arXiv:2301.03077v1 [stat.ML] 8 Jan 2023

针对（弱）对数凹后验分布的随机朗之万蒙特卡洛。

Marelys Crespo Navas[1], S´ebastien Gadat[2,3], Xavier Gendre[11]
ISAE-SUPAERO，图卢兹大学[2]图卢兹经济学院（法国国家科学研究中心 UMR 5314），图卢兹第一大学卡皮托尔分校[3] 法国大学研究院

2023年1月10日

摘要

本文研究了随机朗之万蒙特卡洛方法的一种连续时间版本，该方法在传统的过阻尼朗之万扩散过程中引入了一个随机采样步骤。该方法广泛应用于机器学习中的后验分布采样。我们特别关注其计算成本与n （生成后验分布的观测值数量）和 d （感兴趣参数所在环境空间的维度）之间的关系。我们的分析基于弱凸框架，并借助 Kurdyka-Lojasiewicz（KL）不等式进行参数化；该不等式可处理曲率消失的情形，与简单的强凸情形相比限制要宽松得多。我们证明，为获得熵意义下的 ε 近似，最终模拟时域的阶为（$d$log（n）^{2}）$^{(1+r)^{2}}$[log^{2}（ε^{-1}） $+n^{2}d^{2(1+r)}$ log$^{4(1+r)}$（n）]；采用参数为 $\left(n\right.$（$d$log（^{2}））$^{1+r}\left.\right)^{-1}$的泊松子采样，其中参数 r 出现在 KL 不等式中，取值范围为 0（强凸情形）至 1（极限拉普拉斯情形）。

关键词：朗之万蒙特卡洛采样；对数凹模型；弱凸性。AMS 分类：主分类 6265C05；次分类；62C10；65C30；60H3520。[1]

1 Mark马尔可夫随机朗之万动力学与主要结果

1.1 引言

动机 在最近几年中，机器学习领域发展出了大量方法，用于处理包含大量观测值（X_1，...， X_n）且嵌入高维空间 $\mathbb{R}^d$的大规模海量数据集。这些方法通常涉及对数据依赖函数的优化（用于频率学派学习），或对数据依赖测度进行采样（用于具有后验分布的贝叶斯学习）。在这两种方法中，瓶颈都在于 n 和 d的规模，这通常会导致使用标准算法时出现数值困难。本文从贝叶斯视角出发，研究后验分布的模拟问题。统计模型由定义在 $\mathcal{X}$上的密度函数集合（p_θ）$_{\theta\in\Theta}$ 描述，其中感兴趣参数$\theta^\star$ 属于 $\Theta=\mathbb{R}^d$，且假设（X_i）$_{1\leq i\leq n}$ 是在 $\mathcal{X}$ 中根据 $p_{\theta^\star}$分布的独立同分布观测值。标准的贝叶斯方法包括在 Θ 上定义先验分布 π_0，然后借助过阻尼朗之万扩散，通过数值概率近似对记为 μ_n 的后验分布进行采样（下文将其记为 exp(−U_{ν_n})）：

$$\mathrm{d}\theta_t=-\nabla U_{\nu_n}(t)\mathrm{d}t+\sqrt{2}\mathrm{d}B_t.$$

[1]我们感谢 Patrick Cattiaux 和 Arnaud Guillin 在泛函不等式（尤其是弱对数 Sobolev 不等式）方面提供的有益讨论和参考资料。

在本研究中，我们处理了 [39], 提出的朗之万蒙特卡洛（LMC）算法的一种改编版本，该版本利用了 [36]引入的一些随机算法的早期思想：作者不再使用前一个方程，而是提出对扩散过程进行修改，通过在观测集合（X_i）$_{1\leq i\leq n}$中采用采样策略，使 LMC 产生带噪声的漂移。在详细介绍为准确定义该方法所需的具体对象和算法之前，我们首先给出与之相关的一些文献概述。

最新进展 对过阻尼朗之万蒙特卡洛及许多其他采样算法的遍历性和定量混合性质的研究，是一个热门课题。大致而言，这项研究由概率论领域围绕两种策略展开的工作所开创。第一种策略依赖于路径考量和随机动力系统的动力学性质，并借助某种耦合论证和李雅普诺夫控制来构建。我们请参阅开创性研究 [32, 27], , 其中运用了杜布林耦合方法和全变差（TV）界。通过这种李雅普诺夫方法可以推导出许多推广，并得到瓦瑟斯坦界或 $\mathbb{L}^2$ 上界；关于李雅普诺夫条件与遍历性之间联系的说明，请参阅 [8] 及同一作者的相关参考文献。第二种策略源于马尔可夫算子的谱性质，并与著名的泛函不等式（如庞加莱不等式和对数索博列夫不等式）相关。其基本思想是沿时间演化对距离求导，并应用格朗沃尔引理，对半群的长期演化获得定量估计。我们请参阅 [26, 2], 的开创性研究，以及 [3] 对所有可能的不等式及其对马尔可夫半群遍历性的影响所作的近乎详尽的综述。最后要强调的是，正如 [9]所指出的，谱方法与李雅普诺夫方法之间存在紧密联系。如果泛函不等式与混合性质密切相关，尤其是从定量角度来看，那么就有必要发展一套能够仔细评估这些不等式的工具，特别要关注我们的大规模 n 和 d 统计情形；此时目标测度是对数凹但非强对数凹，这是一种贝叶斯后验分布中常见且颇具挑战性的情形。

在统计学方面，过去十年间，人们对朗之万蒙特卡洛的混合性质进行了大量研究，这主要受到机器学习方法的推动，例如 [11], 提出的指数加权聚合；该方法涉及对非对数凹且重尾的后验分布进行采样。Dalalyan 的首篇论文 [12] 在目标测度 ρ 强对数凹时，给出了朗之万蒙特卡洛达到 ε TV 界所需的计算代价，并以 d和 ρ 表示；同时提出了惩罚化朗之万蒙特卡洛，以克服目标分布仅为对数凹时缺乏强对数凹性的问题。自这篇开创性论文以来，相关文献大量涌现。其中， [16] 对离散化朗之万蒙特卡洛进行了细致研究， [14] 探讨了朗之万蒙特卡洛的动力学版本，而 [15] 则深入研究了非强对数凹情形下的惩罚化朗之万蒙特卡洛。在这些论文中，首先，针对缺乏强对数凹性的情况，研究者通过引入替代项和渐近趋于零的惩罚项来修改初始的朗之万蒙特卡洛算法。其次，这些论文假设算法每次迭代时都能获得对数后验的无噪声梯度；但这一假设可能不切实际，尤其是在 n 较大时。

随机 LMC（下文简称 SLMC）已引起多项研究的关注： [39] 首次提出了这一方法，并从数值角度描述了其在贝叶斯学习这一特定情形下的效率，而这正是我们的研究框架。还可以提及一些近期进展和相关工作： [13] 研究了 LMC 的一种带噪版本，并针对存在可能有偏噪声的强对数凹后验分布，推导出了采样策略的若干非渐近上界（以瓦瑟斯坦距离表示）。 [40] 近期的工作也与我们的研究相关: 作者开发了一套研究 SLMC 的方法，其基础主要是 Poincar´é 不等式；但他们处理 LMC 中所涉及的谱隙下界的方式似乎并不恰当。特别是，(随机)-LMC 中涉及的扩散所处温度极低，与 $1/n$成正比，这给非强对数凹框架下谱隙的大小带来了严重问题。在 [35], 中，作者针对优化目的推导了与我们框架相近的一些界，并指出其论文中记为 λ_* 的谱隙对他们引入的温度水平 $1/\beta$ 存在重要依赖。在某些一般情形下，他们得到的界非常悲观（见其 [35][第 4]节中的讨论）；最后，他们以迫切需要找到一些非平凡情形作结，在这些情形下或许可以推导出 λ_* 的更优下界。

事实上， [35][第 4]节的最后一条评论与广为人知的亚稳态现象有关：在低温下，许多可逆与不可逆马尔可夫半群的混合速率会因低温设置而显著恶化；在观测值数量较多的贝叶斯后验采样问题中，这种低温设置会被隐式诱导出来。在方差噪声处于 $O(\beta^{-1})$ 量级的情形下，对不变测度的大偏差原理的首次研究可追溯至 [18] ；作者在那里证明，过阻尼朗之万扩散的谱隙渐近形式为 $\exp(-I\beta)$（ [18][第 6]章），其中 I 是一个取决于吉布斯场势函数的显式常数。 [26], 进一步深入拓展了这一结果，从而开启了所谓模拟退火方法的首次精确分析（例如见 [24, 33]）。这些工作以及在随机设置下研究不可逆动力系统的更多近期成果（[22, 19]）表明，在亚稳态情形下，就谱隙的渐近行为而言，几乎没有什么值得期待的结果，间接地，就混合速率而言也是如此。因此，唯一可能带来合理结果的情形，是介于（几乎）平凡的强对数凹情形与多势阱亚稳态情形之间的中间情形。这正是弱对数凹情形的研究目的；该情形由一族 Kurdyka-Lojasiewicz 不等式 [28, 30] 描述，这些不等式已被证明对于优化理论 [5] 中的随机优化 [20] 或采样 [21]十分有效。我们还可参考近期的工作 [6] ；该工作在一个中间框架内推导了若干泛函不等式，在这一框架中，曲率 ρ 与其关键函数 α 相关，而后者控制着他们所研究的泛函不等式中涉及的常数。

综合考虑统计考量与局限性，本文旨在研究连续时间随机朗之万蒙特卡洛过程。下一段将对该过程作出精确定义，并介绍以实数 r 为参数的 Kurdyka-Lojasiewicz 框架，其中 r 的取值介于 0（强凸情形）与 1（拉普拉斯渐近尾部的极限情形）之间。我们将证明，为获得 ε 近似，最终的模拟时域量级为：

$$(d\log(n)^2)^{(1+r)^2}[\log^2(\varepsilon^{-1}) + n^2 d^{2(1+r)}\log^{4(1+r)}(n)]$$

采用参数为 $n\ d^{\frac{1}{2}}$（对数（n））$^{1+r}$。的泊松子采样

引言其余部分将依次介绍第 1.2 小节中的算法定义、第 1.3 小节中通过熵准则评估结果质量的方法，以及第 1.4 小节中的定量弱对数凹假设。最后，我们将在第 1.5 小节陈述主要结果。

1.2 连续时间演化

下面，我们简要回顾用于贝叶斯学习的连续时间 SLMC 算法，其离散化形式已在 [39] 中提出。为此，我们考虑一个借助函数 $(x,\theta) \longmapsto p_\theta(x)$ 构建的统计模型，其中 $\theta \in \mathbb{R}^d$ 编码统计模型的参数， x 表示取值于记作 $\mathcal{X}$ 的空间中的观测值。接着假设我们有 n 个独立同分布观测值，记作 $(\mathbf{X}_1,...,\mathbf{X}_n)$，其服从分布 p_{θ°。给定关于$\mathbb{R}^d$的先验分布 π_0，后验分布 μ_n 定义如下：

$$\mu_n(\theta) \propto \pi_0(\theta) \times \prod_{i=1}^n p_\theta(\mathbf{X}_i).$$

我们引入对数参数化，从而得到吉布斯形式：

$$U_x(\theta) = -[\log\pi_0(\theta) + n\log p_\theta(x)],$$

接着我们观察到：

$$\mu_n(\theta) \propto \exp\left(-\frac{1}{n}\sum_{i=1}^n U_{\mathbf{X}_i}(\theta)\right) = \exp\left(-U_{\nu_n}(\theta)\right),$$

其中， ν_n 表示经验分布，而 $X \sim \nu_n$时 $U_X(\theta)$的平均值为 U_{ν_n}：

$$\nu_n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{X}_i}(x) \quad \text{and} \quad U_{\nu_n}(\theta) = \mathbb{E}_{X\sim\nu_n}[U_X(\theta)].$$

标准朗之万蒙特卡洛方法依赖于随机微分方程的遍历行为:

$$\mathrm{d}\theta_t = -\nabla U_{\nu_n}(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \tag{1}$$

在一些温和假设下，该过程具有唯一的不变分布 μ_n。

SLMC 算法结合了利用随机微分方程进行采样的优势，以及漂移项均质化的优势；后者可写成关于 X 的期望，其中 X 根据 ν_n 在观测集合中均匀采样。其核心思想是，将 U_{ν_n} 中依赖于整个观测集合 $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$ 的期望，替换为单个观测值；该观测值在随机微分方程的整个演化过程中均匀随机选取，并根据马尔可夫指数时钟进行修改。由此，我们可以对该算法作如下明确的形式化定义。定义 $\left(\xi_j^{(n)}\right)_{j\geq 1}$ 为一个由指数随机变量构成的无限序列，其均值为 α_n^{-1}，具体值将在后文确定。

我们还考虑一个序列 $\left\{V_{(j}^{n}\right\}_{j>0}$，其中的随机变量是独立同分布的，且在 $\{1, 2, \ldots, n\}$ 上均匀分布。接着，我们将过程 $(X_t)_{t\geq 0}$ 定义为一个取值于 $\{1, 2, \ldots, n\}$ 的跳跃过程，使其满足:

$$X_t = \begin{cases} \mathbf{X}_{V_1^{(n)}}, & \text{if} \quad 0 \leq t < \xi_1^{(n)}, \\ \mathbf{X}_{V_j^{(n)}}, & \text{if} \quad \sum_{k=1}^{j-1} \xi_k^{(n)} \leq t < \sum_{k=1}^{j} \xi_k^{(n)}, \quad j > 1. \end{cases} \tag{2}$$

非正式地，$(X_t)_{t\geq 0}$ 应理解如下：该过程以强度为 α_n 的指数等待时间，均匀随机选取观测值 n X_1、.....、X_n 中的一个作为取值。我们考虑的随机朗之万过阻尼扩散由联合演化 $(\theta_t, X_t)_{t\geq 0}$ 给出，其定义如下:

$$\mathrm{d}\theta_t = -\nabla_\theta U_{X_t}(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad t > 0, \tag{3}$$

其中，$(B_t)_{t\geq 0}$ 是多元标准布朗运动。

算法 1：随机朗之万过阻尼算法

数据：$(\mathrm{X}_1$、.....、$\mathrm{X}_n)$ 独立同分布观测值，n_0 初始分布，π_0 先验分布

$t_0 = 0$

根据 n_0 生成 θ_0

对于 $k = 0, 1,$ .....执行

  从 $\{\mathrm{X}_1, \ldots, \mathrm{X}_n\}$ 中均匀选取 X_k

  按照均值为 α_n^{-1} 的指数分布生成 ξ_k

  $t_{k+1} = t_k + \xi_k$

  $\theta_{t_{k+1}} = \theta_{t_k} - \int_{t_k}^{t_{k+1}} \nabla_\theta U_{X_k}(\theta_s)\mathrm{d}s + \sqrt{2}B_{\xi_k}$

结束

返回 $\lim_{k\to\infty} \theta_{t_k}$

1.3 熵散度

为评估 SLMC 的长时间行为，我们引入一些与二元组 $(\theta_t, X_t)_{t\geq 0}$ 相关的记号。下文中，我们用 λ_d 表示 $\mathbb{R}^d$ 上的勒贝格测度。由 $\mathcal{L}$ 诱导的半群在 θ 坐标上是椭圆的，在 x 坐标上显然不可约且具有有限支撑，因此只要 $t > 0$，(θ_t, X_t) 的分布就关于测度 $\lambda_d \otimes \nu_n$ 绝对连续。

我们引入 m_t 这一记号，用于表示时刻 t 相对于 $\lambda_d \otimes \nu_n$ 的 (θ_t, X_t) 联合密度。同时，n_t 表示 θ_t 的边缘分布，而 $m_t(\cdot|\theta)$ 表示在给定 $\theta_t = \theta$ 的条件下 X_t 的条件分布。即:

$$\mathrm{Law}(\theta_t, X_t) = m_t, \quad n_t(\theta) = \sum_{i=1}^{n} m_t(\theta, \mathbf{X}_i), \quad m_t(x|\theta) = \frac{m_t(\theta, x)}{n_t(\theta)}, \tag{4}$$

对于 $\theta \in \mathbb{R}^d$ 和 $x \in \{X_1, \ldots\ldots, X_n\}$。

为了证明 SLMC 算法能够恢复正确的渐近行为，即当 $t \longrightarrow \infty$ 时， $n_t(\theta) \longrightarrow \mu_n$，我们考虑 n_t 相对于 μ_n 的相对熵（或 Kullback–Leibler 散度）。由于椭圆性，该相对熵定义良好，其定义如下：

$$J_t = Ent_{\mu_n}\left(\frac{n_t}{\mu_n}\right) = \int_{\mathbb{R}^d} \log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right) \mathrm{d}n_t(\theta). \tag{5}$$

J_t 衡量任意时刻 $t > 0$ 过程在时刻 t 的瞬时分布与该过程（θ_t， X_t）的（假定）不变分布 μ_n 之间的散度。也可以用 $\mathbb{L}^2$ 或 χ 平方距离来衡量这两个分布之间的差异，并借助泛函分析进行理论分析，但这需要对函数 U_{ν_n} 作出更强的假设。

同时，我们还引入一个加权 $\mathbb{L}^2$ 距离，用于衡量给定 $\theta_t = \theta$ 时 X_t 的条件分布与测度 ν_n 之间的差异。该距离记作 I_t，定义如下：

$$I_t = \int_{\mathbb{R}^d} \sum_{i=1}^{n} \left(\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)} - 1\right)^2 \nu_n(\mathbf{X}_i)\mathrm{d}n_t(\theta). \tag{6}$$

该量衡量在时间 t，给定 θ 时 x 的条件分布与 ν_n之间的平均接近程度（相对于 θ）。

1.4 主要假设

弱凸性 我们将在弱凸框架下研究 SLMC，即假设 U_{ν_n} 为凸函数，但不一定是强凸函数。近年来，SLMC 在机器学习领域受到广泛关注，其显式欧拉离散形式已在涉及泛函不等式的多种情形下得到深入研究。我们参考 [38] （一致对数 Sobolev 不等式）、 [35] （一致 Poincar´e 不等式；作者在其中对算法进行了 Wasserstein-2 分析）以及 [40] （一致 Poincar´e 不等式）。在这些研究中，泛函不等式对于分析 SLMC 的行为至关重要，并且这些不等式被作为假设，这是一项重要的假设。值得注意的是，Poincar´e 或对数 Sobolev 不等式并非无关紧要：为了使其对维度的依赖合理，它们通常要求凸性（例如参见 [4, 3]）；若要与维度无关，则甚至要求强凸性。否则，这些泛函不等式中涉及的常数会随着"温度"（在我们的情形下为 $n^{-1}(d\log^{2\beta}(n))^{-(1+r)}$ ）和维度（对我们而言为 d ）呈指数级恶化，如 [26] 所示。

在我们的工作中，我们选择借助 Kurdyka- Lojasiewicz 不等式 [28, 30], 来刻画强凸性的缺失。该不等式是优化领域中用于描述凸性与强凸性之间转变的标准工具，并能使界限更加明确。这一假设使我们能够观察熵如何随 KL 不等式中涉及的关键指数而变化。具体而言，它使我们能够理解强凸性缺失所产生的影响；在以往工作所假设的一致 Poincar´e 不等式或对数索博列夫不等式中，这种影响或多或少被隐藏了起来。我们按照 [20] 引入 KL 不等式的一种参数化形式。

为此，对于任意 V 二次可微函数，我们将 V 的海森矩阵的谱记为 $Sp(\nabla^2 V(\theta))$。此外，若 V 为凸函数，我们记：

$$\underline{\lambda}_{\nabla^2 V}(\theta) = \inf Sp(\nabla^2 V(\theta)).$$

假设 $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 若函数 $V\colon \mathbb{R}^d \to \mathbb{R}$ 满足 $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 条件，则称其满足：

a) V 是一个 $\mathcal{C}^2$-函数。

b) V 是一个凸函数，且 $\min_{\theta\in\mathbb{R}^d} V(\theta) = V(\theta^*) > 0$。

c) ∇V 是 L-Lipschitz 连续的。

d) 存在一些常数 $0 \le r < 1$ 和 $\mathfrak{c} > 0$，使得：

$$\mathfrak{c}V^{-r}(\theta) \le \underline{\lambda}_{\nabla^2 V(\theta)} \quad \forall \theta \in \mathbb{R}^d. \tag{7}$$

我们简要说明一下这个假设。

- 在 [21], 中，通过引入另一个指数q（与$\overline{\lambda}_{\nabla^2 V(\theta)} = \sup Sp(\nabla^2 V(\theta))$相关），采用了略有不同的参数化。作者还假设上界为$\overline{\lambda}_{\nabla^2 V(\theta)} \le \tilde{\mathfrak{c}} V^{-q}(\theta)$。这里，我们选择简化这一假设，并用利普希茨常数 L 给出的海森矩阵特征值粗略上界，即在最后一个不等式中，我们只使用 $\tilde{\mathfrak{c}} = L$ 和 $q = 0$。

- 我们将看到，若 $V(\theta) = (1 + \|\theta\|_2^2)^p$ 且 $p \in [1/2, 1]$，则 V满足 $\mathcal{H}_{\mathbf{rKL}}(\mathfrak{c}, L)$，其中$r = \frac{1-p}{p}$且 $\mathfrak{c} = 2p(1 - 2(1 - p))$；更多细节请参见 [21] 中的注释7。特别地， p越大， r越小，这意味着势函数 V的曲率更佳。

- 当 $r = q$ 时，我们得到全局标准 KL 不等式（见 [20, 5]）；当 $r = 1$ 时，它对应于极限拉普拉斯情形。
- 显然，情形 $r = 0$ 与强凸情形相关，此时函数的曲率有统一的下界 $\mathfrak{c}$。

因此，预期 SLMC 的复杂度会随着曲率的减弱而增加，也就是说，它是 r 的递增函数。

在第4节中，我们回顾引理15（见 [21]）中 KL 不等式的一些重要推论。特别是，任何满足 $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 的函数，其增长都可由到其极小化器的距离的某个正幂从下方和上方界定。

如果不等式 (7) 对常数 $\mathfrak{c}$ 成立，那么它对小于 $\mathfrak{c}$ 的所有正值也成立。因此，在第5节中，我们假设 $\mathfrak{c} \le \left(\frac{8L}{1+r}\right)$ 。

先验分布的假设 π_0 我们将在下文陈述一个“总体” $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$假设的重要推论，不过在此之前，先对 π_0陈述一些较弱的假设。

假设 $\mathcal{H}_{\pi_0}(\ell_0)$ π_0 是一个对数凹的 $\mathcal{C}^2$-函数，满足 $\min_{\theta \in \mathbb{R}^d} -\log \pi_0(\theta) > 0$ 且 $\theta \mapsto \nabla \log \pi_0(\theta)$是 ℓ_0-利普希茨连续的。

由于先验分布由用户选择，我们的 $\mathcal{H}_{\pi_0}(\ell_0)$假设并不苛刻；一些典型例子满足这些条件，例如高斯分布、威布尔分布和伽马分布（它们的形状参数均大于 1）、冈贝尔分布等。

命题 1.1。我们假设 $\mathcal{H}_{\pi_0}(\ell_0)$，并且存在$(\mathfrak{c}, r)$，使得对于任意 x： $\theta \longmapsto -\log p_\theta(x)$满足 $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，那么 U_{ν_n} 满足$\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r}, nL + \ell_0\right)$；特别地，对于任意 $\mathbf{X}_i$, $U_{\mathbf{X}_i}$ 满足 $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r}, nL + \ell_0\right)$。

我们引入记号 $a \lesssim_{uc} b$ ($a \gtrsim_{uc} b$)，表示 $a \le cb$ ($a \ge cb$)，其中 c 是一个通用常数，即一个与 n 和 d 无关的正数。

我们假设函数 $U_{\mathbf{X}_i}$ 的极小点包含在一个球内，其半径取决于 n 和 d。此外，我们认为 $\min_{\theta \in \mathbb{R}^d} U_{\mathbf{X}_i}$ 的阶至多为 d。

假设 $\mathcal{H}_{\min}$ ： 存在 $\beta \ge 0$ ，使得：

$$\max_i \| \arg\min U_{\mathbf{X}_i} \|_2 \lesssim_{uc} \sqrt{d} \log^\beta(n) \quad \text{and} \quad \max_i \min_{\theta \in \mathbb{R}^d} U_{\mathbf{X}_i}(\theta) \lesssim_{uc} d.$$

假设 $\mathcal{H}_{\min}$ 并不苛刻。在 $d = 1$维度下，对于许多集中的独立同分布样本（X_i）$_{1 \le i \le n}$ ，只要其具有适当的类次高斯行为，$\min$ U_{X_i} 的拉普拉斯变换就满足如下上界：

$$\mathbb{E}[\exp(\lambda \min U_{X_i})] \le \exp(\sigma^2 \lambda^k), \quad \forall \lambda > 0.$$

前面的上界意味着，在这种情况下， $\mathcal{H}_{\min}$ 中涉及的 β 由 $\beta = \frac{k-1}{k}$ 给出。特别地，当 $k = 2$时，我们恢复了 $\beta = 1/2$ 的情形。对于更高的维度，可以利用 $\|X\|_2^2 \le d\max_{1\le j\le d}(X^j)^2$, 来扩展该结果，其中 X^j 是 X的第 j个分量。我们应牢记上述讨论的结论：即使 $\mathcal{H}_{\min}$ 对 $\beta > 0$的任意取值都有定义（且有意义），它通常也只对 $\beta \le 1$成立。

这一 $\mathcal{H}_{\min}$ 假设与 $\mathcal{H}_{\pi_0}(\ell_0)$一起，使极小化器和 U_{ν_n}的最小值表现出几乎相似的行为。详情见命题 4.4。

1.5 长时间熵收敛

对于任意时刻 $t \ge 0$ ，我们引入 Law(θ_t) 关于 μ_n的密度，其定义为 :

$$f_t(\theta) = \frac{n_t(\theta)}{\mu_n(\theta)},$$

并且 n_0 的选取满足 $\|f_0\|_\infty < +\infty$。以下假设保证了这一结果，证明见命题 3.5。

假设 $\mathcal{H}_{\mathbf{n_0}}(\mathrm{L}\ \ell_0)$ 存在一个正数常数 σ^2，使得 $n_0 = \mathcal{N}(0, \sigma^2 I_d)$。此外，存在两个通用常数 c_1 和 c_2 ，使得 $0 < c_1 \le c_2 < 1$ 并且

$$\frac{c_1}{nL + \ell_0} \le \sigma^2 \le \frac{c_2}{nL + \ell_0}.$$

此外，在命题 3.5 中，作为 $\|f_0\|_\infty$有界性的直接推论，我们得到 $J_0 \lesssim_{uc} nd^{1+r}\log^{2\beta(1+r)}(n) + d\log\left(\frac{d}{n}\right)$。

下一个结果通过熵的递减来评估混合性质，因此说明了 n_t 趋向正确测度 μ_n的收敛性。

定理 1.1. 假设 $\mathcal{H}_{\pi_0}(\ell_0)$、 $\mathcal{H}_{\min}$, $\mathcal{H}_{\mathbf{n_0}}(\mathrm{L}\ \ell_0)$，且每个 $\theta \mapsto -\log p_\theta(\mathrm{X}_i)$ 都满足 $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，则

- U_{ν_n} 满足常数为 $C_P(\mu_n)$ 的 Poincar´e 不等式，统称为 C_P。
- *Define* $c_{n,d} := n^4\left(d\log^{2\beta}(n)\right)^{1+r}$ *and* $O_{n,d} := \left(\frac{C_1 d}{n}\right)^{\frac{dr}{2}}\exp\left(C_2 n\left(d\log^{2\beta}(n)\right)^{1+r}\right)$, *where* C_1 *and* C_2 *are universal constants, then for any* $t > 0$*:*

$$J_t \lesssim_{uc} \left(J_0 + \frac{c_{n,d}}{\alpha_n}\left[1 + \left(\frac{C_P}{\alpha_n} + \sqrt{C_P}\right)e^{\frac{\sqrt{C_P}}{\sqrt{a}} + \frac{C_P}{3\alpha_n}}\right] + O_{n,d}\right)(1+t)^{1/4}e^{-\frac{\sqrt{C_p}}{\sqrt{a}}(\sqrt{1+t}-1)}. \quad (8)$$

- 对于任意 $\varepsilon > 0$，若 $\alpha_n = \frac{1}{n\ d\ \ ^{2\beta}(\log(n))^{1+r}}$，则:

$$t \gtrsim_{uc} \left(d\log^{2\beta}(n)\right)^{(1+r)^2}\left[\log^2(\varepsilon^{-1}) + n^2\left(d\log^{2\beta}(n)\right)^{2(1+r)} + d^2\log^2 d\right] \Longrightarrow J_t \le \varepsilon.$$

若将 t_ε 记为使 $J_{t_\varepsilon} \le \varepsilon$ 成立的最小值，则选择 $\alpha_n = \frac{1}{n\ d\log^{2\beta}\ ((n))^{1+r}}$可保证过程 $(X_t)_{0\le t\le t_\varepsilon}$ 的平均跳跃次数 $\alpha_n t_\varepsilon$ 达到最小。

为证明主要结果，我们首先在第2节介绍与马尔可夫半群相关的经典工具，熟悉该主题的读者可以跳过本节。在第3节中，我们证明主要结果。第4节和第5节分别 посвящены $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 假设和 U_{ν_n}，以及马尔可夫动力学的技术性结果。

2 马尔可夫工具

借助鞅问题（下文简称 MP），可以直接验证 $(\theta_t, X_t)_{t\geq 0}$ 的联合演化存在，并且具有弱唯一性（在分布意义下）。为此，我们首先定义算子 $\mathcal{L}$，该算子作用于任意函数 $f \in \mathcal{C}^2(\mathbb{R}^d \times \mathcal{X})$，具体如下：

$$\mathcal{L}f(\theta, x) = \underbrace{-\langle \nabla_\theta U_x(\theta), \nabla_\theta f(\theta, x)\rangle + \Delta_\theta f(\theta, x)}_{:=\mathcal{L}_1 f(\theta,x)} + \underbrace{\frac{\alpha_n}{n}\sum_{i=1}^{n}[f(\theta, \mathbf{X}_i) - f(\theta, x)]}_{\mathcal{L}_2 f(\theta,x)}, \tag{9}$$

对所有 $(\theta, x) \in \mathbb{R}^d \times \mathcal{X}$。

算子 $\mathcal{L}$ 分为两项，$\mathcal{L}_1$ 作用于分量 θ，并与扩散部分相关；而 $\mathcal{L}_2$ 是作用于 x 分量的跳跃算子。由于观测值的数量有限 $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$，我们可以应用 [17] 第4章第4节和第5节的结果，并推导出以下结论：

命题 2.1. 假设对于任意 $x \in \mathcal{X}$，U_x 是 $\mathcal{C}^2(\mathbb{R}^d)$，且 $\nabla_\theta U_x$ 是 L_x-Lipschitz，那么对于 $\mathbb{R}^d \times \mathcal{X}$上的任意初始分布 ν，鞅问题 $(\mathcal{L}, \nu)$ 是适定的。

相应的（弱）唯一过程 $(\theta_t, X_t)_{t\geq 0}$ 是与半群 $\mathcal{L}$相关的费勒马尔可夫过程。特别地，θ 分量满足随机微分方程 (3)。

若将 $\mathcal{L}^\star$ 记为 $\mathbb{L}^2(\mathbb{R}^d) \times \nu_n$ 中 $\mathcal{L}$ 的伴随算子，则由后向柯尔莫哥洛夫方程可得：

$$\partial_t m_t(\theta, x) = \mathcal{L}^\star m_t(\theta, x). \tag{10}$$

利用半群 $\mathcal{L}$ 在坐标 θ 上的椭圆性，我们可以运用 [25] 的结果，并推得对于任意 $t > 0$，$n_t \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R})$，而不可约性可推出 $\forall t \geq 0, n_t > 0$。我们将在命题 3.5 中证明一些充分条件，以推出 $\|f_0\|_\infty = \| {}^{n_0(\theta)}\mu_n(\theta) \|_\infty < +\infty$。最大值原理的一个重要且标准的推论如下：若 $\|f_0\|_\infty \leq M$，则

$$\forall t \geq 0, \qquad \|f_t\|_\infty \leq M.$$

我们将该结果的细节留待命题 3.5 讨论，因为它们与我们的分析并非直接相关，而且相当技术性。

借助这个主方程，我们可以计算半群作用于 θ 的某个时变函数时的导数。为此，我们引入两个关键算子。第一个描述了在时刻 t，在 X_t 的平均效应下，对坐标 θ 的无穷小作用，其中 X_t 将 $\forall f \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ 映射为：

$$\mathcal{G}_t f(\theta) = -\sum_{i=1}^{n}\langle \nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\mathbf{X}_i|\theta) + \Delta_\theta f(\theta). \tag{11}$$

第二种情况与第一种情况非常接近，不同之处在于，X_t 的平均效应被目标理想分布 ν_n取代。由此得到定义 $\forall f \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$：

$$\mathcal{G} f(\theta) = -\sum_{i=1}^{n}\langle \nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle \nu_n(\mathbf{X}_i) + \Delta_\theta f(\theta) = -\langle \nabla_\theta f(\theta), \nabla_\theta U_{\nu_n}(\theta)\rangle + \Delta_\theta f(\theta). \tag{12}$$

T其导数见下一个结果，证明推迟到附录中给出。

引理 2.1. 设 h_t 是一个二次可微函数，且其在 $\mathbb{R}^d$上的一阶和二阶导数均一致有界，则对于 $t > 0$：

$$\partial_t \left\{ \int_{\mathbb{R}^d} h_t(\theta) \mathrm{d}n_t(\theta) \right\} = \int_{\mathbb{R}^d} \partial_t\{h_t(\theta)\} \mathrm{d}n_t(\theta) + \int_{\mathbb{R}^d} \mathcal{G}_t h_t(\theta) \mathrm{d}n_t(\theta), \tag{13}$$

wh其中，$\mathcal{G}_t$ 是 X_t的平均效应下的扩散算子，定义见公式（11 ）.

3 主要结果的证明

3.1 熵的演化 J_t

熵满足以下微分不等式。

命题 3.1. 假设 $\mathcal{H}_{\min}$, $\mathcal{H}_{\pi_0}(\ell_0)$，且对于每个 $\mathbf{X}_i$， $\theta \to -\log p_\theta(\mathbf{X}_i)$ 满足 $\mathcal{H}^r_{\mathbf{KL}}(\mathfrak{c}, L)$，则存在一个“通用”常数 C（与 n 和 d 无关），使得 $\forall t > 0$:

$$\partial_t\{J_t\} \leq -\int_{\mathbb{R}^d} \left\| \nabla_\theta \left(\sqrt{\frac{n_t(\theta)}{\mu_n(\theta)}} \right) \right\|_2^2 \mathrm{d}\mu_n(\theta) + C I_t^{\frac{1}{3}} n^{\frac{11}{3}} \left(d \log^{2\beta}(n) \right)^{1+r}.$$

证明. 我们将使用一个标准的预备估计，该估计可由椭圆扩散的[29] 公式（3.14）推导得出，以便将引理 2.1 应用于 $f_t = \log(n_t \mu_n^{-1})$。由公式（13），可得:

$$\partial_t\{J_t\} = \int_{\mathbb{R}^d} \partial_t \left\{ \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \right\} \mathrm{d}n_t(\theta) + \int_{\mathbb{R}^d} \mathcal{G}_t \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \mathrm{d}n_t(\theta),$$

第一项为零，因为:

$$\begin{aligned} \int_{\mathbb{R}^d} \partial_t \left\{ \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \right\} \mathrm{d}n_t(\theta) &= \int_{\mathbb{R}^d} \frac{\partial_t\{n_t(\theta)\}}{n_t(\theta)} \mathrm{d}n_t(\theta) \\ &= \int_{\mathbb{R}^d} \partial_t \left\{ n_t(\theta) \right\} \mathrm{d}\theta \\ &= \partial_t \left(\int_{\mathbb{R}^d} \mathrm{d}n_t(\theta) \right) \\ &= 0. \end{aligned}$$

于是，导数可化简为第二项，从而得到:

$$\begin{aligned} \partial_t\{J_t\} &= \int_{\mathbb{R}^d} \mathcal{G}_t \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \mathrm{d}n_t(\theta), \\ &= \underbrace{\int_{\mathbb{R}^d} \mathcal{G} \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \mathrm{d}n_t(\theta)}_{J_{1,t}} + \underbrace{\int_{\mathbb{R}^d} (\mathcal{G}_t - \mathcal{G}) \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \mathrm{d}n_t(\theta)}_{J_{2,t}}. \end{aligned} \tag{14}$$

我们分别研究这两项 $J_{1,t}$ 和 $J_{2,t}$ 。

- 研究 $J_{1,t}$。由于 $\mathcal{G}$ 是一个扩散算子，而 μ_n 是与 $\mathcal{G}$相关的不变测度，因此我们可以利用 $J_{1,t}$ 与狄利克雷形式之间的经典联系（见 [3]）:

$$\begin{aligned} \int_{\mathbb{R}^d} \mathcal{G} \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \mathrm{d}n_t(\theta) &= \int_{\mathbb{R}^d} \frac{n_t(\theta)}{\mu_n(\theta)} \, \mathcal{G} \log\left(\frac{n_t(\theta)}{\mu_n(\theta)} \right) \mathrm{d}\mu_n(\theta) \\ &= -4 \int_{\mathbb{R}^d} \left\| \nabla_\theta \left(\sqrt{\frac{n_t(\theta)}{\mu_n(\theta)}} \right) \right\|_2^2 \mathrm{d}\mu_n(\theta). \end{aligned} \tag{15}$$

- J_2 的研究, t。对于任意二阶可微函数 f，我们使用 $\mathcal{G}$ 与 $\mathcal{G}_t$ 之间的差:

$$\begin{aligned} (\mathcal{G}_t - \mathcal{G}) f(\theta) &= -\sum_{i=1}^n \langle \nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta) \rangle \left[m_t(\mathbf{X}_i|\theta) - \nu_n(\mathbf{X}_i) \right] \\ &= -\sum_{i=1}^n \langle \nabla_\theta f(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta) \rangle \left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)} - 1 \right] \nu_n(\mathbf{X}_i). \end{aligned}$$

那么，项 $J_{2,\ t}$ 可计算如下：

$$\begin{aligned}|J_{2,t}| &= \left|\int_{\mathbb{R}^d}(\mathcal{G}_t-\mathcal{G})\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\mathrm{d}n_t(\theta)\right| \\ &= \left|\int_{\mathbb{R}^d}\sum_{i=1}^n\langle\nabla_\theta\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right),\nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]\nu_n(\mathbf{X}_i)\ \mathrm{d}n_t(\theta)\right|.\end{aligned}$$

在第一行中对测度 $\nu_n(\mathbf{X}_i)\times\mathrm{d}n_t(\theta)$ 使用柯西-施瓦茨不等式，在第二行中使用 $2ab\le a^2+b^2$，并在第三行中使用 $\nabla\log f=2\nabla\log\sqrt{f}=2\frac{\nabla\sqrt{f}}{\sqrt{f}}$，我们得到：

$$\begin{aligned}|J_{2,t}| &\le \left(\int_{\mathbb{R}^d}\left\|\nabla_\theta\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\right\|_2^2\mathrm{d}n_t(\theta)\right)^{\frac12}\left(\int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^2\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\nu_n(\mathbf{X}_i)\ \mathrm{d}n_t(\theta)\right)^{\frac12} \\ &\le \frac34\int_{\mathbb{R}^d}\left\|\nabla_\theta\log\left(\frac{n_t(\theta)}{\mu_n(\theta)}\right)\right\|_2^2\mathrm{d}n_t(\theta)+\frac13\int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^2\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\nu_n(\mathbf{X}_i)\ \mathrm{d}n_t(\theta) \\ &\le 3\int_{\mathbb{R}^d}\left\|\nabla_\theta\left(\sqrt{\frac{n_t(\theta)}{\mu_n(\theta)}}\right)\right\|_2^2\mathrm{d}\mu_n(\theta)+\frac13\int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^2\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\nu_n(\mathbf{X}_i)\ \mathrm{d}n_t(\theta).\end{aligned}$$

使用方程（15）和上一行可得：

$$\partial_t\{J_t\} \le -\int_{\mathbb{R}^d}\left\|\nabla_\theta\left(\sqrt{\frac{n_t(\theta)}{\mu_n(\theta)}}\right)\right\|_2^2\mathrm{d}\mu_n(\theta)+\underbrace{\frac13\int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^2\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\nu_n(\mathbf{X}_i)\ \mathrm{d}n_t(\theta)}_{:=\Delta_t}, \quad (16)$$

接下来我们关注右侧的第二项。为此，我们考虑一个非负函数 $g(t)$，稍后将对其进行固定，并将 Δ_t 拆分为两项：

$$\begin{aligned}\Delta_t &= \int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^2\left(\mathbb{I}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2\le g(t)}+\mathbb{I}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2>g(t)}\right)\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\nu_n(\mathbf{X}_i)\ \mathrm{d}n_t(\theta) \\ &\le g^2(t)I_t+\int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^2\mathbb{I}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2>g(t)}\left[\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right]^2\nu_n(\mathbf{X}_i)\ \mathrm{d}n_t(\theta),\end{aligned}$$

其中，I_t 已在方程（6）中引入，用于衡量 $m_t(\mathbf{X}_i|\theta)$ 与 ν_n 的接近程度。最后，对于最后一项，我们注意到 $0\le m_t(\mathbf{X}_i|\theta)\le 1$ 以及 $\left|\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right|=n\left|m_t(\mathbf{X}_i|\theta)-\frac1n\right|\le n$，这意味着：

$$\Delta_t\le g^2(t)I_t+n^2\underbrace{\frac1n\int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^2\mathbb{I}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2>g(t)}\mathrm{d}n_t(\theta)}_{:=\tilde\Delta_t}. \quad (17)$$

由柯西不等式可得：

$$\begin{aligned}\tilde\Delta_t &\le \left(\frac1n\int_{\mathbb{R}^d}\sum_{i=1}^n\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\right\|_2^4\mathrm{d}n_t(\theta)\right)^{\frac12}\left(\frac1n\int_{\mathbb{R}^d}\sum_{i=1}^n\mathbb{I}_{\|\nabla_\theta U_{\mathbf{X}_i}(\theta)\|_2>g(t)}\mathrm{d}n_t(\theta)\right)^{\frac12} \\ &= \left(\frac1n\sum_{i=1}^n\mathbb{E}\left[\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta_t)\right\|_2^4\right]\right)^{\frac12}\left(\frac1n\sum_{i=1}^n\mathbb{P}\left(\left\|\nabla_\theta U_{\mathbf{X}_i}(\theta_t)\right\|_2>g(t)\right)\right)^{\frac12}. \quad (18)\end{aligned}$$

接下来我们使用命题 4.1，得到：

$$\begin{aligned}\tilde\Delta_t &\le \left(\frac1n\sum_{i=1}^n\mathbb{E}\left[\left(2(nL+\ell_0)U_{\mathbf{X}_i}^2(\theta_t)\right)\right]\right)^{\frac12}\left(\frac1n\sum_{i=1}^n\mathbb{P}\left(2(nL+\ell_0)U_{\mathbf{X}_i}(\theta_t)>g^2(t)\right)\right)^{\frac12} \\ &\le 2(nL+\ell_0)\left(n\mathbb{E}[U_{\nu_n}^2(\theta_t)]\right)^{\frac12}\left(\frac1n\sum_{i=1}^n\frac{2(nL+\ell_0)}{g^2(t)}\mathbb{E}\left[U_{\mathbf{X}_i}(\theta_t)\right]\right)^{\frac12} \\ &\le [2(nL+\ell_0)]^{\frac32}n^{\frac12}\frac{\mathbb{E}\left[U_{\nu_n}^2(\theta_t)\right]^{\frac12}\mathbb{E}\left[U_{\nu_n}(\theta_t)\right]^{\frac12}}{g(t)},\end{aligned}$$

其中我们使用了马尔可夫不等式以及关系 $\|.\|_2 \leq \|.\|_1$ in $\mathbb{R}^n$。我们将命题 5.1 应用于 $\alpha = 2$ 和 $\alpha = 1$，得到存在一个常数 $C > 0$ （其值可能随行而变化），使得:

$$\tilde{\Delta}_t \quad \leq \quad C\frac{n^{\frac{7}{2}}\left(d\log^{2\beta}(n)\right)^{\frac{3(1+r)}{2}}}{g(t)}.$$

我们将这个最后的界代入（17），从而得到:

$$\Delta_t \leq g^2(t)I_t + C\frac{n^{\frac{11}{2}}\left(d\log^{2\beta}(n)\right)^{\frac{3(1+r)}{2}}}{g(t)}.$$

对 $g(t)$ 优化这个最后的界，可得上界:

$$\Delta_t \leq CI_t^{\frac{1}{3}}n^{\frac{11}{3}}\left(d\log^{2\beta}(n)\right)^{1+r}, \quad \forall t \geq 0.$$

□

3.2 加权 $\mathbb{L}^2$ 距离 I_t 的演化

命题 3.1 中涉及的量 I_t 衡量了 ν_n 条件分布 of$X_t|\theta_t$ 与某分布的接近程度。为研究 I_t，我们首先指出，它可以改写为一种更简单的形式。

$$\begin{aligned}
I_t &= \int_{\mathbb{R}^d}\sum_{i=1}^n\left(\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right)^2\nu_n(\mathbf{X}_i)\,\mathrm{d}n_t(\theta)\\
&= \int_{\mathbb{R}^d}\sum_{i=1}^n\left(\frac{m_t^2(\mathbf{X}_i|\theta)}{\nu_n^2(\mathbf{X}_i)}-2\frac{m_t(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}+1\right)\nu_n(\mathbf{X}_i)\,\mathrm{d}n_t(\theta)\\
&= \int_{\mathbb{R}^d}\sum_{i=1}^n\left(\frac{m_t^2(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-2m_t(\mathbf{X}_i|\theta)+\nu_n(\mathbf{X}_i)\right)\mathrm{d}n_t(\theta)\\
&= \int_{\mathbb{R}^d}\left(\sum_{i=1}^n\frac{m_t^2(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}-1\right)\mathrm{d}n_t(\theta)\\
&= \int_{\mathbb{R}^d}\sum_{i=1}^n\frac{m_t^2(\mathbf{X}_i|\theta)}{\nu_n(\mathbf{X}_i)}\mathrm{d}n_t(\theta)-1.
\end{aligned}$$

利用 $m_t(\mathbf{X}_i|\theta)n_t(\theta) = m_t(\theta,\mathbf{X}_i)$，以及 $\nu_n(\mathbf{X}_i) = \frac{1}{n}$ 对于 $i = 1, 2, \ldots, n$，我们得到:

$$I_t = n\int_{\mathbb{R}^d}\sum_{i=1}^n\frac{m_t^2(\theta,\mathbf{X}_i)}{n_t(\theta)}\mathrm{d}\theta - 1. \tag{19}$$

接下来的命题将评估当 $t \longrightarrow +\infty$ 时， I_t 趋近于 0 的速度。

命题 3.2。对于任意 $t \geq 0$:

$$I_t \leq I_0e^{-2\alpha_n t} \leq (n-1)e^{-2\alpha_n t}. \tag{20}$$

证明。我们的出发点是方程（19）。我们计算其关于 t 的导数:

$$\begin{aligned}
\partial_t\{I_t\} &= 2n\int_{\mathbb{R}^d}\sum_{i=1}^n\frac{m_t(\theta,\mathbf{X}_i)}{n_t(\theta)}\partial_t m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta - n\int_{\mathbb{R}^d}\sum_{i=1}^n\frac{m_t^2(\theta,\mathbf{X}_i)}{n_t^2(\theta)}\partial_t n_t(\theta)\mathrm{d}\theta\\
&= 2n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t(\mathbf{X}_i|\theta)\partial_t m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta - n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t n_t(\theta)\mathrm{d}\theta.
\end{aligned}$$

在第一行使用柯尔莫哥洛夫后向方程，并在第二行使用 $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$，其中 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 定义于公式（9），我们得到：

$$\begin{aligned}
\partial_t\{I_t\} &= 2n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}m_t(\mathbf{X}_i|\theta)\ m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta - n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t n_t(\theta)\mathrm{d}\theta \\
&= \underbrace{2n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}_1 m_t(\mathbf{X}_i|\theta)\ m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta}_{:=I_{3,t}} + \underbrace{2n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}_2 m_t(\mathbf{X}_i|\theta)\ m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta}_{:=I_{1,t}} \\
&\quad \underbrace{-n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t n_t(\theta)\mathrm{d}\theta}_{:=I_{2,t}}. \qquad (21)
\end{aligned}$$

然后，$\partial_t\{I_t\}$ 可拆分为三项，分别进行研究。

- 研究$I_{1,t}$。我们观察到：

$$\mathcal{L}_2 m_t(\mathbf{X}_i|\theta) = \frac{\alpha_n}{n}\sum_{j=1}^n [m_t(\mathbf{X}_j|\theta) - m_t(\mathbf{X}_i|\theta)] = \frac{\alpha_n}{n} - \alpha_n\ m_t(\mathbf{X}_i|\theta). \qquad (22)$$

然后，我们将这个最后的方程代入 $I_1(t)$的定义，得到：

$$\begin{aligned}
I_{1,t} &= 2n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}_2 m_t(\mathbf{X}_i|\theta)\ m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta \\
&= 2\alpha_n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta - 2\alpha_n n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t(\mathbf{X}_i|\theta)m_t(\theta,\mathbf{X}_i)\mathrm{d}\theta \\
&= 2\alpha_n - 2\alpha_n n\int_{\mathbb{R}^d}\sum_{i=1}^n \frac{m_t^2(\theta,\mathbf{X}_i)}{n_t(\theta)}\mathrm{d}\theta \\
&= -2\alpha_n I_t. \qquad (23)
\end{aligned}$$

- 研究$I_{2,t}$。根据 n_t的定义，我们得到：

$$\begin{aligned}
I_{2,t} &= -n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t n_t(\theta)\mathrm{d}\theta \\
&= -n\int_{\mathbb{R}^d}\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t\left(\sum_{j=1}^n m_t(\theta,\mathbf{X}_j)\right)\mathrm{d}\theta \\
&= -n\int_{\mathbb{R}^d}\sum_{j=1}^n\sum_{i=1}^n m_t^2(\mathbf{X}_i|\theta)\partial_t m_t(\theta,\mathbf{X}_j)\mathrm{d}\theta \\
&= -n\int_{\mathbb{R}^d}\sum_{j=1}^n\left(\sum_{i=1}^n \mathcal{L}m_t^2(\mathbf{X}_i|\theta)\right) m_t(\theta,\mathbf{X}_j)\mathrm{d}\theta \\
&= -n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta).
\end{aligned}$$

其中，第四行用到了柯尔莫哥洛夫后向方程，最后一行再次用到了n_t 的定义。同样，分解 $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ 可得：

$$I_{2,t} = -n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta) - n\int_{\mathbb{R}^d}\sum_{i=1}^n \mathcal{L}_2 m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta).$$

我们重复一些与公式（22）中类似的计算，以研究由 $\mathcal{L}_2$ 诱导的跳跃分量对 m_t^2的作用。由此可得：

$$\mathcal{L}_2 m_t^2(\mathbf{X}_i|\theta) = \frac{\alpha_n}{n}\sum_{k=1}^{n}[m_t^2(\mathbf{X}_k|\theta) - m_t^2(\mathbf{X}_i|\theta)] = \frac{\alpha_n}{n}\sum_{k=1}^{n} m_t^2(\mathbf{X}_k|\theta) - \alpha_n\ m_t^2(\mathbf{X}_i|\theta).$$

利用最后这个方程，我们得到：

$$\begin{aligned}
I_{2,t} &= -n\int_{\mathbb{R}^d}\sum_{i=1}^{n}\mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta) - \alpha_n\int_{\mathbb{R}^d}\sum_{i=1}^{n}\sum_{k=1}^{n} m_t^2(\mathbf{X}_k|\theta)\ \mathrm{d}n_t(\theta)\\
&\quad +\alpha_n n\int_{\mathbb{R}^d}\sum_{i=1}^{n} m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta)\\
&= -n\int_{\mathbb{R}^d}\sum_{i=1}^{n}\mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta) - \alpha_n n\int_{\mathbb{R}^d}\sum_{k=1}^{n} m_t^2(\mathbf{X}_k|\theta)\ \mathrm{d}n_t(\theta)\\
&\quad +\alpha_n n\int_{\mathbb{R}^d}\sum_{i=1}^{n} m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta)\\
&= -n\int_{\mathbb{R}^d}\sum_{i=1}^{n}\mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta)\ \mathrm{d}n_t(\theta). \qquad (24)
\end{aligned}$$

- 研究 $I_{2,\ t} + I_{3,\ t}$。我们注意到，这个求和只涉及 $\mathcal{L}_1$（见公式（9））。我们先计算：

$$\mathcal{L}_1 m_t(\mathbf{X}_i|\theta) = -\langle\nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t(\mathbf{X}_i|\theta)\rangle + \Delta_\theta m_t(\mathbf{X}_i|\theta),$$

类似地：

$$\begin{aligned}
\mathcal{L}_1 m_t^2(\mathbf{X}_i|\theta) &= -\langle\nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t^2(\mathbf{X}_i|\theta),\rangle + \Delta_\theta m_t^2(\mathbf{X}_i|\theta)\\
&= -2m_t(\mathbf{X}_i|\theta)\langle\nabla_\theta U_{\mathbf{X}_i}(\theta), \nabla_\theta m_t(\mathbf{X}_i|\theta)\rangle + 2\|\nabla_\theta m_t(\mathbf{X}_i|\theta)\|_2^2 + 2m_t(\mathbf{X}_i|\theta)\Delta_\theta m_t(\mathbf{X}_i|\theta).
\end{aligned}$$

将这两个方程代入 $I_{2,\ t} + I_{3,\ t}$和 $m_t(\mathbf{X}_i|\theta)n_t(\theta) = m_t(\theta, \mathbf{X}_i)$，得到：

$$\begin{aligned}
\frac{I_{2,t} + I_{3,t}}{n} &= 2\int_{\mathbb{R}^d}\sum_{i=1}^{n}\langle\nabla_\theta m_t(\mathbf{X}_i|\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta\\
&\quad - 2\int_{\mathbb{R}^d}\sum_{i=1}^{n}\|\nabla_\theta m_t(\mathbf{X}_i|\theta)\|_2^2\ n_t(\theta)\mathrm{d}\theta - 2\int_{\mathbb{R}^d}\sum_{i=1}^{n}\Delta_\theta m_t(\mathbf{X}_i|\theta)\ m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta\\
&\quad - 2\int_{\mathbb{R}^d}\sum_{i=1}^{n}\langle\nabla_\theta m_t(\mathbf{X}_i|\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta + 2\int_{\mathbb{R}^d}\sum_{i=1}^{n}\Delta_\theta m_t(\mathbf{X}_i|\theta)\ m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta\\
&= -\int_{\mathbb{R}^d}\sum_{i=1}^{n}\|\nabla_\theta m_t(\mathbf{X}_i|\theta)\|_2^2\ \mathrm{d}n_t(\theta) \le 0.
\end{aligned}$$

将最后这个不等式与（23）一并代入公式（21），得到：

$$\partial_t\{I_t\} \le -2\alpha_n I_t.$$

注意到 $I_0 \le n-1$，直接应用格朗沃尔引理即可得出结论。 □

3.3 泛函（弱）对数-Sobolev 不等式

3.3.1 泛函不等式相关工作

命题 3.1 和命题 3.2 的一个直接推论是，相对熵 J_t满足以下微分不等式：

$$\partial_t\{J_t\} \le -\int_{\mathbb{R}^d}\left\|\nabla_\theta\left(\sqrt{\frac{n_t(\theta)}{\mu_n(\theta)}}\right)\right\|_2^2\mathrm{d}\mu_n(\theta) + c_{n,d}e^{-\frac{2\alpha_n}{3}t}, \qquad (25)$$

其中$c_{n,\ d}$ 定义为:

$$c_{n,d} \lesssim_{uc} n^4 \left(d \log^{2\beta}(n)\right)^{1+r}. \tag{26}$$

在这一阶段，我们应当注意到，一种标准方法是寻找一个泛函不等式，将如下定义的关键狄利克雷形式 $\mathcal{E}(f)$ 与其他量联系起来:

$$\mathcal{E}(f) = \int_{\mathbb{R}^d} \|\nabla_\theta f(\theta)\|_2^2 \mathrm{d}\mu_n(\theta), \tag{27}$$

即相对于 μ_n 的熵本身 $Ent_{\mu_n}(f^2)$。这些方法基于 [23] 的早期工作，他们在其中引入了对数索об列夫不等式（简称 LSI）。LSI 对指数遍历性的影响随后成为一个广泛的研究领域，关于这一主题的综述请参见 [3] 。确保 LSI 成立的一个常见充分条件是目标测度具有对数强凸性（另见 [2] 等文献）；大量研究聚焦于这些泛函不等式、半群的遍历性、传输不等式与李雅普诺夫条件之间的联系。相关文献请参见 [8, 1] （这两篇工作远非穷尽所有相关研究）。近来，LSI 在机器学习和统计学中的重要性也日益凸显， [31, 34] 关于蒙特卡洛采样的最新研究便是明证。一种将 LSI 从强凸情形推广到更一般情形的常见方法，是借助 [26] 开创性工作中的扰动论证，采用“球外强凸性”这一假设。尽管这种方法适用于研究 [33], [26], 中的模拟退火过程，但对于研究满足 $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 的凸势函数所对应的采样问题，它似乎并不可靠，因为在这种设定下，当 $\|\theta\|_2$ 取较大值时，并不能推出 $\theta \longmapsto U(\theta)$ 具有渐近强凸性。话虽如此，这种方法或许还有一个更严重的后果：采用扰动方法时，对数索об列夫不等式常数不可避免地依赖维度，从而导致收敛速率随环境空间维度增加而严重呈指数级退化。

为克服这些困难，我们选择采用一种略有不同的泛函不等式。它可视为对 LSI 的无害改动，但事实上似乎很适用于通过 $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 假设描述的弱对数凹情形。为此，我们将使用弱对数索об列夫不等式（下文简称 WLSI），该不等式由 [37] 引入，其作用已在许多研究中得到广泛探讨，并用于获得指数次线性的混合速率，例如 [7] 等。为推导这类不等式，我们将以 [10] 的贡献为出发点，该工作建立了李雅普诺夫条件与 WLSI 之间的联系。我们基于 $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 的方法无疑与 [6] 最近的工作有一些相似之处；后者在可变曲率界的框架下得到了若干泛函不等式（Poincar´eand 传输不等式）。

3.3.2 弱对数索об列夫不等式

我们简要介绍关键的理论要素，详细内容见 [3]。我们引入以下假设，该假设适用于有界函数的情形。

定义 3.1（弱对数 Sobolev 不等式）。对于任意可测空间 $(\Omega\ \mathcal{F}, \mu)$ 和任意良好的函数 f，定义:

$$Ent_\mu(f^2) := \int_\Omega f^2 \log(f^2) \mathrm{d}\mu - \int_\Omega f^2 \mathrm{d}\mu \log\left(\int_\Omega f^2 \mathrm{d}\mu\right).$$

若存在一个非增函数 $\varphi_{WLS}\colon (0, +\infty) \mapsto \mathbb{R}_+$，使得对任意 $f \in \mathcal{C}_{\not\partial}(\Omega)$，测度 μ 满足弱对数 Sobolev 不等式:

$$Ent_\mu(f^2) \leq \varphi_{WLS}(s)\mathcal{E}(f) + s\, Osc^2(f), \tag{28}$$

其中 $Osc(f) :=$ 上确界 $f -$ 下确界 f。

在说明如何使用这一泛函不等式之前，我们先陈述 Poincar´e 不等式与 WLSI 之间的重要关系。

命题 3.3. 假设 μ 满足常数为 C_P 的 Poincar´e 不等式，即对于任意光滑可积函数 f:

$$C_p(\mu)Var_\mu(f) = C_p(\mu)\int_\Omega (f-\mu[f])^2 \mathrm{d}\mu \le \int_\Omega |\nabla f|^2 \mathrm{d}\mu,$$

那么，如果 $\log c = \frac{3}{14e^2}\left(\frac{1}{e}+\frac{1}{2}\right)+1+\log\left(\frac{14}{3}\right)$，则 μ满足弱对数 Sobolev 不等式，
且:

$$\varphi_{WLS}(s) = \begin{cases} 0, & s > \frac{1}{e}+\frac{1}{2} \\ \frac{32}{C_P}\log\left(\frac{c}{s}\right), & s \le \frac{1}{e}+\frac{1}{2} \end{cases}.$$

为便于阅读，我们引入一个通用的 $\mathfrak{a} > 0$，使得:

$$\varphi_{WLS}(s) = \begin{cases} 0, & s > \frac{1}{e}+\frac{1}{2} \\ \mathfrak{a}\frac{1+\log\left(\frac{1}{s}\right)}{C_P}, & s \le \frac{1}{e}+\frac{1}{2} \end{cases}. \tag{29}$$

命题 3.3 的证明。在定义 28 所述的有界情形下，Poincar´é 不等式如何推出弱对数 Sobolev 不等式，其证明在此给出以供参考。此处省略技术细节，相关内容请参阅下列参考文献。我们使用测度-容量不等式（见 [3], 第 8.3 节）。我们知道，Poincar´é 不等式推出一个容量不等式（ [3]中的命题 8.3.1），其常数等于 $2C_P$。然后，我们可以应用 [7] 的定理 2.2，该定理给出了基于命题陈述中所给函数 φ_{WLS} 的弱对数 Sobolev 不等式。 □

3.3.3 在 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 条件下的弱对数 Sobolev 不等式

当然，在前述结果中，唯一重要的依赖关系是由 C_P 引起的，因此有必要在假设 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 下专门研究这一点。下文将把 32 和 $\log(c)$ 视为“通用常数”。

下一个命题在 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 框架下给出了 Poincar´e 常数的两个下界。无论随机采样得到的 $(X_1, \ldots, X_n)$ 取何值，第一个下界始终成立。第二个下界则是在对抽样过程 $(X_1, \ldots, X_n)$ 而言具有高概率的意义下考虑的。

命题 3.4. 假设 $\mathcal{H}_{\min}$,$\mathcal{H}_{\mathbf{no}}$(L ℓ_0), $\mathcal{H}_{\pi_0}(\ell_0)$，并且对于任意 $x, \theta \mapsto -\log\, p_\theta(x)$ 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，则:

i) 对于任意样本 $(X_1, \ldots, X_n)$，有:

$$C_P(\mu_n) \gtrsim_{uc} \frac{1}{\left(d\log^{2\beta}(n)\right)^{(1+r)^2}}$$

ii) 假设 $\theta \mapsto \mathbb{P}_\theta$ 是单射，且存在 θ_0 使得 $(X_1, \ldots, X_n) \sim \mathbb{P}_{\theta_0}$。如果在 θ_0 附近的局部范围内，$\theta \mapsto |\theta-\theta_0|^{-\alpha} W_1(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})$ 不为零，则:

$$\mathbb{E}_{(X_1,\ldots,X_n)\sim\mathbb{P}_{\theta_0}}[C_P(\mu_n)] \gtrsim_{uc} \left(\frac{n}{Ld\log n}\right)^{\alpha}.$$

最后，我们需要对 (28) 中引入的 WLSI 所涉及函数的振荡给出上界，也就是说，我们要为任意时刻 $t > 0$ 下的 $Osc^2\left(\sqrt{\frac{n_t}{\mu_n}}\right)$ 求一个上界。为此，我们注意到马尔可夫半群蕴含 $f_t = \frac{n_t}{\mu_n} = P_t f_0$，其中 $f_0 = \frac{n_0}{\mu_n}$。下一个命题表明，当 n_0 选为协方差矩阵经过精心调整的高斯分布时，f_t 在 $\mathbb{R}^d$ 上有界。

命题 3.5。假设 $\mathcal{H}_{\min}$,$\mathcal{H}_{\mathbf{no}}$(L ℓ_0)、 $\mathcal{H}_{\pi_0}(\ell_0)$，并且对于任意 x， $\theta \mapsto -\log\, p_\theta(x)$ 满足$\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，则:

i) 两个 p 存在正常数 C_1 和 C_2，它们与 n 和 d无关，并且满足：

$$\|f_0\|_\infty \lesssim_{uc} \left(\frac{C_1 d}{n}\right)^{\frac{dr}{2}} \exp\left(C_2 n d^{1+r} \log^{2\beta(1+r)}(n)\right).$$

ii) 因此：

$$Osc^2(\sqrt{f_t}) \leq Osc^2(\sqrt{f_0}) \lesssim_{uc} \left(\frac{C_1 d}{n}\right)^{\frac{dr}{2}} \exp\left(C_2 n d^{1+r} \log^{2\beta(1+r)}(n)\right).$$

iii) 此外，由 *i*) 可直接得到：

$$J_0 = \int_{\mathbb{R}^d} \log\left(f_0(\theta)\right) \mathrm{d}n_0(\theta) \lesssim_{uc} n d^{1+r} \log^{2\beta(1+r)}(n) + d \log\left(\frac{d}{n}\right).$$

3.4 SLMC 的熵收敛

本段旨在证明本文的主要结果，即定理 1.1，该定理保证了 SLMC 算法的收敛性。

定理 1.1 的证明。我们的出发点是与泛函加权对数 Sobolev 不等式 (28) 相关的半群不等式 (25)。利用 $c_{n,d}$ 在 (26) 中定义的量，我们得到对任意 $s > 0$:

$$\begin{aligned}
\partial_t\{J_t\} &\leq -\mathcal{E}\left(\sqrt{\frac{n_t}{\mu_n}}\right) + c_{n,d} e^{-\frac{2\alpha_n}{3}t} \\
&\leq -\frac{J_t}{\varphi_{WLS}(s)} + \frac{s}{\varphi_{WLS}(s)} Osc^2\left(\sqrt{\frac{n_t}{\mu_n}}\right) + c_{n,d} e^{-\frac{2\alpha_n}{3}t} \\
&\leq -\frac{J_t}{\varphi_{WLS}(s)} + \frac{s\, O_{n,d}}{\varphi_{WLS}(s)} + c_{n,d} e^{-\frac{2\alpha_n}{3}t},
\end{aligned}$$

其中最后一行应用了命题 3.5，其中 $O_{n,d} \lesssim_{uc} \left(\frac{C_1 d}{n}\right)^{\frac{dr}{2}} \exp\left(C_2 n d^{1+r} \log^{2\beta(1+r)}(n)\right)$和 C_1 以及 C_2 是两个通用常数。随后我们选择 s（它取决于 t），使得：

$$s_t = e^{-A\sqrt{t+1}} \quad \text{with} \quad A > 1 \quad \text{that will be chosen later on.}$$

我们注意到 $s_t < e^{-1} + 1/2$，因此命题 3.3 的方程 (29) 得出：

$$\varphi_{WLS}(s_t) = \mathfrak{a} \frac{1 + \log\left(\frac{1}{s_t}\right)}{C_P} = \mathfrak{a} \frac{1 + A\sqrt{1+t}}{C_P}.$$

我们引入 $\psi(t) = \exp\left(\frac{C_P}{\mathfrak{a}} \int_0^t \mathrm{d}\frac{u}{1+A\sqrt{1+u}}\right)$，并推导出

$$\psi(t) = \exp\left(\frac{C_P}{\mathfrak{a}} \frac{2A(\sqrt{1+t}-1) - 2\log\left(\frac{1+A\sqrt{1+t}}{1+A}\right)}{A^2}\right) \leq \exp\left(\frac{2C_P}{\mathfrak{a}A}(\sqrt{1+t}-1)\right).$$

现在应用格朗沃尔引理：

$$\begin{aligned}
\partial_t\left\{\psi(t) J_t\right\} &= \left(\frac{C_P}{\mathfrak{a}(1+A\sqrt{1+t})} J_t + J_t'\right)\psi(t) \\
&\leq \left[\frac{C_P O_{n,d}}{\mathfrak{a}} \frac{e^{-A\sqrt{t+1}}}{1+A\sqrt{1+t}} + c_{n,d} e^{-\frac{2\alpha_n}{3}t}\right]\psi(t) \\
&\leq \frac{C_P O_{n,d}}{\mathfrak{a}} e^{-(A - \frac{2C_P}{\mathfrak{a}A})\sqrt{1+t}} + c_{n,d} e^{\frac{2C_P}{\mathfrak{a}A}(\sqrt{1+t}-1) - \frac{2\alpha_n}{3}t}.
\end{aligned}$$

我们用 t_0 表示满足方程 $\frac{2C_P}{\mathfrak{a}A}\sqrt{1+t_0}=\frac{\alpha_n t_0}{3}$. 的正实数值。然后我们观察到:

$$\begin{aligned}\int_0^t e^{\frac{2C_P}{\mathfrak{a}A}(\sqrt{1+u}-1)-\frac{2\alpha_n}{3}u}\mathrm{d}u &\le \int_0^{t_0} e^{\frac{2C_P}{\mathfrak{a}A}\sqrt{1+u}}\mathrm{d}u+\int_{t_0}^{+\infty} e^{-\frac{\alpha_n}{3}u}\mathrm{d}u\\ &\le t_0 e^{\frac{2C_P}{\mathfrak{a}A}\sqrt{1+t_0}}+\frac{3}{\alpha_n}=t_0 e^{\frac{\alpha_n t_0}{3}}+\frac{3}{\alpha_n}.\end{aligned}$$

若选择 A 使得 $A>\frac{2C_P}{\mathfrak{a}A}$，则可推得:

$$\begin{aligned}J_t &\le \left(J_0+c_{n,d}t_0e^{\frac{\alpha_n t_0}{3}}+\frac{3c_{n,d}}{\alpha_n}\right)\psi(t)^{-1}+\frac{C_PO_{n,d}}{\mathfrak{a}}\psi(t)^{-1}\int_0^t e^{-\left(A-\frac{2C_P}{\mathfrak{a}A}\right)\sqrt{1+u}}\mathrm{d}u\\ &\le \left(J_0+c_{n,d}t_0e^{\frac{\alpha_n t_0}{3}}+\frac{3c_{n,d}}{\alpha_n}\right)\psi(t)^{-1}+\frac{2C_PO_{n,d}}{\mathfrak{a}\left(A-\frac{2C_P}{\mathfrak{a}A}\right)^2}\psi(t)^{-1},\end{aligned}$$

其中，上一行用到了以下界:

$$\int_0^t e^{-b\sqrt{1+u}}\mathrm{d}u\le\int_0^{+\infty}e^{-b\sqrt{1+u}}\mathrm{d}u\le\frac{2}{b^2}.$$

为得到最小的上界，我们需要选择 A ，使得 $\frac{2C_P}{\mathfrak{a}A}$尽可能大且小于 A，因此自然得到如下选择:

$$\frac{2C_P}{\mathfrak{a}A}=\frac{A}{2}\Longrightarrow A=\frac{2}{\sqrt{\mathfrak{a}}}\sqrt{C_P}.$$

将 A 的该取值代入前面的界中，可见 $t_0\le\frac{3\sqrt{C_P}}{\alpha_n\sqrt{\mathfrak{a}}}+\frac{C_P}{\alpha_n^2}$，因此存在常数 C使得:

$$J_t\le C\left(J_0+\frac{c_{n,d}}{\alpha_n}\left[1+\left(\frac{C_P}{\alpha_n}+\sqrt{C_P}\right)e^{\frac{\sqrt{C_P}}{\sqrt{\mathfrak{a}}}+\frac{C_P}{3\alpha_n}}\right]+O_{n,d}\right)(1+t)^{1/4}e^{-\frac{\sqrt{C_p}}{\sqrt{a}}(\sqrt{1+t}-1)}. \tag{30}$$

在命题 3.4 中，我们得到了 $C_P\ge\frac{\kappa}{(d\log^{2\beta}(n))^{(1+r)^2}}$。若不使用常数 C_P，而直接使用 $\frac{\kappa}{(d\log^{2\beta}(n))^{(1+r)^2}}$，其中 $\kappa<1$，则之前的所有计算均保持不变，只需将C_P 替换为其下界，并且:

$$J_t\le C\left(J_0+\frac{c_{n,d}}{\alpha_n}e^{\frac{\sqrt{\kappa}\left(\frac{1}{\sqrt{a}}+\frac{1}{3\alpha_n}\right)}{(d\log^{2\beta}(n))^{(1+r)^2/2}}}+O_{n,d}\right)(1+t)^{1/4}e^{-\frac{\sqrt{\kappa}(\sqrt{1+t}-1)}{\sqrt{a}(d\log^{2\beta}(n))^{(1+r)^2/2}}}. \tag{31}$$

使用 $O_{n,d}$、 $c_{n,d}$ 的取值以及 J_0的上界，最终可见，若 $\alpha_n=n\,d\log^{2\beta}(n)^{1+r}$，则:

$$t\ge\aleph\left(d\log^{2\beta}(n)\right)^{(1+r)^2}\left[\log^2(\varepsilon^{-1})+n^2\left(d\log^{2\beta}(n)\right)^{2(1+r)}+d^2\log^2 d\right]\Longrightarrow J_t\le\varepsilon.$$

□

4 KL 与 U_{ν_n}的技术结果

4.1 Kurdyka-Lojasiewicz 不等式下的增长性质

这里回顾 KL不等式的一些重要推论，它们揭示了函数与其梯度范数之间的若干关系。这些不等式的证明见 [21] 的引理15（原文中有一处小错误，我们在本文中对表述作了修正，补上了因子 2）。

命题 4.1. 假设函数 V 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，则：

$$\frac{2\mathfrak{c}}{1-r}\left[V^{1-r}(\theta)-\min(V)^{1-r}\right] \le \|\nabla V(\theta)\|_2^2 \le 2L\left[V(\theta)-\min(V)\right], \quad \forall\theta\in\mathbb{R}^d.$$

此外，还可以评估任何满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 的函数的最小和最大增长性质；这类函数必然以到其极小化器的距离的正幂为下界和上界。

命题 4.2。假设函数 V 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，则有 $\forall\theta\in\mathbb{R}^d$：

$$V^{1+r}(\theta)-\min(V)^{1+r} \ge \frac{(1+r)\mathfrak{c}}{2}\|\theta-\arg\min V\|_2^2,$$

and

$$V(\theta)-\min(V) \le \frac{L}{2}\|\theta-\arg\min V\|_2^2.$$

由第一个不等式可直接得出

命题 4.3。假设函数 V 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，则有 $\forall\theta\in\mathbb{R}^d$：

$$V(\theta) \ge 2^{-\frac{r}{1+r}}\left(\min(V)+\left(\frac{(1+r)\mathfrak{c}}{2}\right)^{\frac{1}{1+r}}\|\theta-\arg\min V\|_2^{\frac{2}{1+r}}\right).$$

4.2 U_{ν_n} 的性质

命题 1.1 的证明。首先，我们注意到，若每个 $\theta\mapsto\nabla\log p_\theta(\mathrm{X}_i)$ 都是 L-Lipschitz 的，且 $\theta\mapsto\nabla\log\pi_0$ 是 ℓ_0-Lipschitz 的，那么由三角不等式可得

$$\|\nabla U_{\nu_n}(\theta_1)-\nabla U_{\nu_n}(\theta_2)\|_2 \le (nL+\ell_0)\|\theta_1-\theta_2\|_2.$$

其次，我们考察曲率的下界性质，并观察到：

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}(\theta)} = \inf_{e\in\mathbb{R}^d:|e|=1} e^T(\nabla^2 U_{\nu_n})(\theta)e \ge \frac{1}{n}\sum_{i=1}^n \inf_{e\in\mathbb{R}^d:|e|=1} e^T(\nabla^2 U_{\mathbf{X}_i})(\theta)e.$$

先验的对数凹性推出

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}(\theta)} \ge \frac{1}{n}\sum_{i=1}^n \underline{\lambda}_{\nabla^2(-n\log p_\theta(\mathbf{X}_i))} = \sum_{i=1}^n \underline{\lambda}_{\nabla^2(-\log p_\theta(\mathbf{X}_i))}.$$

然后，将 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$ 性质应用于上述求和中的每一项，并取 $\min_{\theta\in\mathbb{R}^d} -\log\pi_0(\theta) > 0$，可得

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}(\theta)} \ge \mathfrak{c}\sum_{i=1}^n[-\log p_\theta(\mathbf{X}_i)]^{-r} \ge \mathfrak{c}n^r\sum_{i=1}^n U_{\mathbf{X}_i}^{-r}(\theta) = \mathfrak{c}n^{1+r}\left(\frac{1}{n}\sum_{i=1}^n U_{\mathbf{X}_i}^{-r}(\theta)\right).$$

由詹森不等式，最终可推得：

$$\underline{\lambda}_{\nabla^2 U_{\nu_n}(\theta)} \ge \mathfrak{c}n^{1+r}\left(\frac{1}{n}\sum_{i=1}^n U_{\mathbf{X}_i}^{-r}(\theta)\right) \ge \mathfrak{c}n^{1+r}U_{\nu_n}^{-r}(\theta).$$

由此可知，U_{ν_n} 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r}, nL+\ell_0\right)$。对于 $U_{\mathbf{X}_i}$，证明类似。 □

命题 4.4. 我们假设 $\mathcal{H}_{\pi_0}(\ell_0)$、$\mathcal{H}_{\min}$，并且对于任意 x：$\theta\longmapsto-\log p_\theta(x)$ 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，则：

$$\|\arg\min U_{\nu_n}\|_2 \lesssim_{uc} d^{\frac{1+r}{2}}\log^{\beta(1+r)}(n) \quad and \quad \min_{\theta\in\mathbb{R}^d} U_{\nu_n}(\theta) \lesssim_{uc} nd\log^{2\beta}(n).$$

证明。命题 1.1 表明 U_{ν_n} 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r}, nL+\ell_0\right)$。因此，我们可以应用命题 4.2，取 $\theta=0$，并得出：

$$\|\arg\min U_{\nu_n}\|_2^2 \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}}\left(U_{\nu_n}^{1+r}(0)-\min U_{\nu_n}^{1+r}\right).$$

为了得到 $U_{\nu_n}(0)$ 的上界，我们首先利用命题 4.2 对所有 i，给出 $U_{\mathbf{X}_i}(0)$ 的上界，如下：

$$U_{\mathbf{X}_i}(0) \leq \min U_{\mathbf{X}_i} + \frac{nL+\ell_0}{2}\|\arg\min U_{\mathbf{X}_i}\|_2^2 \lesssim_{uc} d + nd\log^{2\beta}(n) \lesssim_{uc} nd\log^{2\beta}(n),$$

于是 $U_{\nu_n}(0) \lesssim_{uc} nd\log^{2\beta}(n)$。由此我们推得：

$$\|\arg\min U_{\nu_n}\|_2^2 \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}}U_{\nu_n}^{1+r}(0) \lesssim_{uc} d^{1+r}\log^{2\beta(1+r)}(n).$$

第二部分来自 $\min U_{\nu_n} \leq U_{\nu_n}(0)$。 □

5 半群的光滑性与有界性

命题 3.4 的证明。 i）。证明依赖于一个以“固定”样本（$X_1,\ldots,X_n$）为基础的论证。我们的出发点是命题 4.2 及 Kurdyka-Łojasiewicz 不等式的推论。由于 $\mathcal{H}_{\pi_0}(\ell_0)$ 且 $\theta \mapsto -\log p_\theta(\mathbf{X}_i)$ 满足 $\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$，命题 1.1 表明 U_{ν_n} 满足$\mathcal{H}^{\mathfrak{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r}, nL+\ell_0\right)$。因此，我们可以应用命题 4.2 并推得：

$$\|\theta-\arg\min U_{\nu_n}\|_2^2 \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}}\left(U_{\nu_n}^{1+r}(\theta)-\min U_{\nu_n}^{1+r}\right) \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}}U_{\nu_n}^{1+r}(\theta).$$

若 I_d 表示恒等映射，我们利用以下事实：对于任意分布 μ，对任意 $a\in\mathbb{R}^d$，都有 $Var[\mu] \leq \mu[\|I_d-a\|_2^2]$，因此，结合 $a=\arg\min U_{\nu_n}$，可直接推出：

$$Var(\mu_n) \leq \int_{\mathbb{R}^d}\|\theta-\arg\min U_{\nu_n}\|_2^2 \mathrm{d}\mu_n(\theta) \leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}}\mu_n[U_{\nu_n}^{1+r}].$$

接下来，我们利用 $(\theta_t)_{t\geq 0}$ 的遍历行为，并注意到存在一个与 n 和 d 无关的常数 C，使得：

$$\begin{aligned} Var(\mu_n) &\leq \frac{2}{(1+r)\mathfrak{c}n^{1+r}}\limsup_{t\geq 0}\mathbb{E}[U_{\nu_n}^{1+r}(\theta_t)] \\ &\leq C\left(d\log^{2\beta}(n)\right)^{(1+r)^2}, \end{aligned}$$

最后一个不等式由命题 5.1 得出。现在，我们对数凹分布的 Poincaŕe 常数使用 Bobkov 界（参见 [4] 的定理 1.2），并由此推得存在一个通用常数 K，使得：

$$C_P(\mu_n) \geq \frac{1}{4K^2 Var(\mu_n)}.$$

利用方差的上界，我们推得存在一个通用常数 $\kappa>0$，使得：

$$C_P(\mu_n) \geq \frac{\kappa}{\left(d\log^{2\beta}(n)\right)^{(1+r)^2}}.$$

ii）。对于第二点，我们考虑对样本取平均的情形，结果利用了后验分布集中在其均值附近这一性质。根据 [21] 的定理 3，我们知道存在一个常数 $c>0$，使得：

$$\mathbb{E}_{(X_1,\ldots,X_n)\sim\mathbb{P}_{\theta_0}}[\mathbb{V}ar(\mu_n)] \leq c\epsilon_{n,d}^2,$$

其中 $\epsilon_{n,\ d}=\left(\frac{Ld}{\text{对数}\ n}\ n\right)^{\alpha^{-1}}$。利用詹森不等式和 Bobkov 界即可得到该结果。 □

命题 3.5 的证明。 i）。我们首先证明 f_0 的有界性。根据我们的假设，应用命题 1.1 可得 U_{ν_n} 满足 $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r},\ nL+\ell_0\right)$。若 $\theta_n^\star = \arg\min U_{\nu_n}$，则由命题 4.2 可推得：

$$f_0(\theta)=\frac{n_0(\theta)}{\mu_n(\theta)}=\frac{Z_n e^{-\frac{\|\theta\|_2^2}{2\sigma^2}+U_{\nu_n}(\theta)}}{(2\pi)^{d/2}\sigma^d}\le\frac{Z_n e^{-\frac{\|\theta\|_2^2}{2\sigma^2}+U_{\nu_n}(\theta_n^\star)+\frac{(nL+\ell_0)}{2}\|\theta-\theta_n^\star\|_2^2}}{(2\pi)^{d/2}\sigma^d}. \tag{32}$$

我们计算 Z_n 的上界，并使用由命题 4.3 得出的 U_{ν_n} 下界：

$$\begin{aligned}Z_n&=\int_{\mathbb{R}^d}e^{-U_{\nu_n}(\theta)}\mathrm{d}\theta\\&\le\int_{\mathbb{R}^d}e^{-2^{-\frac{r}{1+r}}\left[U_{\nu_n}(\theta_n^\star)+n\left(\frac{(1+r)\mathfrak{c}}{2}\right)^{\frac{1}{1+r}}\|\theta-\theta_n^\star\|_2^{\frac{2}{1+r}}\right]}\mathrm{d}\theta\\&\le e^{-2^{-\frac{r}{1+r}}U_{\nu_n}(\theta_n^\star)}\int_{\mathbb{R}^d}e^{-na_r\|\theta\|_2^{\frac{2}{1+r}}}\mathrm{d}\theta,\end{aligned}$$

1</style}其中 $\left(\frac{(1+r)\mathfrak{c}}{2}\right)^{\frac{1}{1+r}}$。利用众所周知的等式：

$$\int_{\mathbb{R}^d}e^{-a|\theta|^\ell}\mathrm{d}\theta=\frac{d\pi^{d/2}\Gamma(d/\ell)}{\ell a^{d/\ell}\Gamma(d/2+1)},\quad\forall a>0,\quad\forall\ell>0.$$

于是，结合 $a=na_r$ 和 $\ell=\frac{2}{1+r}$，我们可推得：

$$Z_n\le e^{-2^{-\frac{r}{1+r}}U_{\nu_n}(\theta_n^\star)}\int_{\mathbb{R}^d}e^{-na_r\|\theta\|_2^{\frac{2}{1+r}}}\mathrm{d}\theta\le\frac{d(1+r)}{2}\frac{\pi^{d/2}}{(na_r)^{\frac{d(1+r)}{2}}}\frac{\Gamma\left(\frac{d(1+r)}{2}\right)}{\Gamma\left(\frac{d}{2}+1\right)}.$$

根据伽马函数的标准关系：

$$Z_n\le 2\left(\frac{2^{1+r}\pi}{\mathfrak{c}n^{1+r}}\right)^{\frac{d}{2}}d^{\frac{dr}{2}}. \tag{33}$$

综合方程（32）和（33），可得：

$$f_0(\theta)\le 2e^{U_{\nu_n}(\theta_n^\star)}\left(\frac{2}{\mathfrak{c}\sigma^2n^{1+r}}\right)^{\frac{d}{2}}d^{\frac{dr}{2}}e^{-\frac{\|\theta\|_2^2}{2\sigma^2}+\frac{(nL+\ell_0)}{2}\|\theta-\theta_n^\star\|_2^2}.$$

对于所有 $\sigma^2<\frac{1}{nL+\ell_0}$，对 θ 进行直接优化可得：

$$\|f_0\|_\infty\le 2e^{U_{\nu_n}(\theta_n^\star)}\left(\frac{2}{\mathfrak{c}\sigma^2n^{1+r}}\right)^{\frac{d}{2}}d^{\frac{dr}{2}}\exp\left(\frac{(nL+\ell_0)}{2(1-\sigma^2(nL+\ell_0))}\|\theta_n^\star\|_2^2\right).$$

然后，选择 $\frac{c_1}{nL+\ell_0}\le\sigma^2\le\frac{c_2}{nL+\ell_0}$，其中 $0<c_1\le c_2<1$ 属于 $\mathcal{H}_{\mathbf{n_0}}(\mathbf{L}\ \boldsymbol{\ell_0})$，并利用命题 4.4 中 $\|\theta_n^\star\|_2^2$ 和 $U_{\nu_n}(\theta_n^\star)$ 的界，可得：

$$\|f_0\|_\infty\le 2\left(\frac{C_1d}{n}\right)^{\frac{dr}{2}}\exp\left(C_2nd^{1+r}\log^{2\beta(1+r)}(n)\right),$$

其中 C_1 和 C_2 是通用常数。

ii）。这一结果几乎是布朗扩散所对应的马尔可夫半群性质的最大值原理的标准推论。对于任意有界可测的 $h>0$，利用马尔可夫性质，我们观察到 $P_th>0$，进而将 g_t 定义为如下函数 $g_t:=\sqrt{P_th}$。随后，我们引入 $\overline{\theta}(t)$ 和 $\underline{\theta}(t)$，定义如下：

$$\overline{\theta}(t)=\arg\max g_t(\theta)\quad\text{ and }\quad\underline{\theta}(t)=\arg\min g_t(\theta).$$

链式法则给出：

$$\begin{aligned}\frac{d}{dt}Osc(g_t) &= \frac{d}{dt}\left(g_t(\overline{\theta}(t)) - g_t(\underline{\theta}(t))\right)\\ &= \frac{dg_t}{dt}(\overline{\theta}(t)) + \left\langle \nabla g_t(\overline{\theta}(t)), \frac{d\overline{\theta}(t)}{dt}\right\rangle - \frac{dg_t}{dt}(\underline{\theta}(t)) - \left\langle \nabla g_t(\underline{\theta}(t)), \frac{d\underline{\theta}(t)}{dt}\right\rangle. \end{aligned} \tag{34}$$

我们计算：

$$\begin{aligned}\frac{dg_t}{dt}(\theta) &= \frac{1}{2\sqrt{P_t h}}\frac{dP_t h}{dt}(\theta)\\ &= \frac{1}{2\sqrt{P_t h}}\mathcal{G}_t P_t h(\theta)\\ &= \frac{1}{2\sqrt{P_t h(\theta)}}\left[-\sum_{i=1}^{n}\langle \nabla_\theta P_t h(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\mathbf{X}_i|\theta) + \Delta_\theta P_t h(\theta)\right]. \end{aligned} \tag{35}$$

现在，我们利用 $\overline{\theta}(t) = \arg\max\ g_t = \arg\max P_t h$，（对 $\underline{\theta}(t)$ 也可作类似论证）：

$$\nabla_\theta g_t(\overline{\theta}(t)) = 0, \quad \nabla_\theta P_t h(\overline{\theta}(t)) = 0 \quad \text{and} \quad \Delta_\theta P_t h(\overline{\theta}(t)) \le 0.$$

于是：

$$\begin{aligned}\frac{d}{dt}Osc(g_t) &= \frac{dg_t}{dt}(\overline{\theta}(t)) - \frac{dg_t}{dt}(\underline{\theta}(t))\\ &= \frac{\Delta_\theta P_t h}{2\sqrt{P_t h}}(\overline{\theta}(t)) - \frac{\Delta_\theta P_t h}{2\sqrt{P_t h}}(\underline{\theta}(t))\\ &\le 0. \end{aligned} \tag{36}$$

因此，我们已证明 $Osc(\sqrt{P_t h})$ 关于 $t \ge 0$ 单调递减，从而完成证明。 □

引理 2.1 的证明。我们按照 [33] 中命题3的做法，证明在推导陈述中涉及的积分时可以使用勒贝格控制收敛定理。于是我们可以推出：

$$\partial_t\left\{\int_{\mathbb{R}^d} f_t(\theta)\mathrm{d}n_t(\theta)\right\} = \int_{\mathbb{R}^d}\partial_t\{f_t(\theta)\}\mathrm{d}n_t(\theta) + \int_{\mathbb{R}^d} f_t(\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta.$$

我们保持第一项不变，现在着重考察第二项：

$$\begin{aligned}\int_{\mathbb{R}^d} f_t(\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta &= \int_{\mathbb{R}^d} f_t(\theta)\partial_t\left\{\sum_{i=1}^{n} m_t(\theta, \mathbf{X}_i)\right\}\mathrm{d}\theta\\ &= \int_{\mathbb{R}^d}\sum_{i=1}^{n} f_t(\theta)\partial_t\{m_t(\theta, \mathbf{X}_i)\}\mathrm{d}\theta\\ &= \int_{\mathbb{R}^d}\sum_{i=1}^{n}\mathcal{L}f_t(\theta)\ m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta, \end{aligned}$$

其中，第一步使用了 n_t 的定义，最后一步使用了柯尔莫哥洛夫后向方程 (10)。由于函数 $f_t(\theta)$ 不依赖于 x，我们注意到 $\mathcal{L}_2 f_t(\theta) = 0$，因此只需计算剩余项 $\mathcal{L}_1 f_t(\theta)$：

$$\begin{aligned}\int_{\mathbb{R}^d} f_t(\theta)\partial_t\{n_t(\theta)\}\mathrm{d}\theta &= \int_{\mathbb{R}^d}\sum_{i=1}^{n}\mathcal{L}_1 f_t(\theta)\ m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta \qquad (37)\\ &= \int_{\mathbb{R}^d}\sum_{i=1}^{n}\left[-\langle \nabla_\theta f_t(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle + \Delta_\theta f_t(\theta)\right] m_t(\theta, \mathbf{X}_i)\mathrm{d}\theta\\ &= -\int_{\mathbb{R}^d}\sum_{i=1}^{n}\langle \nabla_\theta f_t(\theta), \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\mathbf{X}_i|\theta)\mathrm{d}n_t(\theta) + \int_{\mathbb{R}^d}\Delta_\theta f_t(\theta)\mathrm{d}n_t(\theta)\\ &= \int_{\mathbb{R}^d}\mathcal{G}_t f_t(\theta)\mathrm{d}n_t(\theta), \qquad (38) \end{aligned}$$

这里我们用到了如下事实： $m_t(\theta, \mathbf{X}_i) = m_t(\mathbf{X}_i|\theta)n_t(\theta)$。 □

5.1 矩的上界

命题 5.1。假设 $\mathcal{H}_{\mathbf{n_0}}(\mathrm{L}\ \ell_0)$、 $\mathcal{H}_{\pi_0}(\ell_0)$、 $\mathcal{H}_{\min}$，并且对于每个 $\mathbf{X}_i$， $\theta \mapsto -\log\ p_\theta(\mathbf{X}_i)$ 满足$\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}(\mathfrak{c}, L)$。则：

i) 存在三个正数常数 C_1、 C_2 和 C_3，它们与 n 和 d无关，使得对于任意 $t > 0$:

$$\mathbb{E}\left[e^{\frac{(1+r)n\mathfrak{c}}{16}\frac{1}{1+r}(\|\theta_t\|_2^2+1)^{\frac{1}{1+r}}}\right] \leq C_1\left(d\log^{2\beta}(n)\right)^{\frac{r}{1+r}} e^{C_2 nd\log^{2\beta}(n)} + C_3^d e^{\frac{(1+r)n\mathfrak{c}}{16}\frac{1}{1+r}}.$$

ii) 对任意 $t > 0$ 以及任意 $\alpha \geq 1$:

$$\mathbb{E}[U^\alpha_{\nu_n}(\theta_t)] \lesssim_{uc} n^\alpha\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)}.$$

i) 的证明。考虑函数 $f(\theta) = \exp\left(\frac{a}{2}(\|\theta\|_2^2+1)^\rho\right)$，其中 $0 < \rho < 1$。该函数二阶可微。 f 的梯度计算如下:

$$\nabla f(\theta) = a\rho(\|\theta\|_2^2+1)^{\rho-1}f(\theta)\theta.$$

拉普拉斯算子表示为:

$$\Delta f(\theta) = a\rho(\|\theta\|_2^2+1)^{\rho-2}f(\theta)\left[a\rho(\|\theta\|_2^2+1)^\rho\|\theta\|_2^2 + (d+2\rho-2)\|\theta\|_2^2 + d\right].$$

于是我们推得，对于任意 $\theta \in \mathbb{R}^d$:

$$\begin{aligned}
\mathcal{G}_t f(\theta) &= -\sum_{i=1}^n \langle \nabla U_{\mathbf{X}_i}, \nabla f(\theta)\rangle m_t(\mathbf{X}_i|\theta) + \Delta f(\theta) \\
&= a\rho(\|\theta\|_2^2+1)^{\rho-2}f(\theta)\Big[-(\|\theta\|_2^2+1)\sum_{i=1}^n \langle \theta, \nabla_\theta U_{\mathbf{X}_i}(\theta)\rangle m_t(\mathbf{X}_i|\theta) \\
&\quad + a\rho(\|\theta\|_2^2+1)^\rho\|\theta\|_2^2 + (d+2\rho-2)\|\theta\|_2^2 + d\Big] \\
&\leq a\rho(\|\theta\|_2^2+1)^{\rho-2}f(\theta)\Big[-(\|\theta\|_2^2+1)\sum_{i=1}^n (U_{\mathbf{X}_i}(\theta) - U_{\mathbf{X}_i}(0))\, m_t(\mathbf{X}_i|\theta) \\
&\quad + a\rho(\|\theta\|_2^2+1)^{\rho+1} + d\left(\|\theta\|_2^2+1\right)\Big] \\
&\leq a\rho(\|\theta\|_2^2+1)^{\rho-1}f(\theta)\left[-\sum_{i=1}^n (U_{\mathbf{X}_i}(\theta) - U_{\mathbf{X}_i}(0))\, m_t(\mathbf{X}_i|\theta) + a\rho(\|\theta\|_2^2+1)^\rho + d\right],
\end{aligned}$$

其中我们利用了 U_x 对任意位置 x 的凸性。

下面建立 $U_{\mathbf{X}_i}(\theta)$ 和 $U_{\mathbf{X}_i}(0)$ 的界。记 $\theta_i = \operatorname{argmin} U_{\mathbf{X}_i}$，根据假设 $\mathcal{H}_{\min}$，存在两个与 n 和 d 无关的正数常数 $\mathcal{K}_1$ 和 $\mathcal{K}_2$，使得:

$$\max_i \|\theta_i\|_2^2 \leq \mathcal{K}_1 d\log^{2\beta}(n) \quad \text{and} \quad \max_i U_{\mathbf{X}_i}(\theta_i) \leq \mathcal{K}_2 d.$$

我们将命题 4.2 应用于每个满足 $\mathcal{H}^{\mathrm{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r}, nL+\ell_0\right)$的非负函数 $U_{\mathbf{X}_i}$，于是得到:

$$U_{\mathbf{X}_i}(\theta) \geq n\left[\frac{(1+r)\mathfrak{c}}{2}\right]^{\frac{1}{1+r}}\|\theta-\theta_i\|_2^{\frac{2}{1+r}}.$$

由于 $\frac{2}{1+r} > 1$，根据詹森不等式可得 $(u+v)^{\frac{2}{1+r}} \leq 2^{\frac{1-r}{1+r}}\left[u^{\frac{2}{1+r}} + v^{\frac{2}{1+r}}\right]$，对所有 $(u,v) \in \mathbb{R}_+^2$ 均成立，因此我们推得:

$$\|\theta-\theta_i\|_2^{\frac{2}{1+r}} \geq 2^{\frac{r-1}{1+r}}\|\theta\|_2^{\frac{2}{1+r}} - \|\theta_i\|_2^{\frac{2}{1+r}} \geq 2^{\frac{r-1}{1+r}}\|\theta\|_2^{\frac{2}{1+r}} - \left(\mathcal{K}_1 d\log^{2\beta}(n)\right)^{\frac{1}{1+r}}.$$

然后我们利用这个不等式得到 $U_{\mathbf{X}_i}$ 的下界:

$$U_{\mathbf{X}_i}(\theta) \geq 2n\left[\frac{(1+r)\mathfrak{c}}{8}\right]^{\frac{1}{1+r}}\|\theta\|_2^{\frac{2}{1+r}} - n\left[\frac{(1+r)\mathfrak{c}}{2}\right]^{\frac{1}{1+r}}(\mathcal{K}_1 d\log^{2\beta}(n))^{\frac{1}{1+r}}.$$

此外，max $U_{\mathbf{X}_i}(0)$ 的一个上界可由命题 1.1 和 4.2 得出，如下所示:

$$U_{\mathbf{X}_i}(0) \leq U_{\mathbf{X}_i}(\theta_i) + \frac{nL+\ell_0}{2}\|\theta_i\|_2^2 \leq \mathcal{K}_2 d + \frac{\mathcal{K}_1(nL+\ell_0)d\log^{2\beta}(n)}{2}.$$

利用前面的界以及 $\sum_{i=1}^n m_t(\mathbf{X}_i|\theta) = 1$ 这一事实，可得:

$$\begin{aligned}
&\sum_{i=1}^n (U_{\mathbf{X}_i}(\theta) - U_{\mathbf{X}_i}(0))\, m_t(\mathbf{X}_i|\theta) \\
\geq\ & 2n\left[\frac{(1+r)\mathfrak{c}}{8}\right]^{\frac{1}{1+r}}\|\theta\|_2^{\frac{2}{1+r}} - n\left[\frac{(1+r)\mathfrak{c}}{2}\right]^{\frac{1}{1+r}}(\mathcal{K}_1 d\log^{2\beta}(n))^{\frac{1}{1+r}} - \mathcal{K}_2 d - \frac{\mathcal{K}_1(nL+\ell_0)d\log^{2\beta}(n)}{2} \\
\geq\ & \frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4}\|\theta\|_2^{\frac{2}{1+r}} - n\mathfrak{c}^{\frac{1}{1+r}}(\mathcal{K}_1 d\log^{2\beta}(n))^{\frac{1}{1+r}} - \mathcal{K}_2 d - \frac{\mathcal{K}_1(nL+\ell_0)d\log^{2\beta}(n)}{2},
\end{aligned}$$

其中我们使用了一些一致上界（当 $r \in [0,1)$ 。接着，我们选择 $\rho = \frac{1}{1+r}$，由此可得:

$$\begin{aligned}
\mathcal{G}_t f(\theta) \leq\ & \frac{a}{1+r}(\|\theta\|_2^2+1)^{-\frac{r}{1+r}} f(\theta)\left[-\frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4}\|\theta\|_2^{\frac{2}{1+r}} + n\mathfrak{c}^{\frac{1}{1+r}}(\mathcal{K}_1 d\log^{2\beta}(n))^{\frac{1}{1+r}} + \mathcal{K}_2 d\right. \\
&\left. + \frac{\mathcal{K}_1(nL+\ell_0)d\log^{2\beta}(n)}{2} + \frac{a}{(1+r)}(\|\theta\|_2^2+1)^{\frac{1}{1+r}} + d\right] \\
\leq\ & \frac{a}{1+r}(\|\theta\|_2^2+1)^{-\frac{r}{1+r}} f(\theta)\left[-\left(\frac{n\mathfrak{c}^{\frac{1}{1+r}}}{4} - \frac{a}{(1+r)}\right)\|\theta\|_2^{\frac{2}{1+r}} + n\mathfrak{c}^{\frac{1}{1+r}}(\mathcal{K}_1 d\log^{2\beta}(n))^{\frac{1}{1+r}}\right. \\
&\left. + (\mathcal{K}_2+1)d + \frac{\mathcal{K}_1(nL+\ell_0)d\log^{2\beta}(n)}{2} + \frac{a}{(1+r)}\right],
\end{aligned}$$

其中第二行用到了 $(\|\theta\|_2^2+1)^{\frac{1}{1+r}} \leq \|\theta\|_2^{\frac{2}{1+r}} + 1$ 。

现在固定 $a = {}^{n}\!\!\not{(1+r)\mathfrak{c}}_{8}^{\frac{1}{1+r}}$，并可得:

$$\begin{aligned}
\frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq\ & \frac{n^2\mathfrak{c}^{\frac{2}{1+r}}}{64}(\|\theta\|_2^2+1)^{-\frac{r}{1+r}}\left[-\|\theta\|_2^{\frac{2}{1+r}} + 8(\mathcal{K}_1 d\log^{2\beta}(n))^{\frac{1}{1+r}} + \right. \\
&\left. + \frac{8(\mathcal{K}_2+1)d + 4\mathcal{K}_1(nL+\ell_0)d\log^{2\beta}(n)}{n\mathfrak{c}^{\frac{1}{1+r}}} + 1\right].
\end{aligned} \tag{39}$$

接下来，我们研究两种互补的情形。下文中，我们用 $K_{n,d}$ 表示前述 Lyapunov 收缩所涉及的关键紧集的半径:

$$K_{n,d}^{\frac{2}{1+r}} = Cd\log^{2\beta}(n).$$

- 当 $\|\theta\|_2$ 足够大时（$\|\theta\|_2 \geq K_{n,\,d}$），我们发现存在一个足够大的 $C>0$，它与n 和 d无关，使得:

$$\|\theta\|_2^{\frac{2}{1+r}} \geq Cd\log^{2\beta}(n) \Longrightarrow \frac{\mathcal{G}_t f(\theta)}{f(\theta)} \leq -\frac{n^2\left(d\log^{2\beta}(n)\right)^{\frac{1}{1+r}}\mathfrak{c}^{\frac{2}{1+r}}}{128} = -a_{n,d}. \tag{40}$$

• 当 $\|\theta\|_2$ 有上界时（$\|\theta\|_2 \le K_{n,\ d}$），我们使用公式（39）中给出的上界，并得到存在一个通用的 C_1（其值可能随行变化），使得：

$$\begin{aligned}
&\|\theta\|_2^{\frac{2}{1+r}} \le Cd\log^{2\beta}(n) \Longrightarrow \\
&\mathcal{G}_t f(\theta) \le C_1 n^2 f(\theta)\left[8(\mathcal{K}_1 d\log^{2\beta}(n))^{\frac{1}{1+r}} + \frac{8(\mathcal{K}_2+1)d + 4\mathcal{K}_1(nL+\ell_0)d\log^{2\beta}(n)}{n\mathfrak{c}^{\frac{1}{1+r}}} + 1\right] \\
&\le C_1 n^2 d\log^{2\beta}(n)\exp\left(\frac{(C+1)c^{\frac{1}{1+r}} nd\log^{2\beta}(n)}{8}\right) \le b_{n,d}e^{\delta_{n,d}}. \qquad (41)
\end{aligned}$$

然后我们按如下方式使用公式（40）和（41）。我们将函数 $\psi_{n,d}$ 定义为 $\psi_{n,d}(t) = \mathbb{E}[f(\theta_t)]$，并使用引理 2.1：

$$\begin{aligned}
\psi'_{n,d}(t) &= \mathbb{E}[\mathcal{G}_t f(\theta_t)] \\
&= \mathbb{E}\left[\mathcal{G}_t f(\theta_t)\left[\mathbb{I}_{\|\theta_t\|_2 \ge K_{n,d}} + \mathbb{I}_{\|\theta_t\|_2 \le K_{n,d}}\right]\right] \\
&\le \mathbb{E}\left[-a_{n,d} f(\theta_t)\mathbb{I}_{\|\theta_t\|_2 \ge K_{n,d}} + b_{n,d}e^{\delta_{n,d}}\mathbb{I}_{\|\theta_t\|_2 \le K_{n,d}}\right] \\
&\le -a_{n,d}\psi_{n,d}(t) + a_{n,d}\sup_{\|\theta\|_2 \le K_{n,d}} f(\theta) + b_{n,d}e^{\delta_{n,d}} \\
&\le -a_{n,d}\psi_{n,d}(t) + (a_{n,d}+b_{n,d})e^{\delta_{n,d}}.
\end{aligned}$$

应用格朗沃尔引理，可得：

$$\forall t > 0 \qquad \psi_{n,d}(t) \le \left(1 + \frac{b_{n,d}}{a_{n,d}}\right)e^{\delta_{n,d}} + \psi_{n,d}(0)e^{-a_{n,d}t}. \qquad (42)$$

利用 n_0 是在假设 $\mathcal{H}_{\mathbf{n_0}}(\mathrm{L}\ \ell_0)$ 中固定的高斯分布，我们得到 $\psi_{n,d}(0) = E[f(\theta_0)] = \int_{\mathbb{R}^d} f(\theta)\mathrm{d}n_0(\theta)$ 的一个上界如下：

$$\begin{aligned}
\psi_{n,d}(0) &= \left(2\pi\sigma^2\right)^{-\frac{d}{2}}\int_{\mathbb{R}^d} e^{\frac{a}{2}\left(\|\theta\|_2^2+1\right)^{\frac{1}{1+r}} - \frac{\|\theta\|_2^2}{2\sigma^2}}\mathrm{d}\theta \\
&\le \left(2\pi\sigma^2\right)^{-\frac{d}{2}} e^{\frac{a}{2}}\int_{\mathbb{R}^d} e^{-\frac{\|\theta\|_2^2}{2}\left(\frac{1}{\sigma^2}-a\right)}\mathrm{d}\theta,
\end{aligned}$$

若 $\sigma^2 \le \frac{1}{a} = \frac{8}{n\ 1+r\ ()\mathfrak{c}^{\frac{1}{1+r}}}$，则上述积分有限。由于 $c_2 < 1 \le \frac{8L}{(1+r)\mathfrak{c}^{\frac{1}{1+r}},}$ 它保证了$\sigma^2 < \frac{1}{a}$，因此：

$$\begin{aligned}
\psi_{n,d}(0) &\le \left(1 - a\sigma^2\right)^{-\frac{d}{2}} e^{\frac{a}{2}} \\
&\le C_3^d e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}},
\end{aligned}$$

其中 C_3 是一个与 n 和 d 无关的常数。

最后，将 (42) 中 $a_{n,d}$ 和 $b_{n,d}$ 的值代入，可推得：

$$\mathbb{E}\left[e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}(\|\theta_t\|_2^2+1)^{\frac{1}{1+r}}}\right] \le C_1\left(d\log^{2\beta}(n)\right)^{\frac{r}{1+r}} e^{C_2 nd\log^{2\beta}(n)} + C_3^d e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}}, \quad \forall t > 0.$$

其中，C_2 是另一个通用常数，由此完成证明。 □

ii的证明）。我们考虑 $\alpha > 1$，下文中，$C > 0$ 表示一个与 n 和 d无关的“常数”，其值可能随行变化。我们的出发点是 i中得到的指数矩上界。命题 1.1 表明 U_{ν_n} 满足 $\mathcal{H}^{\mathbf{r}}_{\mathbf{KL}}\left(\mathfrak{c}n^{1+r}、\ nL+\ell_0\right)$，因此由命题 4.2 可得：

$$\mathbb{E}[U^{\alpha}_{\nu_n}(\theta_t)] \le \mathbb{E}\left[\left(\min U_{\nu_n} + Cn\|\theta_t - \theta^*_n\|_2^2\right)^{\alpha}\right] \le \mathbb{E}\left[\left(\min U_{\nu_n} + Cn\|\theta^*_n\|_2^2 + Cn\|\theta_t\|_2^2\right)^{\alpha}\right],$$

其中 $\theta_n^* = \arg\min\ U_{\nu_n}$。

利用命题 4.4 以及由詹森不等式 $(a+b)^\beta \le c_\beta(a^\beta+b^\beta)$ 推导出的不等式，对于 $(a,b)\in\mathbb{R}_+^2$ 和 $\beta\ge 1$，我们得到：

$$
\begin{aligned}
&(\min U_{\nu_n} + Cn\|\theta_n^*\|_2^2 + Cn\|\theta_t\|_2^2)^\alpha \\
&\qquad \le C\left[nd\log^{2\beta}(n) + nd^{1+r}\log^{2\beta(1+r)}(n) + n\|\theta_t\|_2^2\right]^\alpha \\
&\qquad \le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + \|\theta_t\|_2^{2\alpha}\right] \\
&\qquad \le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + k^{-\alpha(1+r)}\log^{\alpha(1+r)}\left(e^{k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right] \\
&\qquad \le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + k^{-\alpha(1+r)}\log^{\alpha(1+r)}\left(e^{\alpha(1+r)-1+k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right].
\end{aligned}
$$

詹森不等式以及 $p\ge 1$ 时 $x\mapsto \log^p(x)$ 在 $[e^{p-1},+\infty[$ 上的凹性给出 ld

$$
\begin{aligned}
&\mathbb{E}[U_{\nu_n}^\alpha(\theta_t)] \\
&\le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + k^{-\alpha(1+r)}\mathbb{E}\left[\log^{\alpha(1+r)}\left(e^{\alpha(1+r)-1+k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right]\right] \\
&\le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + k^{-\alpha(1+r)}\log^{\alpha(1+r)}\left[\mathbb{E}\left(e^{\alpha(1+r)-1+k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right]\right] \\
&\le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + k^{-\alpha(1+r)}\left[\alpha(1+r)-1+\log\mathbb{E}\left(e^{k\|\theta_t\|_2^{\frac{2}{1+r}}}\right)\right]^{\alpha(1+r)}\right] \\
&\le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + k^{-\alpha(1+r)}\left[\alpha(1+r)-1+\log\mathbb{E}\left(e^{k(\|\theta_t\|_2^2+1)^{\frac{1}{1+r}}}\right)\right]^{\alpha(1+r)}\right],
\end{aligned}
$$

其中，在最后一个不等式中，我们用到了 $\|\theta\|_2^2 \le \|\theta\|_2^2+1$。

[1]接着，我们应用命题 5.1 中的 i)，取 $k = \frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}$，得到：

$$
\begin{aligned}
&\mathbb{E}[U_{\nu_n}^\alpha(\theta_t)] \\
&\qquad \le Cn^\alpha\left[\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)} + \frac{1}{n^{\alpha(1+r)}}\left[1+\log\mathbb{E}\left(e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}(\|\theta_t\|_2^2+1)^{\frac{1}{1+r}}}\right)\right]^{\alpha(1+r)}\right] \\
&\qquad \le C\left[n^\alpha\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)}\right. \\
&\qquad \left.+\frac{1}{n^{\alpha r}}\left[1+\log\left[C_1\left(d\log^{2\beta}(n)\right)^{\frac{r}{1+r}}e^{C_2nd\log^{2\beta}(n)} + C_3^d e^{\frac{(1+r)n\mathfrak{c}^{\frac{1}{1+r}}}{16}}\right]\right]^{\alpha(1+r)}\right] \\
&\qquad \le Cn^\alpha\left(d\log^{2\beta}(n)\right)^{\alpha(1+r)},
\end{aligned}
$$

其中，在前面的推导中，我们用到了简单的代数运算，以及当 $a\ge 1$和 $b\ge 1$ 时 $\log(a+b)\le\log(2)+\log(a)+\log(b)$。证明至此完成。 □

参考文献

[1] Bakry, D.、Cattiaux, P. 和 Guillin, A.：《遍历连续马尔可夫过程的收敛速率：Lyapunov 与 Poincar´e》。《泛函分析杂志》254 卷第 3 期（2008 年），727–759 页。

[2] Bakry, D. 和 Emery, M.：《超收缩扩散》。S ´eminaire de probabilit ´es 1123, XIX, (1985), 177–206.

[3] Bakry, D.、Gentil, I. 和 Ledoux, M.：《马尔可夫扩散算子的分析与几何》。Springer. 103, (2014).[4] Bobkov, S. G.：《对数凹概率测度的等周与解析不等式》。《概率年刊》27, (1999), 1903–1921.[5] Bolte, J.、Daniilidis, A.、Ley, O. 和 Mazet, L.：《Lojasiewicz不等式的刻画：次梯度流、谷径与凸性》。《美国数学会汇刊》362, (2010), 3319–3363.[6] Cattiaux, P.、Fathi, M. 和 Guillin, A.：《利用变曲率界改进Bakry-Émery判据，以获得Poincar´e不等式和Wasserstein收缩》。Journal de Math ´ematiques Pures et Appliqu ´ees, (2022).[7] Cattiaux, P.、Gentil, I. 和 Guillin, A.：《弱对数 Sobolev 不等式与熵收敛》。《概率论及相关领域》139, 3, (2007), 563–603.[8] Cattiaux, P. 和 Guillin, A.：《命中时间、泛函不等式、李雅普诺夫条件与一致遍历性》。《泛函分析杂志》272, 6, (2017), 2361–2391.[9] Bakry, D.、Cattiaux, P. 和 Guillin, A.：《遍历连续马尔可夫过程的收敛速率：李雅普诺夫条件与Poincar´e不等式》。《泛函分析杂志》254, 3, (2008), 727–759.[10] Cattiaux, P.、Guillin, A.、Wang, F. 和 Wu, L.：《Super Poincar´e不等式的李雅普诺夫条件》。《泛函分析杂志》256, 6, (2009), 1821–1841.[11] Dalalyan, A. 和 Tsybakov, A.：《通过聚合和朗之万蒙特卡洛进行稀疏回归学习》。《计算机与系统科学杂志》，78, 5, (2012), 1423–1443.[12] Dalalyan, A.：《从光滑对数凹密度进行近似采样的理论保证》。《英国皇家统计学会会刊B辑》，79, (2017), 651–676.[13] Dalalyan, A. 和 Karagulyan, A.：《使用不精确梯度进行朗之万蒙特卡洛的实用保证》。《随机过程及其应用》，129, 12, (2019), 5278–5311.[14] Dalalyan, A. 和 Riou-Durand, L.：《使用动理学Langevin扩散从对数凹密度中采样》。《伯努利》，26, 3, 1956–1988.[15] Dalalyan, A.、Karagulyan, A. 和 Riou-Durand, L.：《非强对数凹目标分布的离散化Langevin算法误差界》。《机器学习研究杂志》，23, 235, (2022), 1–38.[16] Durmus, A. 和 Moulines, E.：《通过未校正Langevin算法进行高维贝叶斯推断》，《伯努利》，25, 4A, (2019), 2854–2882.[17] Ethier, S. N. 和 Kurtz, T. G.：《马尔可夫过程——刻画与收敛》，John Wiley & Sons Inc. Wiley概率与数理统计丛书：概率与数理统计，纽约，(1986).[18] Freidlin, M. 和 Wentzell, A.：《动力系统的随机扰动》，Springer Verlag, 1984.[19] Gadat, S. 、 Gavra, I. 和 Risser, L.：《如何计算加权图的重心》。《运筹学数学》，43, 4, (2018).[20] Gadat, S. 和 Panloup, F.：《无强凸性条件下Ruppert-Polyak平均的最优非渐近界》。《随机过程及其应用》，156, (2022), 312–348.[21] Gadat, S.、Panloup, F. 和 Pellegrini, C.：《对数凹模型的贝叶斯后验均值策略的代价》. 预印本，(2022).

[22] Gadat, S. and Panloup, F. and Pellegrini, C.：记忆梯度扩散不变分布的大偏差原理。《电子概率杂志》，81，（2013），1–34。[23] Gross, L.：对数 Sobolev 不等式。《美国数学杂志》，4，97，（1975），1061–1083。[24] Hajeck, B.：最优退火的冷却方案。《运筹学数学》，12，2，（1988），311–329。[25] H¨ormander, L.：亚椭圆二阶微分方程。《数学学报》119，（1967），147–171。[26] Holley, R. and Stroock, D.：通过 Sobolev 不等式实现模拟退火。《数学物理通讯》115，4，（1988），553–569。[27] Khasminskii , R.：微分方程的随机稳定性。《随机建模与应用概率》，Springer，（2012）。[28] Kurdyka, K.：关于 o-极小结构中可定义函数的梯度。《格勒诺布尔傅里叶研究所年刊》48，3，（1998），769–783。[29] Kusuoka, S. and Stroock, D.：Malliavin演算的应用，第一部分。随机分析。Elsevier 32，北荷兰数学丛书，（1984），271–306。[30] Lojasiewicz, S.：实解析子集的一个拓扑性质。法国国家科学研究中心出版社，巴黎，《偏微分方程》。（1963），87–89。[31] Ma, Y. and Chen, Y. and Jin, C. and Flammarion, N. and Jordan, M. I：采样可能比优化更快。《美国国家科学院院刊》116，42，20881–20885。[32] Meyn, S. and Tweedie, R.：马尔可夫链与随机稳定性。Springer Science & Business Media。（2012）。[33] Miclo, L.：模拟退火。关于自由能演化的研究。《IHP概率与统计年刊》28，2，（1992），235–266。[34] Mou, W. and Flammarion, N. and Wainwright, M. J. and Bartlett, P. L.：朗之万扩散离散化的改进界：无需凸性的近最优速率。《伯努利》28，3，（2022），1577–1601。[35] Raginsky, M. and Rakhlin, A. and Telgarsky, M.：通过随机梯度朗之万动力学实现非凸学习：非渐近分析。《机器学习研究论文集》，65，（2017），1–30。[36] Robbins, H. and Monro, S.：一种随机逼近方法。《数学统计年刊》22，3，（1951）：400-407。[37] Wang, F.：本质谱为空的泛函不等式。《泛函分析杂志》170，1，（2000），219–245。[38] Wang, B. and Zou, D. and Gu, Q. and Osher, S. J.：拉普拉斯平滑随机梯度马尔可夫链蒙特卡罗。《SIAM科学计算杂志》43，1，（2021），A26–A53。[39] Welling, M. and Teh, Y. W.：通过随机梯度朗之万动力学进行贝叶斯学习。第28届国际机器学习会议（ICML-11）论文集28，3，（2011），681–688。[40] Xu, P. and Chen, J. and Zou, D. and Gu, Q.：非凸优化中基于朗之万动力学的算法的全局收敛。第32届神经信息处理系统国际会议论文集。NIPS’18。Curran Associates Inc.（2018），3126—3137。