 方法。reSGLD 和 f-reSGLD cons 的表现相近，而 f-reSGLD nonp 的表现远优于前两者，也更接近真实值。这表明，我们得到的方差估计量越好，方法的表现就越好。

5.2 逆偏微分方程

接下来，我们展示一个逆偏微分方程示例。该示例旨在说明，所提出的采样方法能够在似然函数和梯度估计中存在较大噪声时，捕捉多模态逆感兴趣量（iQoI）。具体而言，我们表明，与 reSGLD 相比，所提出的方法在注入噪声大幅降低的情况下，仍能达到相同的准确性和有效性（采样次数）；而未经噪声校正的方法则无法捕捉所有 iQoI。

我们对问题进行设计，使其具有无限多个 iQoI。逆偏微分方程依赖于

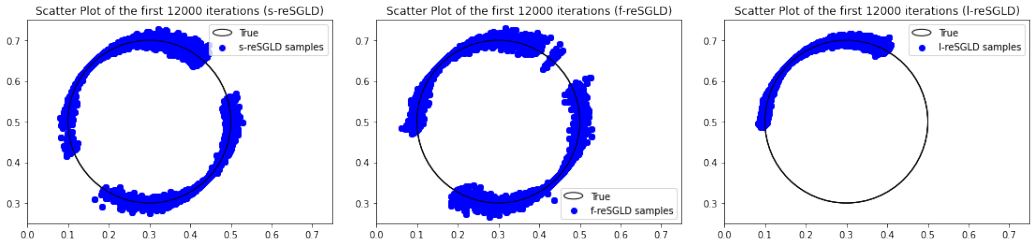


图 3：三种方法各自的前 12,000 个样本。左：注入少量噪声的 reSGLD（s-reSGLD）。中：所提出的注入大量噪声的方法（f-reSGLD）。右：注入与所提出方法（f-reSGLD）相同大量噪声的 reSGLD（l-reSGLD）。

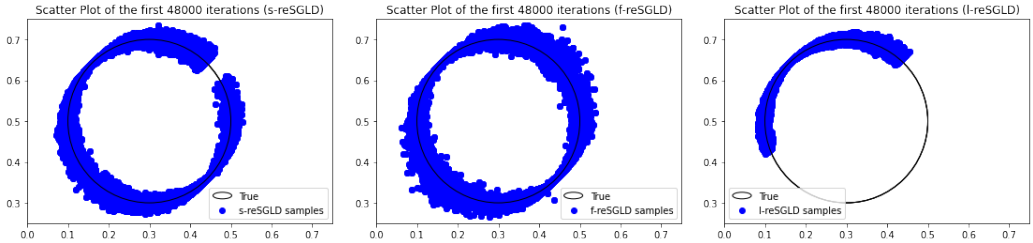


图 4：三种方法各自的前 48,000 个样本。左：注入少量噪声的 reSGLD（s-reSGLD）。中：所提出的注入大量噪声的方法（f-reSGLD）。右：注入与所提出方法（f-reSGLD）相同大量噪声的 reSGLD（l-reSGLD）。

以下模型方程，

$$
\begin{aligned}
&u_t = \nabla \cdot \nabla u + f, x \in \Omega = [0,1]^2, , t \in [0,T] \\
&u(x,t) = 0, x \in \partial\Omega, \\
&u(x,0) = \beta e^{-(x-x_0)^2/\alpha}.
\end{aligned}
$$

初始条件未知，即 x_0 未知。目标是在终端时间 T，利用单个传感器在位置 x_s 的测量值 u（x，t），追踪 x_0。我们将精确解设为 u（x，t）$= \beta e^{-(x-x_0)^2/\alpha} e^{-t}$；如果我们只在 x_s 放置一个传感器，iQoI 将是一个以 x_s 为圆心的圆。在本文中，$\beta = 1/(2\pi h^2)$，$\alpha = 2h^2$，其中 $h = 0.1$。$T = 0.03$，传感器放置在 (0.3，0.5)。

我们进行了三组实验：采用小噪声的原始 reSGLD（s-reSGLD）、采用大噪声的快速 reSGLD（f-reSGLD），以及采用大噪声的原始 reSGLD（l-reSGLD）。我们在似然函数和梯度估计中手动注入噪声。具体而言，我们在 s-reSGLD 的似然函数中注入高斯噪声 $\mathcal{N}(0, 0.1_2)$，同时在 f-reSGLD 和 l-reSGLD 的似然函数中注入高斯噪声 $\mathcal{N}(0, 0.8_2)$。对于似然函数的梯度，我们在 s-reSGLD 中注入高斯噪声 $\mathcal{N}(0, 0.1_2)$，同时在 f-reSGLD 和 l-reSGLD 中注入高斯噪声 $\mathcal{N}(0, 2^2)$。此外，在这三组实验中，低温均为 0.08，高温均为 0.5，据此推导出所提方法的有效温度。结果如图 3 和图 4 所示。

从图 3 和图 4 可以看出，所提出的方法即使在似然函数和似然函数梯度中存在较大噪声，仍能捕捉真实的 iQoI；然而，具有相同噪声的原始 reSGLD 无法捕捉所有 iQoI。

6 结论

在这项工作中，我们提出了一种方法，可以降低对能量函数估计器准确性的要求，同时保持收敛性。具体而言，与原始 reSGLD 相比，我们的估计降低了梯度估计的误差。此外，我们还在定理 4.2 中给出了强凸假设下另一种关于收敛性的估计。与非凸情形相比，随着学习率趋于零，梯度误差将消失。此外，我们提出了一种新的交换率估计器，并证明它是无偏的。与 m-reSGLD 中的原始交换率相比，我们的交换率无需额外的似然评估，速度也快得多。若干数值示例表明，该方法能够容忍梯度估计中的噪声，并实现相同的收敛性；而原始 m-reSGLD 则可能无法收敛。未来的工作将探讨如何在去除凸性假设后获得更紧的界。

致谢

GL 和 ZZ 衷心感谢美国国家科学基金会（DMS- 1555072、DMS-2053746 和 DMS-2134209）、布鲁克海文国家实验室分包合同 382247，以及美国能源部（DOE）科学办公室先进科学计算研究计划 DE-SC0021142 和 DE-SC0023161 的支持。

参考文献

Andrieu, C., Roberts, G.O., 2009. 高效蒙特卡罗计算的伪边际方法。《统计学年鉴》37, 697–725. Besag, J., 1994. 对 U. Grenander 和 M. I. Miller 所著《复杂系统中的知识表征》的评论。《皇家统计学会志：B辑（统计方法）》56, 4. Bhattacharya, R., 1978. 多维扩散过程复返性及不变测度存在性的判据。《应用概率年刊》, 541–553. Chen, Y., Chen, J., Dong, J., Peng, J., Wang, Z., 2019. 通过副本交换朗之万扩散加速非凸学习，载于：第七届国际学习表征会议，ICLR 2019。Chung, E., Efendiev, Y., Leung, W.T., Pun, S.M., Zhang, Z., 2020. 多智能体强化学习加速多尺度反演问题中的马尔可夫链蒙特卡罗。arXiv预印本 arXiv:2011.08954。Dalalyan, A., 2017. 采样与优化之间更深入、更有力的类比：朗之万蒙特卡洛与梯度下降，载于：学习理论会议，PMLR，第678–689页。Dalalyan, A.S., Karagulyan, A., 2019. 对梯度不准确的朗之万蒙特卡洛方法提供易于使用的保证。《随机过程及其应用》129, 5278–5311. Deng, W., Feng, Q., Gao, L., Liang, F., Lin, G., 2020. 通过副本交换随机梯度马尔可夫链蒙特卡罗进行非凸学习，载于：国际机器学习会议，PMLR，第2474–2483页。Dupuis, P., Liu, Y., Plattner, N., Doll, J.D., 2012. 关于并行回火的无限交换极限。《多尺度建模与仿真》10, 986–1022。

Dwivedi, R., Chen, Y., Wainwright, M.J., Yu, B., 2018. 对数凹采样：Metropolis-Hastings算法速度很快！，载于：学习理论会议，PMLR，第793–797页。Efendiev, Y., Hou, T., Luo, W., 2006. 使用粗尺度模型对马尔可夫链蒙特卡罗模拟进行预条件。SIAM科学计算杂志28, 776–803. Gramacy, R., Samworth, R., King, R., 2010. 重要性退火。统计学与计算20, 1–7. Lin, G., Moya, C., Zhang, Z., 2021. 加速的副本交换随机梯度朗之万扩散增强贝叶斯DeepONet，用于求解含噪参数偏微分方程。arXiv预印本 arXiv:2111.02484。Lin, G., Wang, Y., Zhang, Z., 2022. 通过贝叶斯PINN求解逆问题和正问题的多方差副本交换随机梯度马尔可夫链蒙特卡罗。《计算物理学杂志》460, 111173. Machta, J., 2009. 并行回火的优势与不足。《物理评论E》80, 056706. Mangoubi, O., Smith, A., 2021. 强对数凹分布上的哈密顿蒙特卡洛混合：连续动力学。《应用概率年刊》31, 2019–2045. Na, O., Zhang, Z., Lin, G., 2022. 一种用于贝叶斯逆问题的副本交换预条件Crank-Nicolson朗之万动力学蒙特卡罗方法。arXiv预印本 arXiv:2210.17048。Nguyen, T.H., Simsekli, U., Richard, G., 2019. 用于非凸优化的分数阶朗之万蒙特卡洛的非渐近分析，载于：国际机器学习会议，PMLR，第4810–4819页。Raginsky, M., Rakhlin, A., Telgarsky, M., 2017. 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，PMLR，第1674–1703页。Roberts, G.O., Rosenthal, J.S., 1998. 朗之万扩散离散近似的最优缩放。《皇家统计学会志：B辑（统计方法）》60, 255–268. Roberts, G.O., Stramer, O., 2002. 朗之万扩散与Metropolis-Hastings算法。《应用概率方法论与计算》4, 337–357. Roberts, G.O., Tweedie, R.L., 1996. 朗之万分布及其离散近似的指数收敛。《伯努利》, 341–363. Simsekli, U., Zhu, L., Teh, Y.W., Gurbuzbalaban, M., 2020. 分数阶欠阻尼朗之万动力学：在重尾梯度噪声下重新定位带动量的随机梯度下降，载于：国际机器学习会议，PMLR，第8970–8980页。Stuart, A.M., 2010. 逆问题：贝叶斯视角。《数值学报》19, 451–559. Teh, Y.W., Thiery, A.H., Vollmer, S.J., 2016. 随机梯度朗之万动力学的一致性与波动。《机器学习研究杂志》17. Vollmer, S.J., Zygalakis, K.C., Teh, Y.W., 2016. 探究随机梯度朗之万动力学的（非）渐近偏差与方差。《机器学习研究杂志》17, 5504–5548. Wainwright, M.J., 2019. 高维统计学：非渐近视角。第48卷。剑桥大学出版社。

Welling, M., Teh, Y.W., 2011年。基于随机梯度朗之万动力学的贝叶斯学习，载于：第28届国际机器学习会议（ICML-11）论文集，Citeseer。第681–688页。

Zhang, Y., Liang, P., Charikar, M., 2017年。随机梯度朗之万动力学的命中时间分析，载于：学习理论会议，PMLR。第1980–2022页。

定理 4.2 的证明

由于位置交换等价于温度交换（Dupuis等人，2012；Chen 等人，2019年），我们可以用以下随机微分方程对 reLD 建模：

$$d\beta_t = -\nabla G(\beta_t)dt + \Sigma_t dW_t,$$

其中 $\beta_t = (\theta_t^{(1)}, \theta_t^{(2)})$，$G(\beta_t) = [U(\theta_t^{(1)}) \; U(\theta_t^{(2)})]$，$W_t \in \mathbb{R}^{2p}$是布朗运动，而 Σ_t是一个在这些对角矩阵之间进行交换的随机矩阵

$$M_1 = \begin{bmatrix} \sqrt{2\tau_1 I_p} & 0 \\ 0 & \sqrt{2\tau_2 I_p} \end{bmatrix} \text{ and } M_2 = \begin{bmatrix} \sqrt{2\tau_2 I_p} & 0 \\ 0 & \sqrt{2\tau_1 I_p} \end{bmatrix}$$

以概率 $aS(\theta_t^{(1)}, \theta_t^{(2)})dt$，$I_p$ 是 p 维单位矩阵。

回顾一下，我们的更新过程为

$$\widetilde{\beta}_{k+1,\eta} = \widetilde{\beta}_{k,\eta} - \eta \nabla \widetilde{G}(\widetilde{\beta}_{k,\eta}) + \sqrt{\eta}\widetilde{\Sigma}_{k,\eta}\xi_k,$$

其中 $\widetilde{G}(\widetilde{\beta}_{k,\eta}) = [\widetilde{U}(\widetilde{\theta}_k^{(1),\eta}) \; \widetilde{U}(\widetilde{\theta}_k^{(2),\eta})]$，且 $\widetilde{\Sigma}_{k,\eta}$是一个随机矩阵，以概率 $a\eta\widetilde{S}(\widetilde{\theta}_k^{(1),\eta}, \widetilde{\theta}_k^{(2),\eta})$在对角矩阵 M_1 与 M_2 之间切换。不失一般性，我们假设 $\widetilde{}$我们的初始点为 $\beta_{0,\eta} \sim \mu_0$。回顾一下，$\mu_k$ 是 $\beta_{k,\eta}$的分布。为便于后续使用，我们将去噪更新定义为

$$\beta_{k+1,\eta} = \beta_{k,\eta} - \eta \nabla G(\beta_{k,\eta}) + \sqrt{\eta}\Sigma_{k,\eta}\xi_k,$$

其中 $\Sigma_{k,\eta}$ is a 在 M_1 和 M_2 之间以概率 $a\eta S(\theta_k^{(1),\eta}, \theta_k^{(2),\eta})$ 切换的随机矩阵 .

$\widetilde{}$证明。令 L_0 为从 π_{re} 中抽取的随机向量，使得 $W_2(\mu_0, \pi_{\text{re}}) = \mathbb{E}[\|\beta_{0,\eta} - L_0\|_2]$。我们将随机过程 $\{L_t\}_{t\geq 0}$ 定义为

$$L_t = L_0 - \int_0^t \nabla G(L_s)\, ds + \int_0^t \Sigma_s\, dW_s.$$

由于 π_{re}是平稳分布，因此我们知道，对于所有 $t \geq 0$，都有 $L_t \sim \pi_{\text{re}}$。注意到

$$L_{(k+1)\eta} = L_{k\eta} - \int_{k\eta}^{(k+1)\eta} \nabla G(L_s)\, ds + \int_{k\eta}^{(k+1)\eta} \Sigma_s\, dW_s.$$

~记 $\Delta_k = L_{k\eta} - \beta_{k,\ \eta}$，则

$$\begin{aligned}
\Delta_{k+1} &= L_{k\eta} - \int_{k\eta}^{(k+1)\eta} \nabla G(L_s)\, ds + \int_{k\eta}^{(k+1)\eta} \Sigma_s\, dW_s \\
&\quad - (\widetilde{\beta}_{k,\eta} - \eta \nabla \widetilde{G}(\widetilde{\beta}_{k,\eta}) + \sqrt{\eta}\widetilde{\Sigma}_{k,\eta}\xi_k) \\
&= \Delta_k + \eta \nabla \widetilde{G}(\widetilde{\beta}_{k,\eta}) - \int_{k\eta}^{(k+1)\eta} \nabla G(L_s)\, ds + \int_{k\eta}^{(k+1)\eta} (\Sigma_s - \widetilde{\Sigma}_{k,\eta})\, dW_s \\
&= \Delta_k - \eta \underbrace{(\nabla G(L_{k\eta}) - \nabla G(\widetilde{\beta}_{k,\eta}))}_{:=A_1} - \underbrace{\int_{k\eta}^{(k+1)\eta} (\nabla G(L_s) - \nabla G(L_{k\eta}))\, ds}_{:=A_2} \\
&\quad + \eta \underbrace{(\nabla \widetilde{G}(\widetilde{\beta}_{k,\eta}) - \nabla G(\widetilde{\beta}_{k,\eta}))}_{:=A_3} \\
&\quad + \underbrace{\int_{k\eta}^{(k+1)\eta} (\Sigma_s - \Sigma_{k,\eta})\, dW_s}_{:=A_4} + \underbrace{\int_{k\eta}^{(k+1)\eta} (\Sigma_{k,\eta} - \widetilde{\Sigma}_{k,\eta})\, dW_s}_{:=A_5}.
\end{aligned} \tag{23}$$

注意到 $A_3 = (\varphi_k^{(1)}\zeta_k^{(1)},\ \varphi_k^{(2)}\zeta_k^{(2)})$，其中 $\zeta_k^{(1)},\ \zeta_k^{(2)} \sim \mathcal{N}(0,\ I_p)$。根据 $\varphi_k^{(1)}$和$\varphi_k^{(2)}$ 的定义，给定 $\widetilde{\beta}_{k,\ \eta}$ 和 $\mathbb{E}[A_3|\widetilde{\beta}_{k,\eta}] = 0$ 时，A_3 与所有其他项相互独立。此外，我们有

$$\mathbb{E}[\|A_3\|_2^2] = \mathrm{trace}(\psi_k \psi_k^\top). \tag{24}$$

~注意到，给定 $\beta_{k,\ \eta}$ 时，A_4 也与所有其他项相互独立。根据 Itˆo ~积分的定义，我们知道 $\mathbb{E}[A_4|\beta_{k,\eta}] = 0$。根据 Itˆo 等距性，我们有

$$\begin{aligned}
&\mathbb{E}\left[\left\|\int_{k\eta}^{(k+1)\eta} (\Sigma_s - \Sigma_{k,\eta})\, dW_s\right\|_2^2\right] \\
=&\sum_{j=1}^{2p} \int_{k\eta}^{(k+1)\eta} \mathbb{E}[(\Sigma_s(j) - \Sigma_{k,\eta}(j))^2]\, ds \\
=&\sum_{j=1}^{2p} 2(\sqrt{\tau_2} - \sqrt{\tau_1})^2 \int_{k\eta}^{(k+1)\eta} P(\Sigma_s(j) \neq \Sigma_{k,\eta}(j))\, ds.
\end{aligned}$$

按照 Chen 等人（2019年）；Deng 等人（2020年）的讨论，根据条件概率

$$P(\Sigma_s(j) \neq \Sigma_{k,\eta}(j)|\widetilde{\beta}_{k,\eta}) = aS(\widetilde{\theta}_{k,\eta}^{(1)}, \widetilde{\theta}_{k,\eta}^{(1)})(s - k\eta) + ao(s - k\eta),$$

其中 $o(s - k\eta)$ 是关于 $s - k\eta$ 的高阶余项。因此，

$$\begin{aligned}
\mathbb{E}[\|A_4\|_2^2] &= \sum_{j=1}^{2p} 2(\sqrt{\tau_2} - \sqrt{\tau_1})^2 a \left[\int_{k\eta}^{(k+1)\eta} (s - k\eta) + o(s - k\eta)\, ds\right] \\
&\leq \delta_1(\tau_1, \tau_2, a) p \eta^2,
\end{aligned} \tag{25}$$

其中 $\delta_1(\tau_1, \tau_2, a)$ 是一个取决于 τ_1、τ_2 和 a 的常数。

~由于 $A_5^- = \sqrt{\eta}(\Sigma_{k,\eta} - \Sigma_{k,\eta})\zeta$，其中 $\zeta \sim \mathcal{N}(0, I_{2p})$，因此 $\mathbb{E}[A_5] = 0$。此外， $\mathbb{E}[\|A_5\|_2^2]$ 的上界可由下式得到

$$\begin{aligned}\mathbb{E}[\|A_5\|_2^2] &= \sum_{j=1}^{2p} 2(\sqrt{\tau_2} - \sqrt{\tau_1}) \int_{k\eta}^{(k+1)\eta} P(\Sigma_{k,\eta}(j) \neq \widetilde{\Sigma}_{k,\eta}(j))\, ds \\ &= \sum_{j=1}^{2p} 2(\sqrt{\tau_2} - \sqrt{\tau_1})^2 a\eta \mathbb{E}[|S(\widetilde{\theta}_{k,\eta}^{(1)}, \widetilde{\theta}_{k,\eta}^{(2)}) - \widetilde{S}(\widetilde{\theta}_{k,\eta}^{(1)}, \widetilde{\theta}_{k,\eta}^{(2)})|] \\ &\leq \delta_2(\tau_1, \tau_2, a) p\eta \mathbb{E}[|\psi_k|],\end{aligned} \tag{26}$$

其中， $\delta_2(\tau_1, \tau_2, a)$是一个依赖于 τ_1、 τ_2和 a的常数。界（26）最初由 Deng 等人（2020年）给出，这里为方便读者再次列出。

~由于在给定 β的条件下， A_3 和 A_4 的期望均为 0，并且结合 $\mathbb{E}[A_5] = 0$以及（24）、（25）和（26）中的结果，我们有

$$\begin{aligned}\mathbb{E}[\|\Delta_{k+1}\|_2^2] &= \mathbb{E}[\|\Delta_k - \eta A_1 - A_2 + \eta A_3 + A_4 + A_5\|_2^2] \\ &= \mathbb{E}[\|\Delta_k - \eta A_1 - A_2\|_2^2] + \eta^2 \mathbb{E}[\|A_3\|_2^2] + \mathbb{E}[\|A_4\|_2^2] + \mathbb{E}[\|A_5\|_2^2] \\ &\leq (\mathbb{E}[\|\Delta_k - \eta A_1\|_2] + \mathbb{E}[\|A_2\|])^2 \\ &\quad + \eta^2 \operatorname{trace}(\psi_k \psi_k^\top) + \delta_1 p \eta^2 + \delta_2 p \eta \mathbb{E}[|\psi_k|].\end{aligned} \tag{27}$$

根据 Dalalyan 和 Karagulyan（2019）中的引理 2，由于 U 是 m-强凸的，且 ∇U 是M-利普希茨的，并且 $\eta < 1/(M + m)$，我们有

$$\|\theta - \theta' - \eta(\nabla U(\theta) - \nabla U(\theta'))\|_2 \leq (1 - m\eta)\|\theta - \theta'\|_2.$$

因此，在我们的情形下，

$$\begin{aligned}&\|\beta - \beta' - \eta(\nabla G(\beta) - \nabla G(\beta'))\|_2^2 \\ =&\|\theta^{(1)} - \theta'^{(1)} - \eta(\nabla U(\theta^{(1)}) - \nabla U(\theta'^{(1)}))\|_2^2 \\ &+ \|\theta^{(2)} - \theta'^{(2)} - \eta(\nabla U(\theta^{(2)}) - \nabla U(\theta'^{(2)}))\|_2^2 \\ \leq&(1 - m\eta)^2 \|\theta^{(1)} - \theta'^{(1)}\|_2^2 + (1 - m\eta)^2 \|\theta^{(2)} - \theta'^{(2)}\|_2^2 \\ =&(1 - m\eta)^2 \|\beta - \beta'\|_2^2.\end{aligned}$$

因此，

$$\begin{aligned}\|\Delta_k - \eta A_1\|_2 =& \|L_{k\eta} - \widetilde{\theta}_{k,\eta} - \eta(\nabla G(L_{kh}) - \nabla G(\widetilde{\theta}_{k,h}))\|_2 \\ &\leq (1 - m\eta)\|\Delta_k\|_2.\end{aligned} \tag{28}$$

根据 Dalalyan 和 Karagulyan（2019）中的引理 3，由于函数 U 连续可微，且 U 的梯度是常数为 M的利普希茨函数，因此

$$\mathbb{E}[\|\nabla U(\theta)\|_2^2] \leq Mp.$$

对于术语 A_2，由于所有 L_t 都服从相同的分布，我们只需考虑以下情况

当 $k=0$ 时。然后，

$$\begin{aligned}
\mathbb{E}[\|A_2(0)\|_2] &:= \mathbb{E}\left[\left\|\int_0^\eta (\nabla G(L_s) - \nabla G(L_0))\, ds\right\|_2\right] \\
&\le \int_0^\eta \mathbb{E}[\|\nabla G(L_s) - \nabla G(L_0)\|_2]\, ds \\
&\le M \int_0^\eta \mathbb{E}[\|L_s - L_0\|_2]\, ds \\
&= M \int_0^\eta \mathbb{E}\left[\left\|-\int_0^s \nabla G(L_t)\, dt + \int_0^s \Sigma_t\, dW_t\right\|_2\right] ds \\
&\le M \underbrace{\int_0^\eta \int_0^s \mathbb{E}[\|\nabla G(L_t)\|_2]\, dt\, ds}_{:=B_1} + M \underbrace{\int_0^\eta \mathbb{E}\left[\left\|\int_0^s \Sigma_t\, dW_t\right\|_2\right] ds}_{:=B_2}.
\end{aligned}$$

对于项 B_1，由于对所有 $t \ge 0$， L_t 都服从分布 π，根据引理 3,

$$\mathbb{E}[\|\nabla G(L_t)\|_2^2] = \mathbb{E}[\|\nabla G(L_0)\|_2^2] = \mathbb{E}[\|\nabla U(L_0^{(1)})\|_2^2] + \mathbb{E}[\|\nabla U(L_0^{(2)})\|_2^2] \le 2Mp.$$

因此，

$$B_1 \le \int_0^\eta \int_0^s \sqrt{2Mp}\, dt\, ds = \frac{\sqrt{2Mp}\eta^2}{2}. \tag{29}$$

由于 Σ_t 是一个对角矩阵，其元素对任意 t 均为 $\sqrt{2\tau_1}$ 或 $\sqrt{2\tau_2}$，我们有

$$B_2 \le \int_0^\eta \sqrt{2\tau_2}\mathbb{E}[\|W_s\|_2]\, ds \le \sqrt{2\tau_2} \int_0^\eta \sqrt{2ps}\, ds = \frac{4\sqrt{p\tau_2}\eta^{3/2}}{3}. \tag{30}$$

结合式 (29) 和 (30)，我们有

$$\mathbb{E}[\|A_2\|_2] \le \frac{\sqrt{2}M^{3/2}p^{1/2}\eta^2}{2} + \frac{4Mp^{1/2}\eta^{3/2}}{3}.$$

根据假设 $\eta < 1/M$，我们有

$$\mathbb{E}[\|A_2\|_2] \le \delta_3 M\eta^{3/2}p^{1/2}. \tag{31}$$

将式 (28) 和 (31) 代入式 (27)，我们有

$$\begin{aligned}
\mathbb{E}[\|\Delta_{k+1}\|_2^2] \le &\left((1-m\eta)\mathbb{E}[\|\Delta_k\|_2] + \delta_3 M\eta^{3/2}p^{1/2}\right)^2 \\
&+ \eta^2 \max_i \operatorname{trace}(\varphi_i\varphi_i^\top) + \delta_1 p\eta^2 + \delta_2 p\eta \max_i \mathbb{E}[|\psi_i|].
\end{aligned}$$

最后，根据 Dalalyan 和 Karagulyan（2019年）中的引理 1，我们得到

$$\begin{aligned}
&W_2(\mu_k, \pi_{\mathrm{re}}) \\
\le &(1-m\eta)^k W_2(\mu_0, \pi_{\mathrm{re}}) + \frac{\delta_3 M\eta^{3/2}p^{1/2}}{m\eta} \\
&+ \frac{\eta^2 \max_i \operatorname{trace}(\varphi_i\varphi_i^\top) + \delta_1 p\eta^2 + \delta_2 p\eta \max_i \mathbb{E}[|\psi_i|]}{\delta_3 M\eta^{3/2}p^{1/2} + \sqrt{m\eta\left(\eta^2 \max_i \operatorname{trace}(\varphi_i\varphi_i^\top) + \delta_1 p\eta^2 + \delta_2 p\eta \max_i \mathbb{E}[|\psi_i|]\right)}} \\
\le &(1-m\eta)^k W_2(\mu_0, \pi_{\mathrm{re}}) \\
&+ C_{\tau_1,\tau_2,a,m,M}\left(\eta \max_i \operatorname{trace}(\varphi_i\varphi_i^\top) + p\eta + p \max_i \mathbb{E}[\psi|]\right)^{1/2}.
\end{aligned}$$

□

附录 B 定理 4.1 的证明

证明。定义

$$\beta_t = \beta_0 - \int_0^t \nabla G(\beta_s)\, ds + \int_0^t \Sigma_s dW_s,$$

其中 $\beta_t = (\theta_t^{(1)},\ \theta_t^{(2)})$，$G(\beta_t) = [U(\theta_t^{(1)})\ U(\theta_t^{(2)})]$，$W_t \in \mathbb{R}^{2p}$ 是布朗运动，Σ_t 是一个在对角矩阵之间切换的随机矩阵

$$M_1 = \begin{bmatrix} \sqrt{2\tau_1 I_p} & 0 \\ 0 & \sqrt{2\tau_2 I_p} \end{bmatrix} \text{ and } M_2 = \begin{bmatrix} \sqrt{2\tau_2 I_p} & 0 \\ 0 & \sqrt{2\tau_1 I_p} \end{bmatrix}$$

以概率 $aS(\theta_t^{(1)},\ \theta_t^{(2)})dt$，$I_p$ 是 p 维单位矩阵。定义 $\beta_{k,\ \eta} := \beta_{k\eta}$，并将 $\beta_{k,\ \eta}$ 的分布记为 ν_k。注意，我们可以将更新式 (17) 改写为

$$\widetilde{\beta}_{k,\eta} = \widetilde{\beta}_0 - \int_0^{k\eta} \nabla \widetilde{G}(\widetilde{\beta}_{\lfloor s/\eta \rfloor,\eta})\, ds + \int_0^{k\eta} \widetilde{\Sigma}_{\lfloor s/\eta \rfloor,\eta} dW_s,$$

其中 $\lfloor x \rfloor$ 是不大于 x 的最大整数。将 $\beta_{k,\ \eta}$ 的分布记为 $\widetilde{\mu}_k$。为后续使用，我们将 $\{\beta_t^\eta\}$ 定义为 $\{\beta_{k,\eta}\}$ 的连续时间插值，它是一个由下式定义的连续时间随机过程：

$$\widetilde{\beta}_t^\eta = \widetilde{\beta}_0 - \int_0^t \nabla \widetilde{G}(\widetilde{\beta}_{\lfloor s/\eta \rfloor,\eta})\, ds + \int_0^t \widetilde{\Sigma}_{\lfloor s/\eta \rfloor,\eta} dW_s.$$

根据三角不等式，我们有

$$W_2(\mu_k, \pi_{\mathrm{re}}) \le W_2(\mu_k, \nu_k) + W_2(\nu_k, \pi_{\mathrm{re}}).$$

对于第二项，根据 Deng 等人（2020）中的引理 5，在满足耗散性 1 和光滑性 2 假设的条件下，我们有

$$W_2(\nu_k, \pi_{\mathrm{re}}) \le C_0 \exp\left\{ -\frac{k\eta(1 + C_{\mathbb{D}})}{C_{\mathcal{E}}} \right\}, \tag{32}$$

其中 $C_{\mathcal{E}} > 0$ 是常数，$C_0 = \sqrt{2C_{\mathcal{E}}\mathcal{E}(\nu_0|\pi)}$ 且

$$C_{\mathbb{D}} = \inf_{t>0} \frac{\mathbb{D}\left(\frac{d\nu_k}{d\pi}\right)}{\mathbb{D}_S\left(\frac{d\nu_k}{d\pi}\right)} - 1$$

是一个非负常数，取决于式 (6) 中定义的交换率 S。

接下来，我们来界定离散化误差项 $W_2(\mu_k,\ \nu_k)$。假设 $\beta_0 \sim \widetilde{\beta}_0$，那么，

$$\beta_t - \widetilde{\beta}_t^\eta = -\int_0^t \left(\nabla G(\beta_s) - \nabla \widetilde{G}(\widetilde{\beta}_{\lfloor s/\eta \rfloor,\eta}) \right) ds + \int_0^t (\Sigma_s - \widetilde{\Sigma}_{\lfloor s/\eta \rfloor,\eta})\, dW_s.$$

根据三角不等式和闵可夫斯基不等式，我们有

$$\begin{aligned} \mathbb{E}[\|\beta_t - \widetilde{\beta}_t^\eta\|_2^2]^{1/2} \le & \mathbb{E}\left[\underbrace{\left\| \int_0^t \left(\nabla G(\beta_s) - \nabla \widetilde{G}(\widetilde{\beta}_{\lfloor s/\eta \rfloor,\eta}) \right) ds \right\|_2^2}_{:=\mathcal{I}_1} \right]^{1/2} \\ & + \mathbb{E}\left[\underbrace{\left\| \int_0^t (\Sigma_s - \widetilde{\Sigma}_{\lfloor s/\eta \rfloor,\eta})\, dW_s \right\|_2^2}_{:=\mathcal{I}_2} \right]^{1/2}. \end{aligned} \tag{33}$$

我们先关注 $\mathcal{I}_1$。注意到

$$\begin{aligned}\mathbb{E}[\mathcal{I}_1] &\le \mathbb{E}\left[t\int_0^t \|\nabla G(\beta_s) - \nabla \widetilde{G}(\widetilde{\beta}_{\lfloor s/\eta\rfloor,\eta})\|_2^2\, ds\right] \\ &\le 3t\left\{\mathbb{E}\left[\underbrace{\int_0^t \|\nabla G(\beta_s) - \nabla G(\widetilde{\beta}^s_\eta)\|_2^2\, ds}_{:=\mathcal{I}_{1a}}\right]\right. \\ &\quad + \mathbb{E}\left[\underbrace{\int_0^t \|\nabla G(\widetilde{\beta}^s_\eta) - \nabla G(\widetilde{\beta}_{\lfloor s/\eta\rfloor,\eta})\|_2^2\, ds}_{:=\mathcal{I}_{1b}}\right] \\ &\quad \left. + \mathbb{E}\left[\underbrace{\int_0^t \|\nabla G(\widetilde{\beta}_{\lfloor s/\eta\rfloor,\eta}) - \nabla \widetilde{G}(\widetilde{\beta}_{\lfloor s/\eta\rfloor,\eta})\|_2^2\, ds}_{:=\mathcal{I}_{1c}}\right]\right\}.\end{aligned} \tag{34}$$

根据光滑性假设 2,

$$\mathbb{E}[\mathcal{I}_{1a}] \le M^2\mathbb{E}\left[\int_0^t \|\beta_s - \widetilde{\beta}^s_\eta\|_2^2\, ds\right]. \tag{35}$$

记 $\kappa = \lfloor t/\eta\rfloor$，根据光滑性假设 2,

$$\begin{aligned}&\mathbb{E}[\mathcal{I}_{1b}] \\ \le& M^2\sum_{i=0}^{\kappa}\mathbb{E}\left[\int_{i\eta}^{(i+1)\eta}\|\widetilde{\beta}^s_\eta - \widetilde{\beta}_{i,\eta}\|_2^2\, ds\right] \\ =& M^2\sum_{i=0}^{\kappa}\mathbb{E}\left[\int_{i\eta}^{(i+1)\eta}\left\|-\nabla\widetilde{G}(\widetilde{\beta}_{i,\eta})(s-i\eta) + \widetilde{\Sigma}_{i,\eta}\int_{i\eta}^{s} dW_t\right\|_2^2 ds\right] \\ \le& 2M^2\sum_{i=1}^{\kappa}\left\{\int_{i\eta}^{(i+1)\eta}(s-i\eta)^2\mathbb{E}[\|\nabla\widetilde{G}(\widetilde{\beta}_{i,\eta})\|_2^2]\, ds + \int_{i\eta}^{(i+1)\eta}\mathbb{E}\left[\left\|\widetilde{\Sigma}_{i,\eta}\int_{i\eta}^{s} dW_t\right\|_2^2\right] ds\right\}\end{aligned} \tag{36}$$

注意到

$$\begin{aligned}\mathbb{E}[\|\nabla\widetilde{G}(\widetilde{\beta}_{i,\eta})\|_2^2] &= \mathbb{E}[\|\nabla G(\widetilde{\beta}_{i,\eta}) + \varphi_i\zeta_i\|_2^2] \\ &\le 2\mathbb{E}[\|\nabla G(\widetilde{\beta}_{i,\eta})\|_2^2] + 4\,\mathrm{trace}(\varphi_i\varphi_i^\top)] \\ &= 2\mathbb{E}[\|\nabla G(\widetilde{\beta}_{i,\eta}) - \nabla G(\beta^*)\|_2^2] + 4\,\mathrm{trace}(\varphi_i\varphi_i^\top)] \\ &\le 2M^2\mathbb{E}[\|\widetilde{\beta}_{i,\eta} - \beta^*\|_2^2] + 4\,\mathrm{trace}(\varphi_i\varphi_i^\top)] \\ &\le 4M^2\mathbb{E}[\|\widetilde{\beta}_{i,\eta}\|_2^2] + 4M^2\mathbb{E}[\|\beta^*\|_2^2] + 4\,\mathrm{trace}(\varphi_i\varphi_i^\top)].\end{aligned}$$

根据 Chen 等人（2019）中的引理 D.2，若 $0 < \eta < \alpha/M^2$，则存在常数 $\widetilde{\delta}_1(\tau_2, M, \alpha, \beta)$，使得

$$\sup_{i\ge 0}\mathbb{E}[\|\widetilde{\beta}_{i,\eta}\|_2^2] \le \widetilde{\delta}_1(\tau_2, M, \alpha, \beta).$$

因此,

$$\mathbb{E}[\|\nabla\widetilde{G}(\widetilde{\beta}_{i,\eta})\|_2^2] \le \widetilde{\delta}_2(\tau_2, M, \alpha, \beta) + 4\,\mathrm{trace}(\varphi_i\varphi_i^\top)],$$

then

$$\int_{i\eta}^{(i+1)\eta}(s-i\eta)^2\mathbb{E}[\|\nabla\widetilde{G}(\widetilde{\beta}_{i,\eta})\|_2^2]\,ds\le\frac{1}{3}\eta^3(\widetilde{\delta}_2(\tau_2,M,\alpha,\beta)+4\,\mathrm{trace}(\varphi_i\varphi_i^\top)]). \tag{37}$$

~由于 $\Sigma_{i,\ \eta}$ 是一个对角矩阵，其对角元素为 $\sqrt{2\tau_1}$ 或 $\sqrt{2\tau_2}$，

$$\mathbb{E}\left[\left\|\widetilde{\Sigma}_{i,\eta}\int_{i\eta}^{s}dW_t\right\|_2^2\right]\le 8\tau_2 p(s-i\eta).$$

因此，

$$\int_{i\eta}^{(i+1)\eta}\mathbb{E}\left[\left\|\widetilde{\Sigma}_{i,\eta}\int_{i\eta}^{s}dW_t\right\|_2^2\right]ds\le 4\tau_2 p\eta^2. \tag{38}$$

将 (37) 和 (38) 代入 (36)，

$$\begin{aligned}\mathbb{E}[\mathcal{I}_{1b}]&\le 2M^2\sum_{i=1}^{\kappa}\left(\frac{1}{3}\eta^3\widetilde{\delta}_2(\tau_2,M,\alpha,\beta)+\frac{4p}{3}\eta^3\,\mathrm{trace}(\varphi_i\varphi_i^\top)]+4\tau_2 p\eta^2\right)\\&\le\widetilde{\delta}_3\kappa p\eta^2(1+\eta\max_i\mathrm{trace}(\varphi_i\varphi_i^\top)]).\end{aligned} \tag{39}$$

根据 φ_k 的定义，我们有

$$\mathcal{I}_{1c}\le\sum_{i=0}^{\kappa}2p\eta\,\mathrm{trace}(\varphi_i\varphi_i^\top)\le 2p\kappa\eta\max_i\mathrm{trace}(\varphi_i\varphi_i^\top). \tag{40}$$

将 (35)、(39) 和 (40) 代入 (34)，

$$\mathbb{E}[\mathcal{I}_1]\le 3tM^2\int_0^t\mathbb{E}[\|\beta_s-\widetilde{\beta}_\eta^s\|_2^2]\,ds+\widetilde{\delta}_4 pt(t+1)\left(\eta+\max_i\mathrm{trace}(\varphi_i\varphi_i^\top)]\right). \tag{41}$$

对于项 $\mathcal{I}_2$，根据伊藤等距公式，我们有

$$\begin{aligned}&\mathbb{E}[\mathcal{I}_2]\\=&\sum_{j=1}^{2p}\int_0^t\mathbb{E}[(\Sigma_s(j)-\widetilde{\Sigma}_{\lfloor s/\eta\rfloor,\eta}(j))^2]\,ds\\\le&\sum_{j=1}^{2p}\sum_{i=1}^{\kappa}\int_{i\eta}^{(i+1)\eta}\mathbb{E}[(\Sigma_s(j)-\widetilde{\Sigma}_{i,\eta}(j))^2]\,ds\\\le&2\sum_{j=1}^{2p}\sum_{i=1}^{\kappa}\left\{\int_{i\eta}^{(i+1)\eta}\mathbb{E}[(\Sigma_s(j)-\Sigma_{i,\eta}(j))^2]\,ds+\int_{i\eta}^{(i+1)\eta}\mathbb{E}[(\Sigma_{i,\eta}(j)-\widetilde{\Sigma}_{i,\eta}(j))^2]\,ds\right\},\end{aligned}$$

其中，我们将对角矩阵 M 中 j 对应的对角元素记为 $M(j)$。按照定理 4.2 的证明中对 (23) 中定义的 A_4 和 A_5 的讨论，我们可以将 $\mathcal{I}_2$ 界定为

$$\mathbb{E}[\mathcal{I}_2]\le\widetilde{\delta}_5(\tau_1,\tau_2,a)p(t+1)(\eta+\max_i\mathbb{E}[|\psi_i|]), \tag{42}$$

~其中 $\delta_5(\tau_1,\ \tau_2,\ a)$ 是一个取决于 τ_1、τ_2 和 a 的常数。

将（41）、（42）代入（33），

$$
\begin{aligned}
&\mathbb{E}[\|\beta_t - \widetilde{\beta}_t^{\eta}\|_2^2] \\
\leq & 6tM^2 \int_0^t \mathbb{E}[\|\beta_s - \widetilde{\beta}_{\eta}^{s}\|_2^2]\, ds + 2\widetilde{\delta}_4 pt(t+1)(\eta + \max_i \operatorname{trace}(\varphi_i \varphi_i^{\top})) \\
& + \widetilde{\delta}_5(\tau_1, \tau_2, a)p(t+1)(\eta + \max_i \mathbb{E}[|\psi_i|]) \\
\leq & 6tM^2 \int_0^t \mathbb{E}[\|\beta_s - \widetilde{\beta}_{\eta}^{s}\|_2^2]\, ds \\
& + \widetilde{\delta}_6 p(t+1)\left((t+1)\eta + t\max_i \operatorname{trace}(\varphi_i \varphi_i^{\top}) + \max_i \mathbb{E}[|\psi_i|].\right)
\end{aligned}
$$

通过应用 Gr¨onwall 积分形式的不等式，我们得到

$$
\mathbb{E}[\|\beta_t - \widetilde{\beta}_t^{\eta}\|_2^2] \leq \widetilde{\delta}_7(\tau_1, \tau_2, a, M, t, p)(\eta + \max_i \operatorname{trace}(\varphi_i \varphi_i^{\top}) + \max_i \mathbb{E}[|\psi_i|]).
$$

因此，

$$
\begin{aligned}
W_2(\mu_k, \nu_k) &\leq \mathbb{E}[\|\beta_{k,\eta} - \widetilde{\beta}_{k,\eta}\|_2] \\
&\leq \widetilde{C}(\tau_1, \tau_2, a, M, k, p)\left(\eta^{1/2} + \max_i \operatorname{trace}(\varphi_i \varphi_i^{\top})^{1/2} + \max_i \mathbb{E}[|\psi_i|]^{1/2}\right),
\end{aligned} \tag{43}
$$

结合式 (32)，证明完成。 □