arXiv:2602.01381v1 [cs.CL] 1 Feb 2026

关于（近似）奖励模型在推理时扩展中的作用

Youheng Zhu[1] **Yiping Lu**[1]

[1]西北大学工业工程与管理科学系**通讯作者：** Yiping Lu <yiping.lu@northwestern.edu>

摘要

推理时扩展近来已成为提升大语言模型推理能力的一种强大范式。在各种方法中，序贯蒙特卡洛（*SMC*）已成为一个尤为重要的框架，它能够迭代地生成、评估、拒绝并重采样中间推理轨迹。这一过程的核心组成部分是奖励模型，它负责评估部分解答，并指导推理过程中的计算资源分配。

然而，在实践中，真实的奖励模型从未可用。所有已部署的系统都依赖近似奖励模型，由此引出一个根本问题：为什么以及何时近似奖励模型足以实现有效的推理时扩展？在这项工作中，我们给出了一个理论答案。我们确定，近似奖励模型的贝尔曼误差是决定基于SMC的推理时扩展效果的关键量。对于长度为 T的推理过程，我们证明，如果近似奖励模型的贝尔曼误差上界为 $O(1/T)$，那么将该奖励模型与SMC结合，可将推理的计算复杂度从关于 T 的指数级降至关于 T的多项式级。尽管使用的只是近似奖励，这仍使推理效率得到指数级提升。

1 引言

过去几年里，大语言模型的推理能力得到了显著提升 [1]。尽管模型规模和训练数据仍是关键因素，但近期出现了一种互补范式——推理时扩展，它已成为一种强有力的机制，能够在不修改模型参数的情况下提升推理性能 [2, 3, 4, 5]。这类方法在推理过程中分配额外计算资源，使模型能够探索、优化并验证中间推理步骤，从而在复杂的多步任务上取得显著提升 [1, 6]。推理时扩展通过生成多个候选答案及多样化的推理路径，并利用奖励模型选择质量更高的输出，从而提升模型性能 [7]。这一策略在广泛的推理任务中展现出显著的实证成效。然而，尽管它在实践中效果显著，对于达到特定性能水平需要多少条推理路径，以及支配这种权衡的原则是什么，我们仍缺乏充分的理论认识。在这项工作中，我们迈出第一步，着手研究奖励模型如何影响推理时扩展的效果，并阐明它在引导选择过程中的作用。尽管实证成效显著，推理路径数量、奖励模型质量与整体性能之间的关系仍未得到充分理解。

在现有的推理时扩展方法中，序贯蒙特卡洛（SMC）作为一种原理清晰且灵活的结构化推理框架，正日益受到关注 [8, 9, 10, 11]。在基于 SMC 的推理中，模型会生成多个候选推理轨迹，利用奖励模型评估其质量，并通过迭代执行拒绝和重采样，将计算资源集中于最有希望的路径。该过程使系统能够以可控且有理论依据的方式，在计算量与解答质量之间进行权衡。SMC 成功背后的一个关键因素，是具备能够可靠评估部分推理状态质量的奖励模型。

然而在实际中，这类奖励模型从来都不可能完全准确。它们是基于有限数据、带噪声的人类反馈或启发式信号训练而成，因此不可避免地会偏离真实任务目标。这引出了一个根本性且在很大程度上尚未得到探索的问题：

> 当只有近似奖励模型可用时，为什么推理时扩展仍然有效?

直观来看，人们可能会认为，奖励估计中的微小误差会沿着较长的推理链不断累积，导致 SMC 放大误差，而非纠正误差。然而，实证结果表明，近似奖励模型似乎足以显著提升推理性能。理解这一现象，对于推理时扩展的理论基础以及鲁棒奖励模型的实际设计都至关重要。

1.1 贡献

推理时扩展的一种统一视角是通过扭曲基础模型动力学，实现奖励倾斜采样。最优扭曲对应于前瞻剩余价值函数，但通常难以计算。这促使我们关注近似价值函数，以及其近似质量（通过贝尔曼误差衡量）如何决定采样的计算效率。正文研究的主要算法见表1。具体而言，我们的贡献包括：

- **下界：次指数复杂度需要较小的贝尔曼误差。**在第4节中，我们证明了两个关于推理时采样的指数级信息论下界：一个T 不带中间引导（定理4.1），另一个带有访问近似奖励模型的权限（推论4.2）。特别地，摆脱对

表 1：在假设 3.1 和 3.2 下，正文中主要算法的时间复杂度比较5.2 推论和定理 6.1中的结果。我们考察句子级推理算法。

方法	时间复杂度
贝尔曼误差假设 3.2 ($\varepsilon = O(1/T)$)	
SP-gSMC+MH	$\tilde{O}(LT^3 \log(\delta^{-1}) \log(\delta_{\mathrm{TV}}^{-1}))$
序贯蒙特卡洛（朴素）	$O(L^6 T^2 \delta_{\mathrm{TV}}^{-1})$
VGB†	$\tilde{O}(LT^3 \log(\delta^{-1}) \log(\delta_{\mathrm{TV}}^{-1}))$

†VGB 指的是 [12]。该算法要求预先了解L（$1+\varepsilon$）以选择超参数。这是因为它们使用了拒绝采样（算法3），该算法要求拒绝比率大于分布之间的覆盖率；详见备注B.2。

备注： 未经校正的单粒子引导序贯蒙特卡洛（SP-gSMC）无法生成具有任意精度的样本，除非奖励模型是完美的（定理 4.3）。

推理时域要求处于贝尔曼误差较小 的情形（例如， $\varepsilon = O(1/T)$）

• 推理时扩展中序贯蒙特卡洛的首个粒子数与时间复杂度界限。 序贯蒙特卡洛（SMC）是一种核心的推理时扩展基础方法，实证效果显著。我们在第5 节给出 首个 端到端保证，证明在我们的贝尔曼误差假设 $\delta_{\text{全变差}}$ 精度下，SMC 使用明确的粒子数 N 即可实现总运行时间关于 $(T,\ \delta^{-1}{}_{\text{全变差}})$ 3.2 的条件下。此外，附录中还讨论了灵活选择提议分布和粒子转移动力学的理论与分析，为未来的算法设计提供了有益指导。

• 带有 Metropolis–Hastings 校正的单粒子引导序贯蒙特卡洛的复杂度。 我们通过将一种广泛使用的朴素引导原语形式化为单粒子引导序贯蒙特卡洛更新来研究它；除非奖励模型是完美的，否则该更新本身无法实现任意小的全变差误差（定理4.3）。为了量化实现任意精度所需的计算调节手段，我们为其增加 Metropolis–Hastings 校正，并证明在一个高概率事件上具有全变差距离的几何收缩，这意味着 $O(\log(\delta^{-1}{}_{\text{全变差}}))$ 个 MH 步骤足以达到全变差精度$\delta_{\text{全变差}}$。这与 [12] 中的定性几何混合行为相符，同时不会对算法施加额外的信息要求。

2 预备知识

本节介绍本文分析中使用的主要工具。我们还请读者参阅附录A，其中列出了全文使用的符号说明。

2.1 Metropolis–Hastings

当我们只能将目标分布计算到一个归一化常数的倍数时，Metropolis–Hastings（MH）提供了一种从目标分布中进行采样的通用方法。在本文中，MH 将作为校正步骤使用，把一个易于采样的提议分布转化为一个平稳分布为所需（经奖励倾斜的）目标分布的马尔可夫链。设 $\tilde{\pi}$ 是状态空间 E 上的目标分布。我们假设 $\tilde{\pi}$ 只能通过未归一化评分 $\tilde{\pi}'$ 获得，使得 $\tilde{\pi}(x) \propto \tilde{\pi}'(x)$（归一化常数是未知的）。设 $q(x,\cdot)$ 是一个我们可以高效采样的提议分布。给定当前状态 $x \in \mathsf{E}$，从 $y \sim q(x,\cdot)$ 中抽取一个提议，并以如下概率接受它

$$\alpha(x,\ y) :=\ 1 \wedge \frac{\tilde{\pi}'(y)\ q(y,\ x)}{\tilde{\pi}'(x)\ q(x,\ y)}.$$

若接受，则设置 $X_{n+1} = y$；否则设置 $X_{n+1} = x$。等价地，MH 转移核可写为

$$K(x, dy) := q(x, dy)\,\alpha(x, y) + \left(1 - \int q(x, dy')\,\alpha(x, y')\right)\delta_x(dy),$$

其中， δ_x 表示在 x处的狄拉克测度。根据构造， K 是$\tilde{\pi}$-可逆的，因此诱导出一个以 $\tilde{\pi}$为不变分布的马尔可夫链。

2.2 序贯蒙特卡洛的 Feynman-Kac 模型

序贯蒙特卡洛（SMC）方法是一类基于模拟的算法，它们通过对加权粒子进行重采样和传播来近似一系列概率分布。

费曼–卡茨模型描述了一系列概率分布，这些分布通过迭代地按势函数重新加权、归一化，并通过马尔可夫核进行传播而获得。形式上，它定义了一个概率测度的（离散）流 $\{\eta_t\}_{t\geq 0}$，其中 $\eta_t \in \mathrm{P}(E_t)$ 是定义在（可能随时间变化的）状态空间 E_t 上的分布，通常称为费曼–卡茨流。

从对偶的角度来看，费曼–卡茨模型通过一系列算子刻画分布的演化。这些算子既可以视为作用于测度，也可以等价地通过其对测试函数的伴随作用来描述。具体而言，测度上的算子 P 可由其在测试函数上的伴随算子 P 通过对偶配对 $\mu, f :=\int f\, d\mu$ 唯一确定，即 $\mu P, f\ = \mu, Pf$。因此，对于每个 $t \in [T]$，设 M_t 是从 E_{t-1} 到 E_t 的马尔可夫核，记作 $M_t(x, dy)$，并令其作用于测试函数 $f\colon E_t \to \mathbb{R}$，定义为 $M_t(\ f)(x) := \int f(y) M_t(x, dy)$。令 $G_{t-1}\colon E_{t-1} \to [0,\infty)$ 为势函数。定义作用于测试函数的未归一化费曼–卡茨算子如下：

$$Q_t(f)(x) := G_{t-1}(x)\, M_t(f)(x), \quad Q_{p,n} := Q_{p+1} \cdots Q_n.$$

根据对偶性， Q_t 在测度上的相应作用定义为 μQ_t， $f\ = \mu$， $Q_t f$，因此 Q_t可解释为分布上的一种未归一化“传播–重新加权”更新。从初始分布 $\eta_0 \in \mathrm{P}(E_0)$出发，定义未归一化流 $\gamma_t \in \mathrm{P}(E_t)$和归一化流 η_t如下

$$\gamma_t := \gamma_{t-1} Q_t, \qquad \eta_t := \frac{\gamma_t}{\gamma_t(1)}.$$

直观而言，递推式 $\gamma_t\ = \gamma_{t-1} Q_t$ 对应于测度的重新加权–传播演化：首先，势函数 G_{t-1}在 E_{t-1} 上对 γ_{t-1} 进行倾斜，然后通过马尔可夫核 M_t 将重新加权后的质量推送至 E_t ，从而得到一个未归一化测度。等价地，$\int \gamma_{t-1}\ dx\ G_{t-1}\ x\ M_t\ x\ dy$ (dy) ( ) ( ) (.)。归一化流 $\eta_t\ =$ $\gamma_t\ \gamma_t\ 1$ / ( ) 是经过重新加权和传播后对应的概率律。在此基础上，还可以考虑

算法1 用于 FK 模型的序贯蒙特卡洛（G_t, M_t）

输入： 时域 T，粒子数 N，在 E_1上的初始分布 η_1，势函数 $\{G_{t-1} : E_{t-1} \to \mathbb{R}_+\}_{t=2}^T$，核 $\{M_t : E_{t-1} \to \mathrm{P}(E_t)\}_{t=2}^T$.

独立地为 $i = 1, \ldots, N$采样 $X_1^i \sim \eta_1$。

对于 $t = 2, 3, \ldots, T$ **do**

$w_{t-1}^j \leftarrow G_{t-1}(X_{t-1}^j)$, $\bar{w}_{t-1}^j \leftarrow \frac{w^j}{\sum_{\ell=1}^N w_{t-1}^\ell}$。

对 $A_{t-1}^i \sim$ 分类分布$(\bar{w}_{t-1}^{1:N})$ 进行独立同分布重采样，适用于 $i \in [N]$。

独立地为 $i \in [N]$ 抽取 $X_t^i \sim M_t(X_{t-1}^{A_{t-1}^i}, \cdot)$。

结束循环

从 $I \sim \mathrm{Unif}(\{1, \ldots, N\})$ 中抽取，并且 **输出** X_T^I。

归一化半群，即将归一化步骤纳入一个非线性算子中

$$\Phi_{p,n}(\mu)(f) := \frac{\mu(Q_{p,n}(f))}{\mu(Q_{p,n}(1))}, \qquad (0 \le p \le n \le T),$$

从而有 $\eta_t = \Phi_t(\eta_{t-1})$，其中 $\Phi_t := \Phi_{t-1,\ t}$。

现在，我们使用非线性 FK流$\eta_{p+1} = \Phi_{p+1}(\eta_p)$描述一种序贯蒙特卡洛算法。具体而言，对于每个 $p \in \{0, \ldots, T-1\}$ 以及每个 $\eta \in \mathrm{P}(E_p)$，令 $K_{p+1,\ \eta}$为从 E_p到E_{p+1} 的马尔可夫核，使得

$$\eta K_{p+1,\eta} = \Phi_{p+1}(\eta). \tag{1}$$

序贯蒙特卡洛通过粒子 $\{X_p^i\}_{i=1}^N \subset E_p$及其经验测度 η_p^N: $= \frac{1}{N}\sum_{i=1}^N \delta_{X_p^i}$来模拟 Feyman-Kae 流，相应的相互作用粒子更新通过乘积核对下一代进行采样

$$(X_{p+1}^1, \ldots, X_{p+1}^N) \sim \prod_{i=1}^N K_{p+1,\eta_p^N}(X_p^i, dx_{p+1}^i), \tag{2}$$

在给定 F_p^N 的条件下相互独立，其中 $\mathrm{F}_p^N := \sigma(X_q^i : 0 \le q \le p,\ 1 \le i \le N)$ 是使截至时间 p 的所有粒子均可测的最小 σ-代数。

方程（1）通常有多种 $K_{p+1,\ \eta}$ 的选取方式。一种简单的选择是采用 x-独立核

$$K_{p+1,\eta}(x, dy) = \Phi_{p+1}(\eta)(dy),$$

此时，粒子是从当前费曼-卡茨目标分布$\Phi_{p+1}(\eta_p^N)$中独立同分布抽取的，整个过程可用算法 1 描述。直观地说，根据序贯蒙特卡洛采样器的表述，它可以看作对费曼-卡茨流中重新加权与归一化步骤的基于粒子的离散模拟：它通过对有限个粒子进行重采样，再进行传播，取代（难以处理的）加权测度更新。

3 问题设定

在本文中，我们将推理时推理建模为在推理轨迹上进行的奖励倾斜采样。令 S_0 表示提示空间，令 S_t 表示第 t 步的推理状态空间（词元/句子）。我们用 $\mathrm{S}_{1:t} := \mathrm{S}_{1:t-1} \times \mathrm{S}_t$（其中 $\mathrm{S}_{1:0} := \{\varnothing\}$）表示前缀空间，因此轨迹 $s_{1:T} \in \mathrm{S}_{1:T}$ 由可变长度的提示 $s_0 \in \mathrm{S}_0$ 通过自回归基础模型 π_{ref}（$s_{1:T}|s_0$）$= \prod_{t=1}^T \pi_{\mathrm{ref}}$（$s_t|s_0,\ s_{1:t-1}$）。遵循 [10]，我们将推理时推理视为从在推理轨迹上的能量倾斜后验中进行采样，其中 π_{ref} 作为先验。由此诱导出的目标分布为

$$\tilde{\pi}(s_{1:T}|s_0) \propto \pi_{\mathrm{ref}}(s_{1:T}|s_0)\, \phi(s_{1:T}),$$

其中，π_{ref} 是预训练模型对自然语言推理路径的内在先验，而 $\phi(s)$ 是任务相关效用（或能量）函数（例如人类偏好模型或可验证奖励），它会使推理偏向高质量解答。这种表述通常称为测试时对齐。对于任意 $t < T$，我们还定义中间推理分布

$$\tilde{\pi}_t(s_{1:t}|s_0) \propto \pi_{\mathrm{ref}}(s_{1:t}|s_0)\, \hat{V}(s_0, s_{1:t}),$$

其中 $\hat{V}(s_0, s_{1:t})$用于评估部分推理前缀的质量，而 $\hat{V}(s_0, s_{1:T}) = \phi(s_{1:T})$。我们的目标是为这一系列中间分布拟合一个费曼-卡茨模型，然后便可将先前介绍的算法 1 直接应用于这种语言模型推理场景。

在详细介绍费曼-卡茨模型及其诱导的序贯蒙特卡洛过程之前，我们先简要展示语言模型推理中使用的序贯蒙特卡洛的一般流程：假设样本从提议分布 π_{p} 中抽取，我们维护 N 个粒子来近似 $\tilde{\pi}_{t-1}$，并通过以下两个步骤迭代更新它们，以得到目标 $\tilde{\pi}_t$:

- **重采样：** 给定前缀 $s_{1:1}^{1:N}, \ldots, s_{1:t-1}^{1:N}$，计算权重 $w_{t-1}^{1:N}$，将其归一化为 $\bar{w}_{t-1}^j := \frac{w_{t-1}^j}{\sum_{l=1}^N w_{l,t-1}}$，从 $\mathrm{Cat}(\bar{w}_{t-1}^{1:N})$ 中独立同分布地采样 $A_{t-1}^i \sim$，并令 $\tilde{s}_{1:t-1}^i := s_1^{A_{t-1}^i}, \ldots, s_{t-1}^{A_{t-1}^i}$。
- **传播：** 对每个 $s_t^i \sim \pi_{\mathrm{p}}(\cdot|s_0, \tilde{s}_{1:t-1}^i)$ 采样，并构造 $s_{1:t}^i := (\tilde{s}_{1:t-1}^i, s_t^i)$。

上述过程会迭代执行，直至终止时间步。

费曼-卡茨模型与序贯蒙特卡洛采样器。基于上述思路，我们首先将 $\{\tilde{\pi}_t\}_{t=1}^T$ 用标准的 Feynman–Kac（FK）记法重新表述，以简化表述并便于进行严谨的理论分析。我们设 $E_t := \mathcal{S}_{1:t}$ 并引入一个提议核 $\pi_{\mathrm{p}}(\cdot|s_{1:t})$ 来定义马尔可夫传播$M_t(s_{1:t}, ds_{1:t+1}) := \pi_{\mathrm{p}}(ds_{t+1}|s_{1:t})\delta_{s_{1:t}}(ds_{1:t})$。换言之，序贯蒙特卡洛采样器在每一步都从提议分布 π_{p} 中采样。然后，如 [10],所示，增量权重（势函数）为

$$G_t(s_{1:t}) := \frac{\pi_{\mathrm{ref}}(s_t \mid s_{1:t-1})}{\pi_{\mathrm{p}}(s_t \mid s_{1:t-1})} \cdot \frac{\hat{V}(s_0, s_{1:t})}{\hat{V}(s_0, s_{1:t-1})}.$$

在额外增加一个终止步骤的情况下，即步骤 T+1 采用 E_T 上的恒等马尔可夫核 $M_{T+1} = \mathrm{Id}$（因此 $E_{T+1} = E_T$），并将最后一个势函数取为

$$G_T \qquad \tilde{\pi}_T\ s_{1:T}|s_0 \propto, \text{目标}\ \pi_{\mathrm{ref}}\ s_{1:T}|s_0\ V\ s_0\ s_{1:T}\ (\ )\ (\ )\hat{}\ (\ ,\ )$$

被实现为所诱导FK流的终止分布，且任意中间分布 $\tilde{\pi}_t$ 等于η_t 经 G_t重新加权后的结果。随后，上述“重采样”和“传播”过程可由算法1严格描述，并指定(G_t, M_t)对。由构造可知，在指定(G_t, M_t)时存在一个自由度，即提议分布π_{p}，我们将在附录 D以及附录 E的命题 E.7中进一步讨论。在我们的主要定理中，我们关注最简单、最直接的提议分布选择π_{p}，即朴素提议分布

$\pi_p = \pi_{ref}$。也就是说，序贯蒙特卡洛采样器从预训练模型自身的输出分布中采样。在此情形下，我们有

$$G_t(s_{1:t}) = \frac{\hat{V}(s_0, s_{1:t})}{\hat{V}(s_0, s_{1:t-1})}.$$

具体而言，在附录 B 中，我们将算法1 应用于算法*5* 中的朴素提议分布。

最优扭曲函数。[10] 表明，能够恢复奖励倾斜分布 $V^*(s_0, s_{1:t})$中间边缘分布的扭曲函数$\tilde{\pi}_{s_{1:t}}(s_{1:t}|s_0) := \sum_{s_{t+1:T}} \tilde{\pi}(s_{1:T}|s_0)$ 应当是以下贝尔曼方程的解

$$\begin{aligned} V^*(s_0, s_{1:t}) &\propto \sum_{s_{t+1:T}} \pi_{\text{ref}}(s_{t+1:T}|s_0, s_{1:t})\phi(s_{1:T}) \\ &\propto \sum_{s_{t+1}} \pi_{\text{ref}}(s_{t+1}|s_0, s_{1:t})V^*(s_0, s_{1:t+1}). \end{aligned} \tag{3}$$

直观而言，扭曲函数起到“前瞻”评估器的作用：给定当前的部分轨迹（s_0, $s_{1:t}$），它会根据参考策略，对所有可能的未来延续及其最终奖励进行加权平均，以估计这个前缀有多好。由于该量衡量的是当前状态基于未来结果的预期价值，因此它自然对应于一个价值函数。

然而在实践中，计算最优扭转函数$V^*(s_0, s_{1:t})$ 是难以处理的，因此通常采用近似方法来学习近似奖励模型。其中一些方法包括：

•对比扭曲学习（CTL）。 [10] 提出了对比扭曲学习，通过匹配目标前缀边缘分布来学习按时间索引的前缀扭曲（剩余奖励）$\{\hat{V}_t^\theta\}_{t=1}^T$。定义扭曲前缀模型 $\tilde{\pi}_t^\theta(s_{1:t}|s_0) \propto \pi_{\text{参考}}(s_{1:t}|s_0)\hat{V}_t^\theta(s_0, s_{1:t})$，CTL 求解 $\min_\theta \sum_{t=1}^T D_{\text{KL}}(\tilde{\pi}_t(s_{1:t}|s_0)\|\tilde{\pi}_t^\theta(s_{1:t}|s_0))$，从而得到对比式（正–负）更新，其期望通过重要性采样 / 序贯蒙特卡洛 [10]进行估计。

• **软强化学习**。回顾一下，最优扭转满足一个递归贝尔曼方程，该方程由（3）定义。因此，软强化学习类方法 [13] 拟合一个价值函数/扭曲函数，使其满足由奖励倾斜目标分布所蕴含的软贝尔曼递归关系（或路径一致性约束 [14]）。这一思路已应用于语言模型文献，例如 [15, 16, 17, 18]。

• 密度比扭曲学习（例如 SIXO）。SIXO [19] 将扭曲学习视为通过噪声对比（二元）分类进行密度比估计。最优前缀扭曲可写成后验与先验之比 $V_t^*(s_0, s_{1:t}) \propto \frac{\tilde{\pi}_t(s_{1:t}|s_0)}{\pi\text{ref}(s_{1:t}|s_0)}$，其中$\tilde{\pi}_t$ 是由奖励倾斜目标分布$\tilde{\pi}(s_{1:T}|s_0) \propto \pi\text{ref}(s_{1:T}|s_0)\phi(s_0, s_{1:T})$诱导的 t步边缘分布。SIXO 拟合一个分类器，用于区分正前缀 $s_{1:t} \sim \tilde{\pi}_t(\cdot|s_0)$ 和负前缀 $s_{1:t} \sim \pi\text{ref}(\cdot|s_0)$；最优对数几率可恢复 $\log V_t^*(s_0, s_{1:t})$，从而得到学习到的扭曲函数。

•CD-FUDGE。[20] 提出的方法采用函数逼近类算法，对目标进行直接回归：$\min_{V\in\mathcal{V}} \sum_{t=1}^{T-1} \mathbb{E}_{\pi_{\text{ref}}(\cdot|s_0,s_{1:t})}\left[(V(s_0, s_{1:t}) - \phi(s_{1:T}))^2\right]$。这等价于针对任意给定前缀，使用蒙特卡洛求解一个同策略评估问题。在这种情况下，误差可能由函数类不足、有限样本误差和优化差距导致。

与贝尔曼误差假设的联系。 这些目标可以通过其隐式控制的贝尔曼误差来理解。软强化学习 / 路径一致性针对沿 FK流的前缀剩余奖励的 局部 （单步）一致性，与假设 3.2相符。相比之下，CD-FUDGE 通过蒙特卡洛展开直接回归终端势函数 $\phi(s_{1:T})$，从而更自然地与附录 H中的 全局 假设集合（假设 H.1）保持一致。通过双重采样，FUDGE式回归也可能被重新表述为局部贝尔曼误差目标。

尽管在实证上取得了显著成功，但推理时计算量（采样的推理路径数量）、奖励模型质量与最终性能之间的相互作用仍缺乏充分理解。具体而言，现有方法表明，即使奖励仅为近似奖励，采样更多轨迹也能带来显著提升；然而，尚不清楚这种近似在何时足以实现有效扩展，以及改进奖励模型何时确实能够进一步提升性能——这些洞见对于高效分配建模工作和推理资源至关重要。为了刻画奖励模型，以及衡量模型对最优扭转的近似程度这一指标，我们提出以下两个假设。

假设 3.1（奖励模型比值界）。 我们假设存在某个 $L > 0$,

$$\max_{t\in[T]} \max_{s_{1:t}\in\mathcal{S}_{1:t}} \max\left\{\frac{\hat{V}(s_0, s_{1:t-1})}{\hat{V}(s_0, s_{1:t})}, \frac{\hat{V}(s_0, s_{1:t})}{\hat{V}(s_0, s_{1:t-1})}\right\} \le L.$$

假设 3.2（贝尔曼误差一致界）。 我们假设存在某个 $\varepsilon > 0$，满足以下条件：

$$\max_{t\in[T]} \max_{s_{1:t}\in\mathcal{S}_{1:t}} \max\left\{\frac{\hat{V}(s_0, s_{1:t})}{\mathbb{E}_{s_{t+1}\sim\pi_{\text{ref}}(\cdot|s_0,s_{1:t})}[\hat{V}(s_0, s_{1:t+1})]}, \frac{\mathbb{E}_{s_{t+1}\sim\pi_{\text{ref}}(\cdot|s_0,s_{1:t})}[\hat{V}(s_0, s_{1:t+1})]}{\hat{V}(s_0, s_{1:t})}\right\} \le 1 + \varepsilon$$

注释 3.3。 贝尔曼误差刻画了奖励模型的质量：理想的奖励模型应满足最优扭曲条件（3），因此其贝尔曼误差为零，二者的差异自然也由贝尔曼误差本身刻画。我们将在定理 4.3 中证明，当贝尔曼误差为零时，也就是奖励模型完美时，单粒子引导序贯蒙特卡洛是精确的。此外，这里采用的假设形式是局部贝尔曼误差，另一种可考虑的假设是全局贝尔曼误差， [12]采用了这种假设。关于后者的讨论见附录H。

我们还对本文讨论的算法作出说明。

注释 3.4（词元级推理与句子级推理）。 在词元级推理中，每一步都会追加一个词元，因此基础模型的概率 $\pi_{\text{ref}}(s_t|s_0, s_{1:t-1})$ 可由变换器明确给出，并且能够精确计算。在句子级推理中，单步会追加一个较长的文本片段（整句或片段），作为一个“推理步骤”；通常无法精确计算此类文本片段在 π_{ref} 下的概率，通常只能通过采样获取。为句子级推理设计的算法可以原样应用于

词元级推理（将每个“句子步骤”视为一个词元即可）。本文始终聚焦于句子级设定。

4 有无近似奖励模型时的下界

本节中，我们建立关于时域呈指数级增长的采样时间复杂度下界，涵盖有和没有近似奖励模型的情况。这些下界指出，贝尔曼误差较小的情形是唯一能够实现次指数复杂度的条件，这与单粒子引导式序贯蒙特卡洛误差分析中得出的要求一致。稍后在第5节的备注5.3中，我们将下界与上界进行比较。

4.1 信息论下界

为了单独考察根本性限制，我们采用一种最简接口，将所有算法细节抽象掉。具体而言，我们允许采样器（i）在给定任意前缀的情况下，从基础模型中抽取下一个词元的样本；以及（ii）针对前缀查询奖励模型。由此得到的下界仅取决于这些信息访问方式，因此适用于任何推理时采样策略。我们通过预言机模型形式化这种访问模式及相应的下界结果。

预言机模型。固定 $B \in \mathbb{N}$ 和时域 T。令 $U := \bigcup_{t=0}^{T-1} [B]^t$ 为前缀集合。预言机是一个表 $o = \{o(u)[k]\}_{u \in U,\ k \geq 1} \in [B]^{U \times \mathbb{N}}$。在 $\mathrm{O}_{B,\ T} := [B]^{U \times \mathbb{N}}$ 上赋予乘积 σ-代数。给定每个 $u \in U$ 上的 $\pi_{\mathrm{ref}}(\cdot \mid u)$，令 $\mathbb{P}_{\pi_{\mathrm{ref}}}$ 为乘积测度，其中，对于每个固定的 u，$\{o(u)[k]\}_{k \geq 1}$ 独立同分布，分布为 $\pi_{\mathrm{ref}}(\cdot \mid u)$，且在不同的 u 之间相互独立。

输入与目标。输入为 $I = B,\ \pi_{\mathrm{ref}},\ \hat{V}$，其中 $\hat{V}$ 是定义在前缀上的奖励模型，且 $\hat{V}(\varnothing) = 1$。它在叶节点上诱导出目标分布 $p_I(s_{1:T}) \propto \pi_{\mathrm{ref}}(s_{1:T})\, \hat{V}(s_{1:T})$。

算法与输出分布。确定性算法 a 接收 $(B, \hat{V})$ 并可访问 o 的预言机，能自适应地查询有限个坐标 $o(u)[k]$，并输出 $\widehat{S}_{1:T} \in [B]^T$。等价地，它诱导出一个可测映射 $\mathsf{a} : \hat{\mathrm{V}} \times \mathbb{N} \times \mathrm{O}_{B,T} \to [B]^T$。随机化算法 A 是此类 a 的一个分布 $R \in \Delta(\mathrm{A})$。令 A 表示满足不得猜测 约束的算法：$\widehat{\ }$每当 a 输出 $S_{1:T}$ 时，它必须沿该路径查询过，$\widehat{\ }$即对所有 $p \leq T$，每个前缀 $S_{1:p}$ 都满足此条件。对于 $A \in \Delta(\mathrm{A})$ 和输入 I，定义输出分布

$$q_{A,I}(\cdot) := \mathbb{E}_{\mathsf{a} \sim R}\left[\mathbb{P}_{o \sim \mathbb{P}_{\pi_{\mathrm{ref}}}}\left(\mathsf{a}(\hat{V}, B, o) \in \cdot\right)\right].$$

定理 4.1（LB1：最坏情况时间复杂度）。任何随机化算法 $A \in \mathcal{A}$ 若对每个满足假设 *3.1* 且具有 $L > 1$ 的输入 $I = \langle B, \pi_{\mathrm{ref}}, \hat{V}\rangle$，其输出叶节点分布满足 $\|q_{A,I} - p_I\|_{\mathrm{TV}} \leq \frac{1}{3}$，并且 A 总是在时间复杂度TC之前停止，则 $\mathrm{TC} = \Omega\left(L^{2T/3}\right)$。

推论4.2（LB2：Bellman 误差界下的最坏情况复杂度）。任意随机化算法 $A \in \mathcal{A}$ 满足：对于每个输入 $I = \langle B, \pi_{\mathrm{ref}}, \hat{V}\rangle$ 满足假设 *3.1* 和 *3.2* 其中 $L > 1$ 且 $\varepsilon \in (0, L-1]$，输出叶分布满足 $\|q_{A,I} - p_I\|_{\mathrm{TV}} \leq \frac{1}{3}$；若 A 始终在时间复杂度TC之前停止，则 $\mathrm{TC} = \Omega\left((1+\varepsilon)^{2T/3}\right)$。

证明概要。详情见附录C。我们构造一个困难实例族 $\{I_u\}$，使得对每个 u，目标叶分布 p_{I_u} 在某个事件 $A_u \subseteq [B]^T$（最大概率叶节点的集合）上分配了 $> 1/2$ 的概率质量。因此，$\|q_{A,\ I_u} - p_{I_u}\|_{\mathrm{TV}} \leq 1/3$ 要求任何有效算法都必须以常数概率命中 A_u。根据 Yao 的极小极大原理，只需对随机困难输入 $U \sim \mathrm{Unif}([B]^{2T/3})$，界定任意确定性算法 a 的平均成功概率。在时间预算 Q 和禁止猜测约束下，a 最多只能“尝试” Q 个候选叶节点/前缀路径，因此根据并集界，$\Pr[\mathsf{a} \text{ 命中 } A_U] \leq Q/B^{2T/3}$。所以，要达到常数成功概率，就必须有 $Q = \Omega(B^{2T/3})$，由此得到 $\mathrm{TC} = \Omega(L^{2T/3})$（在假设 3.2 下则为 $\Omega((1+\varepsilon)^{2T/3})$）。□

4.2 我们能从单粒子引导序贯蒙特卡洛中学到什么

一种常见且朴素的推理时引导基本操作，会执行局部的、逐步的奖励倾斜采样，并在不同领域中以不同名称出现，包括动作级采样 [20, 12]，局部约束解码 [11,21, 22, 23, 24, 25]，以及扩散引导中的基于价值的解码 / *SVDD* [26, 27]。为避免术语过多，我们将这一基本操作称为单粒子引导序贯蒙特卡洛(SP-gSMC)，因为它恰好是使用以下最优提议分布的序贯蒙特卡洛采样器的 $N = 1$ 特例（附录 D）：$\pi_{\mathrm{P}}(s_{t+1} \mid s_0, s_{1:t}) \propto \pi_{\mathrm{ref}}(s_{t+1} \mid s_0, s_{1:t})\, \hat{V}(s_0, s_{1:t+1})$。

令 $\hat{\pi}_t(\cdot \mid s_0)$ 表示通过依次采样 $s_{1:t}$ 而得到的联合分布，其中 $s_k \sim \hat{\pi}(\cdot \mid s_0, s_{1:k-1})$ 对应于 $k = 1, \ldots, t$。因此，单粒子引导序贯蒙特卡洛可视为最优提议序贯蒙特卡洛更新的单轨迹类比，将基于粒子的引导与第6节稍后介绍的基于链的方法联系起来。我们还将在备注 D.3中进一步讨论单粒子引导序贯蒙特卡洛相较于其多粒子对应方法的效率，前提是贝尔曼误差足够小。

我们有如下关于单粒子引导序贯蒙特卡洛的全变差误差控制。

定理 4.3（SP-gSMC 全变差误差）. 假定假设 *3.2* 成立。则对每个 $t \in [T]$，$\left\|\hat{\pi}_t(\cdot \mid s_0) - \tilde{\pi}_t(\cdot \mid s_0)\right\|_{\mathrm{TV}} \leq 2t\varepsilon$.

该结果表明，在朴素引导下，单粒子引导序贯蒙特卡洛输出分布 $\hat{\pi}_T$ 仅当贝尔曼误差满足 $\varepsilon < \frac{1}{2T}$（等价地，$\varepsilon = O(1/T)$）时，才会更接近目标 $\tilde{\pi}_T$（即达到 $D_{\mathrm{TV}}(\hat{\pi}_T, \tilde{\pi}_T) < 1$）。一旦 ε 超过该阈值，朴素引导就会过弱，无法在全变差上产生有意义的改进。此外，若目标采样精度为 δ_{TV}，则当 $\varepsilon \geq \frac{\delta_{\mathrm{TV}}}{2T}$ 时，朴素单粒子引导序贯蒙特卡洛无法达到所需的全变差误差，这促使我们考虑朴素引导之外的其他算法。

贝尔曼误差较小的 $\varepsilon = O(1/T)$ 情形也与推论 4.2 中的指数下界相吻合，这表明该情形对于保证次指数复杂度至关重要；更多细节请参见备注 5.3 。

5 基于序贯蒙特卡洛的推理时扩展复杂度

本节分析序贯蒙特卡洛采样器的计算复杂度，考察其与粒子数、时域以及马尔可夫传播和奖励评估成本之间的关系。该分析基于大语言模型推理时扩展情形，在这种情形下，性能取决于奖励模型评估的次数以及轨迹搜索的范围。

5.1 序贯蒙特卡洛复杂度

序贯蒙特卡洛采样器的基本思想如第 3 节所述，即使用粒子系统来模拟费曼-卡茨流。我们设计了一个带终止步骤的费曼-卡茨模型，使其费曼-卡茨流在步骤 $T+1$ 处与我们的目标分布一致，即 $\eta_{T+1} = \bar{\pi}_T$。相应地，我们将序贯蒙特卡洛采样器的采样分布定义为$\hat{\pi}_T := \bar{\eta}^N_{T+1} := \mathbb{E}[\eta^N_{T+1}]$，也就是说，我们在这次终端重采样步骤之后 输出一个粒子。具体而言，使用朴素提议分布的序贯蒙特卡洛算法详见算法 5。作为研究时间复杂度的第一步，我们提出以下关于粒子复杂度的定理，即序贯蒙特卡洛采样器要以所需精度进行采样，必须维持的粒子数。

定理 5.1（粒子复杂度）。固定时域 $T \geq 2$ 和目标精度 $\delta_{\mathrm{TV}} \in (0,\ 1)$。若假设 *3.1* 和*3.2* 在 L、$\varepsilon > 0$下成立，令 $\bar{\eta}^N_{T+1} := \mathbb{E}[\eta^N_{T+1}]$ 为序贯蒙特卡洛采样器的输出分布（在使用 $M_{T+1} = \mathrm{Id}$执行终端重采样步骤 $T+1$ 之后）。那么，在采用朴素提议分布时，以下充分条件可保证$\|\bar{\eta}^N_{T+1} - \eta_{T+1}\|_{\mathrm{TV}} \leq \delta_{\mathrm{TV}}$。也就是说，粒子数 $N \geq \frac{L^6 T}{2\,\delta_{\mathrm{TV}}}(1+\varepsilon^{\,6(T-1)})$。

该定理的证明使用以下定理E.6以及引理 E.2。详细证明见附录 E。

我们还给出定理 D.1 和命题 E.7，将粒子复杂度结果推广到任意提议分布。请注意，粒子复杂度 N 自然地导出序贯蒙特卡洛式算法的 NT 型基准成本，因为必须在 N 个步骤中传播 T 粒子（每一步都执行重采样/重新加权）。额外开销取决于每步以下操作的成本：(i) 从提议分布 π_{p} 中采样；(ii) 计算重要性比率 $\pi_{\mathrm{ref}}/\pi_{\mathrm{p}}$（以及重新加权时使用的任何辅助量）。因此，这些结果使我们能够灵活地权衡提议分布的难度、重要性比率的计算以及所需的粒子数；参见备注E.8。

尽管更复杂的提议分布可能会带来额外的计算开销π_{p}，但朴素提议 SMC 可以直接将粒子数转换为时间复杂度：每个 T 步骤需要 $O(N)$的工作量，因此定理 5.1立即推出以下时间复杂度推论。

推论 5.2（朴素提议 SMC 的时间复杂度）。 在假设 *3.1* 和*3.2*下，并考虑$\varepsilon = O(1/T)$这一情形。采用朴素提议的 *SMC* 采样器要达到 δ_{TV} 精度，即 $\|\bar{\pi}_T - \hat{\pi}_T\| \leq \delta_{\mathrm{TV}}$，其时间复杂度为$O\left(\frac{L^6 T^2}{\delta_{\mathrm{TV}}}\right)$。

有了相匹配的算法保证，我们现在可以将采用朴素提议的 SMC 采样器所实现的复杂度上界与我们的下界所刻画的最坏情况限制联系起来。

备注 5.3（下界与上界）。 我们的 LB2 表明，对于近似奖励/价值模型（由贝尔曼误差界 ε刻画），任何采样器要达到常数全变差精度，最坏情况时间复杂度仍至少为 $= \Omega((1+\varepsilon)^{2T/3})$。这凸显了近似的关键优势：指数底数变为 $1+\varepsilon$，随着奖励模型的改进，该值可以任意接近 1，即使比率界常数 L 很大也是如此。相比之下，若没有这种近似（只有比率界），目前可得的最佳下界仍为 $\Omega(L^{2T/3})$。

反过来，我们的上界也达到了这一有利的基数（至多相差常数次幂）：粒子复杂度定理 5.1 给出的 N 缩放为 $(1+\varepsilon)^{O(T)}$（乘以 $\mathrm{Poly}(T, L, \delta_{\mathrm{TV}}^{-1})$），与 LB2 相比，指数上的差距至多为常数因子。对于采用 Metropolis-Hastings 校正的单粒子引导序贯蒙特卡洛（或 SP-gSMC+MH，见定理 6.1 和 G.2 的证明），同样的定性 $(1+\varepsilon)^{O(T)}$ 依赖关系也成立；我们将在下一节介绍这种方法。在 $\varepsilon = O(1/T)$ 的情形下，$(1+\varepsilon)^{O(T)} = O(1)$，因此这些指数因子可以被吸收，总体复杂度关于 $(T, \delta_{\mathrm{TV}}^{-1})$ 呈多项式增长。

5.2 关于 *McKean* 核的选择。

我们的全变差保证由定理 E.6 推导得出，该定理具有 核无关的性质：它给出了一个一致上界，对任意满足一致性条件 $K_{p,\eta}$ 的 McKean 转移核 $\eta K_{p+1,\eta} = \Phi_{p+1}(\eta)$ 均成立，而我们此前的所有结果都建立在这种最坏情形控制之上。然而，同一个 FK 模型可以对应许多不同的 McKean 核，其粒子系统可能表现出不同的二阶行为（因此，在结构化情形下，收敛速率可能有所提升）。为求完整，附录 E 给出了 核敏感 的改进结果（命题 E.9），并通过一个具体示例加以说明：一种基于分位数的分层转移（推论 E.10），它可以视为一种特殊的分层规则，专门适配于指定函数的分位数。分层重采样方案在粒子滤波文献中由来已久；例如，参见 [28]。

6 BeyondSMC：基于链的方法

在本节中，我们将讨论另一类广受欢迎的推理时扩展算法 [26, 27, 21, 22, 23, 24, 25, 12]，它不同于基于 SMC 的采样。这类算法以单粒子引导的 SMC 为起点，采用回溯 [12] 或 Metropolis-Hastings [29] 等多种方法来校正有偏采样器。这些校正会诱导马尔可夫链向目标分布混合，使采样分布任意接近目标分布。这种观点将采样视为一个混合问题，而非纯粹的蒙特卡洛平均问题；后者的精度主要受方差控制支配。因此，对于所需全变差精度 δ_{TV}，这种方法会带来性质上截然不同的依赖关系。总体而言，我们考虑构建经 MH 校正的 SP-gSMC 采样器的两种方法。本节重点讨论 SP-gSMC 中基于重采样的提议，这与扩散引导中的 SVDD [26, 27] 类似。另一种方法详见附录G。

算法 2 SP-gSMC M-重采样 MH 校正

输入： T, H, M, $\hat{V}, \pi_{\text{ref}}, s_0, \delta$.
初始化 $(s_{1:T}^{\text{acc}}, w^{\text{acc}}) \leftarrow (\varnothing, 1)$。
对于 $h = 1, \ldots, H$ **do**
  初始化 $(s_{1:T}, w) \leftarrow (\varnothing, 1)$。
  对于 $t = 1, \ldots, T$ **do**
    样本 $s_t^{(j)} \overset{iid}{\sim} \pi_{\text{ref}}(\cdot|s_0, s_{1:t-1})$, $j = 1, \ldots, M$.
    $v_t^{(j)} \leftarrow \hat{V}(s_0, (s_{1:t-1}, s_t^{(j)}))$, $\bar{Z}_t \leftarrow \frac{1}{M}\sum_{j=1}^{M} v_t^{(j)}$。
    绘制 $A_t \sim$分类分布$\left(\frac{v_t^{(1:M)}}{\sum_{\ell=1}^{M} v_t^{(\ell)}}\right)$；设置 $s_t \leftarrow s^{(A_t)}{}_t$。
    $w \leftarrow w \cdot \frac{\hat{V}(s_0,\ s_{1:t})}{\bar{Z}_t}$.
  结束循环
  如果 $h = 1$**则**
    $(s_{1:T}^{\text{acc}},\ w^{\text{acc}}) \leftarrow (s_{1:T},\ w)$；**继续**
  结束条件判断
  抽取 $\alpha \sim$ 均匀分布$[0, 1]$。
  如果 $\alpha < 1 \wedge \frac{w^{\text{acc}}}{w}\frac{\hat{V}(s_{1:T})}{\hat{V}(s_{1:T}^{\text{acc}})}$ **then**
    $(s_{1:T}^{\text{acc}},\ w^{\text{acc}}) \leftarrow (s_{1:T},\ w)$。
  结束 if
结束 for
输出： $s_{1:T}^{\text{acc}}$。

6.1 基于重采样的 *SP-gSMC* 采样与 *Metropolis-Hastings* 校正

重采样池 MH。在第 t 步，理想的 SP-gSMC 提议会对样本 $s_t \sim \pi_{\text{ref}}(\cdot|s_0, s_{1:t-1})$ 按局部评分 $\hat{V}(s_0, s_{1:t})$ 进行倾斜。我们使用大小为 M 的有限重采样池来近似这一步：抽取 M 个独立同分布的候选项 $s_t^{(1:M)} \sim \pi_{\text{ref}}(\cdot|s_0, s_{1:t-1})$，使用奖励模型计算其评分 $v^{(j)}{}_t$，构造经验归一化常数 $\bar{Z}_t = \frac{1}{M}\sum_{j=1}^{M} v_t^{(j)}$，再选择索引$A_t \sim$分类分布$(v^{(1:M)}{}_t / \sum_{\ell=1}^{M} v^{(l)}{}_t)$，并令 $s_t = s^{(A_t)}{}_t$。尽管当 $M < \infty$ 时，s_t 的边缘分布只是理想倾斜分布的近似，但整个随机化过程在包含候选池和索引变量的增广空间上具有精确已知的概率。因此，目标（经奖励倾斜的）路径分布与该重采样池提议之间的似然比可以用闭式形式表示：沿着一条提议轨迹，我们可以更新 $w \leftarrow w \cdot \hat{V}(s_0, s_{1:t}) / \bar{Z}_t$，因此最终的 w 是由用 Z_t 替换 Z_t 所诱导的精确拉东–尼科迪姆修正。随后，我们在轨迹末尾使用这一精确比率执行标准的 MH 接受–拒绝步骤，从而得到一个有效的 MH 核，其不变分布正比于 $\pi_{\text{ref}}(s_{1:T}|s_0)\,\hat{V}(s_0, s_{1:T})$。算法 2 总结了总计 H 次 MH 迭代的完整过程。

因此，第一个结果表明，尽管使用了有限的重采样池 M，我们仍能在 δ_{TV} 中保持指数快速的混合行为。

定理 6.1。对于任意 $0 < \delta \lesssim \delta_{\text{TV}}$，使用时间 $\tilde{O}\left(LT^3 \log\left(\delta^{-1}\right) \log\left(\delta_{\text{TV}}^{-1}\right)\right)$，存在一个良好事件 $\mathcal{E}$，其满足 $\mathbb{P}(\mathrm{E}) \geq 1 - \delta$；在此条件下，算法 *2* 达到 δ_{TV}精度，即 $\|\tilde{\pi}_T - \hat{\pi}_T|_{\mathcal{E}}\|_{\text{TV}} \leq \delta_{\text{TV}}$。

证明见附录F。

为何这是可行的。关键在于，重采样过程允许在增广空间上进行精确的概率计算：尽管当 $M < \infty$ 时，词元上的边际提议分布仅是近似的，但算法会通过 Z_t 的池依赖似然比来进行跟踪，因此 MH 接受比率是精确的。在假设 3.1–3.2 下，在一个良好事件上，接受概率进一步被限制在远离 0 和 1 的范围内，这意味着存在常数阶的 Dobrushin 收缩系数。因此，只需 $H = O(\log(\delta_{\text{TV}}^{-1}))$ 个 MH 步骤即可。最后，由于我们只需要常数阶的收缩，因此只需将每个 Z_t 估计到相对误差 $O(1/T)$，使用 $M = O(T^2)$ 个样本即可实现。

6.2 与多粒子序贯蒙特卡洛的关系及总结

单粒子引导序贯蒙特卡洛的视角阐明了，基于链的方法与粒子方法共享相同的底层更新：二者都实现了一个单粒子序贯蒙特卡洛步骤，主要区别在于如何 利用额外计算来降低误差。这一视角为基于粒子的引导与基于链的引导之间提供了便利的桥梁：前者通过增加粒子数来降低误差，后者则通过马尔可夫链混合来降低误差。序贯蒙特卡洛可以通过扩展粒子数实现任意δ_{TV} 精度，通常会导致对 δ_{TV}^{-1}的多项式依赖。相比之下，Metropolis–Hastings 校正通过混合来逼近相同的极限：相关的计算调节参数是 MH 步骤数 H。当接受概率是精确的，并且在高概率良事件上具有一致界时，所得核在全变差意义下呈几何收缩，从而得到 $H = O\left(\log(\delta_{\text{TV}}^{-1})\right)$。我们的重采样池 MH 构造保留了这种对数依赖关系，而附录G 中讨论的基于拒绝的构造通常会因接受率估计误差的线性累积而失去这一性质。

参考文献

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou 等。思维链提示激发大语言模型的推理能力。神经信息处理系统进展，35:24824–24837，2022年。[2] Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese 等。大语言模型推理前沿综述：推理扩展、推理学习与智能体系统。*arXiv* 预印本 *arXiv:2504.09037*，2025年。[3] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le 等。由浅入深提示使大语言模型能够进行复杂推理。载于第十一届国际学习表征会议，2023年。[4] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song 和 Jacob Steinhardt。使用 MATH 数据集衡量数学问题求解能力。载于第三十五届神经信息处理系统大会数据集与基准赛道（第二轮），2021。

[5] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever 和 Karl Cobbe。让我们逐步验证。载于第十二届国际学习表征会议，2023年。[6] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery 和 Denny Zhou。自洽性提升语言模型中的思维链推理。载于第十一届国际学习表征会议，2023年。[7] Qiyuan Liu, Hao Xu, Xuhong Chen, Wei Chen, Yee Whye Teh 和 Ning Miao。利用奖励模型增强大语言模型推理：分析性综述。*arXiv* 预印本 *arXiv:2510.01925*，2025年。[8] Shengyu Feng, Xiang Kong, Shuang Ma, Aonan Zhang, Dong Yin, Chong Wang, Ruoming Pang 和 Yiming Yang。通过扭曲序贯蒙特卡洛实现数学问题的逐步推理。载于第十三届国际学习表征会议，2025年。[9] Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B Tenenbaum 和 Yejin Choi。面向大语言模型的假设驱动的心智理论推理。*arXiv* 预印本 *arXiv:2502.11881*，2025年。[10] Stephen Zhao, Rob Brekelmans, Alireza Makhzani 和 Roger Baker Grosse。通过扭曲序贯蒙特卡洛实现语言模型中的概率推断。载于国际机器学习会议，第60704–60748页。PMLR，2024年。[11]João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner 等。通过序贯蒙特卡洛对大语言模型进行句法与语义控制。载于国际学习表征会议，2025年。[12] Dhruv Rohatgi, Abhishek Shetty, Donya Saless, Yuchen Li, Ankur Moitra, Andrej Risteski 和 Dylan J Foster。驾驭不完美过程验证器：回溯的采样视角。*arXiv* 预印本 *arXiv:2510.03149*，2025年。[13] Sergey Levine。作为概率推断的强化学习与控制：教程与综述。*arXiv* 预印本 *arXiv:1805.00909*，2018年。[14] Ofir Nachum, Mohammad Norouzi, Kelvin Xu 和 Dale Schuurmans。弥合基于价值与基于策略的强化学习之间的差距。神经信息处理系统进展，30，2017年。[15] Vasileios Lioutas, Jonathan Wilder Lavington, Justice Sefas, Matthew Niedoba, Yunpeng Liu, Berend Zwartsenberg, Setareh Dabiri, Frank Wood 和 Adam Scibior。评论器序贯蒙特卡洛。国际学习表征会议，2023年。[16] Sidharth Mudgal, Jong Lee, Harish Ganapathy, Yaguang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman 等。基于语言模型的受控解码。载于国际机器学习会议，第36486–36503页。PMLR，2024年。

[17] Han Guo、Bowen Tan、Zhengzhong Liu、Eric Xing 和 Zhiting Hu。《数据良好样本有限条件下用于文本生成的高效（软）Q学习》。发表于计算语言学协会论文集：*EMNLP 2022*年会议论文集，第 6969–6991 页，2022 年。[18] Edward J Hu、Moksh Jain、Eric Elmoznino、Younesse Kaddar、Guillaume Lajoie、Yoshua Bengio 和 Nikolay Malkin。《在大语言模型中摊销难以处理的推理》。发表于第十二届国际学习表征会议，2024 年。[19] Dieterich Lawson、Allan Raventós、Andrew Warrington 和 Scott Linderman。《SIXO：使用扭曲目标进行平滑推理》。神经信息处理系统进展，35:38844–38858，2022 年。[20] Kevin Yang 和 Dan Klein。《FUDGE：利用未来判别器进行可控文本生成》。发表于*2021*年计算语言学协会北美分会人类语言技术会议论文集，第 3511–3535 页，2021 年。[21] Richard Shin、Christopher Lin、Sam Thomson、Charles Chen Jr、Subhro Roy、Emmanouil Antonios Platanios、Adam Pauls、Dan Klein、Jason Eisner 和 Benjamin Van Durme。《受约束语言模型可生成少样本语义解析器》。发表于*2021*年自然语言处理实证方法会议论文集，第 7699–7715 页，2021 年。[22] Torsten Scholak、Nathan Schucher 和 Dzmitry Bahdanau。《PICARD：对语言模型的受约束自回归解码进行增量解析》。*arXiv* 预印本 *arXiv:2109.05093*，2021 年。[23] Gabriel Poesia、Oleksandr Polozov、Vu Le、Ashish Tiwari、Gustavo Soares、Christopher Meek 和 Sumit Gulwani。《Synchromesh：基于预训练语言模型的可靠代码生成》。发表于国际学习表征会议，2022 年。[24] Brandon T Willard 和 Rémi Louf。《面向大语言模型的高效引导生成》。*arXiv* 预印本 *arXiv:2307.09702*，2023 年。[25] Shubham Ugare、Tarun Suresh、Hangoo Kang、Sasa Mis- ailovic 和 Gagandeep Singh。《SynCode：通过语法增强实现大语言模型生成》。机器学习研究汇刊，2024 年。[26] Xiner Li、Yulai Zhao、Chenyu Wang、Gabriele Scalia、Gok- cen Eraslan、Surag Nair、Tommaso Biancalani、Shuiwang Ji、Aviv Regev、Sergey Levine 等。《通过软价值解码在连续和离散扩散模型中进行无导数引导》。*arXiv* 预印本 *arXiv:2408.08252*，2024 年。[27] Masatoshi Uehara、Yulai Zhao、Chenyu Wang、Xiner Li、Aviv Regev、Sergey Levine 和 Tommaso Biancalani。《扩散模型中的推理时对齐与奖励引导生成：教程与综述》。*arXiv* 预印本 *arXiv:2501.09685*，2025 年。[28] Randal Douc 和 Olivier Cappé。《粒子滤波中重采样方案的比较》。发表于*ISPA 2005*：第四届国际图像与信号处理及分析研讨会论文集，*2005*年。，第 64–69 页。IEEE，2005 年。

[29] Aayush Karan 和 Yilun Du。《基于采样的推理：你的基础模型比你想象的更聪明》。*arXiv* 预印本 *arXiv:2510.14901*，2025 年。[30] Dylan J Foster、Zakaria Mhammedi 和 Dhruv Rohatgi。《高效强化学习是否需要优质基础模型？基础模型在探索中的计算作用》。*arXiv* 预印本 *arXiv:2503.07453*，2025 年。[31] Alekh Agarwal、Daniel Hsu、Satyen Kale、John Langford、Lihong Li 和 Robert Schapire。《驯服怪兽：一种快速简洁的上下文老虎机算法》。发表于国际机器学习会议，第 1638–1646 页。PMLR，2014 年。[32] Pierre del Moral 和 Emmanuel Rio。《平均场粒子模型的集中不等式》。应用概率年刊，21(3):1017–1052，2011 年。

目录

1. 引言 .1**2. 预备知识** .2**3. 问题设定** .3**4. 有无近似奖励模型时的下界** .5

定理 4.1 .5推论 4.2 .5定理 4.3 .5

5. 基于SMC的推理时间扩展复杂度 .6
定理 5.1 .6推论 5.2 .6备注 5.3 .6

6. 超越 SMC：基于链的方法 .6定理 6.1 .7**A. 符号说明** .12**B. 算法** .14

定理 B.1 .14注 B.2 (关于拒绝采样版本及一项特殊要求的注释).14Freedman 引理 B.3 .15引理 B.4 及证明.15引理 B.5 及证明.16

C. 第 4 .20
定理 *4.1* 的证明.20推论 *4.2* 的证明.21引理 C.1 及 证明 .21定理 *4.3* 的证明.22

D. 最优提议序贯蒙特卡洛 .23
定理 D.1 (最优提议序贯蒙特卡洛的粒子复杂度).23注 D.2 (最优提议分布的优缺点).23注 D.3 (粒子数： $o(T)$ 比 1差.
23命题 D.4 (最优提议序贯蒙特卡洛的时间复杂度).23注 D.5 (最优提议序贯蒙特卡洛的低效率).23

E. 第 5 节和附录 D 中的证明.24 引理 E.1 及其证明（费曼-卡茨模型的一阶稳定性）.24 引理 E.2 及其证明（提议指定的 FK 常数）.26 引理 E.3 及其证明（测度望远镜求和）.27 引理 E.4 及其证明（单步局部偏差与方差）.27 引理 E.5 及其证明（时域放大的局部方差）.28 定理 E.6 及其证明（序贯蒙特卡洛总变差偏差界）.29 定理 5.1 和定理 D.1 的证明.29 命题 E.7 及其证明（任意提议分布下的粒子复杂度）.30 注 E.8（选择提议分布时的权衡）.30 命题 D.4 的证明.30 命题 E.9 及其证明（给定麦基恩核下的序贯蒙特卡洛总变差偏差界）.34

推论 E.10 及 证明 （分层核示例）.35备注 E.11 （从独立同分布 McKean 转移到分层重采样）.37**F. 第 6** 节中的证明.38引理 F.1 及 证明 .38定理 *6.1* 的证明.38

预印本——论（近似）奖励模型对推理时间扩展的基 **G. 基于拒绝的 SP-gSMC 与 Metropolis-Hastings 校正** .42

引理 G.1 及证明.42定理G.2 及证明.42备注G.3 （比较定理6.1 与G.2）.44

H。新假设：全局贝尔曼误差 .45命题 H.2 和证明.45命题 H.3 和证明.46

A 符号说明

符号	含义
推理／语言模型设置	
T	时域（推理步骤数）。
$[T]$	集合 {1、2、.....、T}。
S_t / S_0	时间 t 处的推理／令牌空间/ 初始提示空间。
$\mathsf{S}_{1:t}$	截至时间 t 的前缀空间，递归地 $\mathsf{S}_{1:t} := \mathsf{S}_{1:t-1} \times \mathsf{S}_t$。
$s_{1:t}$ / s_0	$\mathsf{S}_{1:t}$ / 初始提示中的一个前缀（轨迹）。
$\pi_{\text{ref}}(\cdot \mid s_{1:t-1})$	参考/提议策略（例如，基础语言模型）。
$\pi_{\text{p}}(\cdot \mid s_{1:t-1})$	提议分布，采样器从该分布中抽取样本。（朴素情形为 $\pi_{\text{p}} = \pi_{\text{ref}}$）
$\phi(s_{1:T})$	依赖于任务的最终奖励函数。
$\hat{V}(s_0, \ s_{1:t})$	定义在前缀上的（不完美）奖励/价值模型。 $\hat{V}(s_0, s_{1:T}) = \phi(s_{1:T})$
$\tilde{\pi}(s_{1:T}\|s_0)$	倾斜目标分布 $\tilde{\pi}(s_{1:T}\|s_0) \propto \pi_{\text{ref}}(s_{1:T}\|s_0)\ \phi(s_{1:T})$， $\tilde{\pi}_T(s_{1:T}\|s_0) = \tilde{\pi}(s_{1:T}\|s_0)$。
$\tilde{\pi}_t(s_{1:t}\|s_0)$	倾斜中间分布 $\tilde{\pi}_t(s_{1:t}\|s_0) \propto \pi_{\text{ref}}(s_{1:t}\|s_0)$ $\hat{V}(s_0, s_{1:t})$， $\tilde{\pi}_T(s_{1:T}\|s_0) = \tilde{\pi}(s_{1:T}\|s_0)$.
$\hat{\pi}_T$	任意指定算法的采样分布。对于序贯蒙特卡洛，$\hat{\pi}_T := \bar{\eta}^N_{T+1} := \mathbb{E}[\eta^N_{T+1}]$。
L	奖励模型比值界（假设 3.1）。
ε	贝尔曼误差界（假设 3.2）。
C_{act}	局部单步比值常数： $C_{\text{act}} := \sup_{t\in[T]} \sup_{s_{1:t}\in\mathsf{S}_{1:t}} \dfrac{\hat{V}(s_0, \ s_{1:t})}{\mathbb{E}_{s_t\sim\pi_{\text{ref}}(\cdot\|s_{1:t-1})}\hat{V}(s_0, \ s_{1:t})}$, $C_{\text{act}} \le L(1+\varepsilon)$ 在假设3.1 和3.2下。
δ_{TV}	采样分布 $\hat{\pi}_T$ 与目标分布 $\tilde{\pi}_T$之间期望的全变差误差。
信息论下界	
B	抽象词元空间的分支因子（字母表大小） $[B]$。
U	前缀集合： $U := \bigcup_{t=0}^{T-1} [B]^t$。
$o(u)\ [k]$	与前缀 $u \in U$相关的第 k次预言机下一个词元抽样。
$O_{B,T}$	预言空间 $\mathsf{O}_{B,\ T} := [B]^{U\times\mathbb{N}}$ （配备乘积 σ-代数）。
$\mathbb{P}_{\pi_{\text{ref}}}$	由基础模型 $\pi_{\text{ref}}(\cdot \mid u)$ 在 $\mathsf{O}_{B,T}$ 上诱导的乘积测度。
I	输入实例 $I = B, \pi_{\text{ref}},$ $\hat{V}$（带有 $\hat{V}(\varnothing) = 1$）。
p_I	目标叶分布 $p_I(s_{1:T}) \propto \pi_{\text{ref}}(s_{1:T})$ $\hat{V}(s_{1:T})$。
$\mathcal{A}$	满足不猜测约束的（确定性）算法类。
$q_{A,I}$	随机化算法 A 在输入 I 上的输出分布。
费曼–卡茨模型	
E_t	时刻 t 的费曼–卡茨状态空间；在我们的语言模型设定中 $E_t := \mathsf{S}_{1:t}$。
$\mathrm{P}(E)$	可测空间 E 上的概率测度。
$M_t(x, \ dy)$ / $M_t(f)(x)$	从 E_{t-1} 到 E_t（提议动力学）的马尔可夫核，其中 $M_t(f)(x) = \int f(y)\ M_t(x, \ dy)$。
$G_t : \ E_t \to [0,\infty)$	势函数 / 增量权重函数（通常源自 $\hat{V}$).
Q_t	未归一化费曼–卡茨算子： $Q_t(f)(x) := G_{t-1}(x)\ M_t(f)(x)$。
$Q_{p,n}$	多步未归一化半群： $Q_{p,\ n} := Q_{p+1}\cdots Q_n$。
η_t	E_t 上的极限 FK 流（目标）分布。
Φ_t	归一化 FK 映射： $\eta_t = \Phi_t(\eta_{t-1})$；更一般地， $\Phi_t(\mu)(f) = \dfrac{\mu(Q_t(f))}{\mu(Q_t(1))}$.
$\Phi_{p,n}$	归一化半群： $\Phi_{p,n}(\mu)(f) := \dfrac{\mu(Q_{p,n}(f))}{\mu(Q_{p,n}(1))}$。（所以 $\Phi_{p,n} = \Phi_n \circ \cdots \circ \Phi_{p+1}$。）
粒子系统（序贯蒙特卡洛）	
N	粒子数。
$\xi_t^{(N,i)}$	时刻 t的 N 个粒子中的第 i个。
η_t^N	时刻 t 的粒子经验测度（ η_t的随机近似）。
$\bar{\eta}_t^N$	均值测度 $\bar{\eta}_t^N := \mathbb{E}[\eta_t^N]$。
$\mathcal{F}_t^N$	由粒子系统截至时刻 t所生成的自然滤过。
ν_t^N	单步预测测度 $\nu_t^N := \Phi_t(\eta_{t-1}^N)$。
混合／稳定性系数与振荡范数	
振幅(f)	振荡： $\sup f - \inf f$。
振幅$_1(E)$	单位振荡类： $\{f: E \to \mathbb{R} : \text{振幅}(f) \le 1\}$。
$\beta(K)$	积分算子/核 K 的 Dobrushin 系数（用于控制收缩）。

符号	含义
FK 一阶展开（用于附录）	
$D_\eta\Phi$	基于基点 η 的一阶积分算子，其中 $(D_\eta\Phi)(1) = 0$。
$\beta(D\Phi)$	一致 Dobrushin 界：$\beta(D\Phi) := \sup_\eta \beta(D_\eta\Phi) < \infty$。
$R^\Phi(\mu, \eta)$	上述展开中的二阶余项符号测度。
FK 半群专用辅助量	
$P_{p,n}$	由 $Q_{p,n}$ 诱导的马尔可夫核：$P_{p,n}(f) := \dfrac{Q_{p,n}(f)}{Q_{p,n}(1)}$。
$G_{p,n,\eta}$	归一化势：$G_{p,n,\eta} := \dfrac{Q_{p,n}(1)}{\eta(Q_{p,n}(1))}$。
$q_{p,n}$	比率常数：$q_{p,n} := \sup_{x,y\in E_p} \dfrac{Q_{p,n}(1)(x)}{Q_{p,n}(1)(y)} \in [1,\infty]$。

B 算法

B.1 拒绝采样

我们将拒绝采样作为一种基本方法，从倾斜分布中抽取（条件）精确样本

$$\nu(\mathrm{d}z) \propto g(z)\,\mu(\mathrm{d}z),$$

其中，μ 是易于处理的基准分布（例如，$\pi_{\text{ref}}(\cdot|s_0, s_{1:t})$），而 g 是非负评分（例如奖励）。从概念上讲，拒绝采样将未归一化权重 g 转化为从 ν 进行精确采样，只需使用来自 μ 的提议分布，并作出接受或拒绝的决策。

算法 3 拒绝采样 [12]

输入： 函数 $g : \mathrm{Z} \to \mathbb{R}_{\geq 0}$，基分布 μ，阈值 $M > 0$，失败概率 δ **输出：** 样本 $\hat{z} \in$ Z 设置 $N \leftarrow 4M\log(4/\delta)$ 抽取 $z_1, \ldots, z_N \sim \mu$ 独立同分布 计算 $\widehat{Z} \leftarrow \frac{1}{N}\sum_{i=1}^{N} g(z_i)$ **对于** $i = 1$ 到 N **执行** 抽取 $z \sim \mu$ 抽取 $\xi \sim$ 伯努利分布$\left(\min\left\{\frac{g(z)}{M\widehat{Z}},\ 1\right\}\right)$ **如果** $\xi = 1$ **则** **返回** z **结束如果** **结束循环** 抽取 $z \sim \mu$ **返回** z {failure case}

—

关键保证在于，尽管该算法使用经验归一化因子 Z，但其输出在一个高概率事件上是精确的。以下定理对此进行了刻画。

定理 *B.1*（[30]中的定理 *E.1*）。令 $g : Z \to [0, \infty)$ 和 $\mu \in \Delta(Z)$，并定义 $\nu(dz) := g(z)\,\mu(dz)$，其中 $Z := \mathbb{E}_{z\sim\mu}[g(z)] \in (0, \infty)$。令 $C_\infty := \left\|\frac{g}{Z}\right\|_\infty$，固定 $\delta \in (0, 1)$，并假设 $M \geq 4C_\infty$。使用 (g, μ, M, δ) 运行算法 3，并令 $\hat{z}$ 为其输出。那么存在一个事件 E^{accept}，其满足 $\mathbb{P}(E^{\text{accept}}) \geq 1 - \delta$，使得 $\mathbb{P}(\hat{z} \in \cdot\ | E^{\text{accept}}) = \nu(\cdot)$。

注释 B.2。算法 3 是教科书式拒绝采样的一种截断、自归一化变体：它用经验估计值 Z 替代未知的归一化常数 $Z = \mathbb{E}_\mu[g]$，并且最多运行 N 次接受–拒绝试验（若失败则回退到 $z \sim \mu$）。尽管进行了这些修改，但在高概率事件上，输出仍然是精确的（定理 B.1）：在给定 $\mathrm{E}^{\text{accept}}$ 的条件下，返回的样本服从分布 ν。值得注意的是，算法 3 有一个超参数 M，其值必须大于或等于 $4C_\infty$。否则，拒绝采样在数学上无法成立。因此，必须对 C_∞ 有近似了解，才能执行该算法。如果已精确知道这样的 C_∞，那么理想的总体运行时间为 $O(C_\infty \log(\delta^{-1}))$。

假设已知 C_∞ 的上界近似估计，记为 $C_\infty^\star$。为进行采样而执行算法 3 的总运行时间为 $O(C_\infty^\star \log(\delta^{-1}))$，其中超参数 $M = 4C_\infty^\star$取一个保守值。错误的估计可能导致 $C_\infty^\star \gg C_\infty$，造成计算资源浪费。

当算法还要求将归一化常数估计到较小的相对误差 ϵ时，我们可以使用相同的预算 $N = \Theta(M\log(\delta^{-1}))$，将这项估计与算法 3结合起来，既用于构造 $\widehat{Z}$，也用于运行截断接受–拒绝试验。在此情形下，主要成本来自归一化估计，通常需要 $N = \tilde{O}(C_\infty\epsilon^{-2})$。等价地，这对应于选择 $M = \tilde{O}(C_\infty\epsilon^{-2})$；当 ϵ 足够小时（例如算法 6或算法 8 中），这会自动推出 $M \geq 4C_\infty$，因此条件精确性的包络要求不会带来额外限制。特别是，为了确保拒绝采样器在目标精度范围内定义良好，无需另外提供先验上界 $C_\infty^\star$ 关于 C_∞。因此，前述 $C_\infty^\star \gg C_\infty$ 问题不再构成障碍。

B.2 经验期望

算法 4 经验期望

输入： 函数 $g : \mathrm{S}_{1:t} \to \mathbb{R}_{\geq 0}$，预训练 moel π_{ref}，当前状态轨迹 $s_{1:t}$，蒙特卡罗总步数 N_m。
采样 N_m 个独立样本 $s_t^{[i]}$，$i \in [N_m]$ 从 $\pi_{\text{ref}}(\cdot|s_0, s_{1:t-1})$。
返回 $\frac{1}{N_m}\sum_{i=1}^{N_m} g(s_t^{[i]})$

引理 B.3 (Freedman 引理，例如 [31])。 设 $(X_t)_{t\le T}$是关于滤过 $(\mathrm{F}_t)_{t\le T}$适应的实值鞅差序列。若 $|X_t| \le R$几乎必然成立，则对任意 $\eta \in (0,\ 1/R)$，以至少 $1-\delta$的概率，

$$\sum_{t=1}^{T} X_t \le \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t^2] + \frac{\log(\delta^{-1})}{\eta}.$$

其中 $\mathbb{E}_t[\cdot] :\ = \mathbb{E}[\cdot|\mathrm{F}_t]$。

引理 B.4。 设 $(X_t)_{t\le T}$为一个适应序列，且几乎处处有 $0 \le X_t \le R$。记作

$$\mu_t := \mathbb{E}_{t-1}[X_t], \qquad D_t := X_t - \mu_t.$$

那么，对于任意 $\epsilon \in (0,\ 1)$和 $\delta \in (0,\ 1)$，以概率至少 $\ge 1-\delta$，

$$\sum_{t=1}^{T} X_t \le (1+\epsilon) \sum_{t=1}^{T} \mu_t + \frac{R}{\epsilon} \log \frac{4}{\delta}, \tag{4}$$

$$\sum_{t=1}^{T} \mu_t \le \frac{1}{1-\epsilon} \sum_{t=1}^{T} X_t + \frac{R}{\epsilon(1-\epsilon)} \log \frac{4}{\delta}. \tag{5}$$

特别地，当 $\epsilon = 1/2$且我们放宽常数时，这将恢复原始形式，其系数为 3/2 和 2 ，相差一个乘法常数。

证明。定义

$$\mu_t := \mathbb{E}_{t-1}[X_t], \qquad D_t := X_t - \mu_t.$$

那么，(D_t) 是一个鞅　　差序列（相对于 (F_t)），并且由于 $0 \le X_t \le R$，我们有 $|D_t| \le R$。此外　　r

$$\mathbb{E}_{t-1}[D_t^2] = \mathrm{Var}(X_t \mid \mathcal{F}_{t-1}) \le \mathbb{E}_{t-1}[X_t^2] \le R\mathbb{E}_{t-1}[X_t] = R\mu_t. \tag{$*$}$$

固定 $\epsilon \in (0,1)$ 和 $\delta \in (0,1)$。我们将对 (D_t) 和 $(-D_t)$ 应用引理 A.2，选取合适的参数，然后取并集界。

步骤 1：给出 $\sum X_t$ 的上界。对 (D_t)应用**引理 A.2**

$$\eta = \frac{\epsilon}{R}, \qquad \delta' = \frac{\delta}{2}.$$

注意，由于 $\epsilon \in (0,1)$， $\eta \in (0,1/R)$。因此，引理 A.2 表明，以至少 $1-\delta/2$ 的概率，

$$\sum_{t=1}^{T} D_t \ \le\ \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[D_t^2] + \frac{\log((\delta/2)^{-1})}{\eta}.$$

利用 $(*)$ 以及 η 的选取，

$$\eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[D_t^2] \ \le\ \eta R \sum_{t=1}^{T} \mu_t = \epsilon \sum_{t=1}^{T} \mu_t.$$

Also

$$\frac{\log((\delta/2)^{-1})}{\eta} = \frac{R}{\epsilon} \log \frac{2}{\delta} \le \frac{R}{\epsilon} \log \frac{4}{\delta}.$$

因此，在事件 E_1 且 $\Pr(\mathrm{E}_1) \ge 1-\delta/2$ 时，

$$\sum_{t=1}^{T} D_t \le \epsilon \sum_{t=1}^{T} \mu_t + \frac{R}{\epsilon} \log \frac{4}{\delta}.$$

由于 $\sum X_t = \sum \mu_t + \sum D_t$，我们得到

$$\sum_{t=1}^{T} X_t \le (1+\epsilon) \sum_{t=1}^{T} \mu_t + \frac{R}{\epsilon} \log \frac{4}{\delta},$$

这恰好是（4）。

步骤 2：控制 $\sum \mu_t$ 用 $\sum X_t$。 现在将引理 A.2 应用于鞅差序列　　ce

$$Y_t := -D_t = \mu_t - X_t,$$

再次使用相同的 $\eta = \epsilon/R$ 和 $\delta' = \delta/2$。我们有 $|Y_t| \le R$ 和

$$\mathbb{E}_{t-1}[Y_t^2] = \mathbb{E}_{t-1}[D_t^2] \le R\mu_t$$

与之前一样，因此引理 A.2 表明，在事件 E_2 上，$\Pr(\mathrm{E}_2) \ge 1 - \delta/2$，

$$\sum_{t=1}^{T} Y_t \le \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[Y_t^2] + \frac{\log((\delta/2)^{-1})}{\eta} \le \epsilon \sum_{t=1}^{T} \mu_t + \frac{R}{\epsilon} \log \frac{4}{\delta}.$$

由于 $\sum Y_t = \sum \mu_t - \sum X_t$，这变为

$$\sum_{t=1}^{T} \mu_t - \sum_{t=1}^{T} X_t \le \epsilon \sum_{t=1}^{T} \mu_t + \frac{R}{\epsilon} \log \frac{4}{\delta},$$

i.e.

$$(1-\epsilon) \sum_{t=1}^{T} \mu_t \le \sum_{t=1}^{T} X_t + \frac{R}{\epsilon} \log \frac{4}{\delta}.$$

两边除以 $1-\epsilon$，得到

$$\sum_{t=1}^{T} \mu_t \le \frac{1}{1-\epsilon} \sum_{t=1}^{T} X_t + \frac{R}{\epsilon(1-\epsilon)} \log \frac{4}{\delta},$$

这就是 (5)。

步骤 3：并集界。最后，定义 $\mathrm{E} := \mathrm{E}_1 \cap \mathrm{E}_2$。根据并集界，

$$\Pr(\mathcal{E}) \ge 1 - \Pr(\mathcal{E}_1^c) - \Pr(\mathcal{E}_2^c) \ge 1 - \frac{\delta}{2} - \frac{\delta}{2} = 1 - \delta.$$

在 $\mathcal{E}$ 上，(4) 和 (5) 均成立，因此引理得证。 □

引理 B.5. 设 $N_m \ge 8C_{\mathrm{act}} \log \frac{4}{\delta}$。以大于 $1-\delta$ 的概率，使用 *4* 算法并采用 $g := \hat{V}$ 估计期望的相对误差为

$$1 - 2\sqrt{2}\sqrt{\frac{C_{\mathrm{act}} \log \frac{4}{\delta}}{N_m}} \le \frac{\tilde{V}^{N_m}}{\mathbb{E}_{s_t \sim \pi_{\mathrm{ref}}(\cdot|s_{1:t-1}, s_0)}[\hat{V}(s_0, s_{1:t})]} \le 1 + 2\sqrt{2}\sqrt{\frac{C_{\mathrm{act}} \log \frac{4}{\delta}}{N_m}},$$

where $\tilde{V}^{N_m} := \frac{1}{N_m} \sum_{i=1}^{N_m} \hat{V}(s_0, (s_{1:t-1}, s_t^{[i]}))$ *and* $C_{\mathrm{act}} := \sup_{t \in [T]} \sup_{s_{1:t} \in \mathcal{S}_{1:t}} \frac{\hat{V}(s_0, s_{1:t})}{\mathbb{E}_{s_t \sim \pi_{\mathrm{ref}}(\cdot|s_{1:t-1})} \hat{V}(s_0, s_{1:t})}$.

引理 *Le* 的证明 *mma B.5.* 对任意固定前缀 $s_{1:t-1}$，应用引理 B.4，以至少 $1-$ 的概率可得 δ，

$$\tilde{V}^{N_m} \le (1+\epsilon)\mathbb{E}[\hat{V}(s_0, s_{1:t})] + \frac{\sup_{s_t} \hat{V}(s_0, s_{1:t})}{N_m \epsilon} \log \frac{4}{\delta},$$

$$(1-\epsilon)\mathbb{E}[\hat{V}(s_0, s_{1:t})] \le \tilde{V}^{N_m} + \frac{\sup_{s_t} \hat{V}(s_0, s_{1:t})}{N_m \epsilon(1-\epsilon)} \log \frac{4}{\delta}.$$

因此，相对误差满足：

$$1 - \epsilon - \frac{C_{\mathrm{act}}}{N_m \epsilon(1-\epsilon)} \log \frac{4}{\delta} \le \frac{\tilde{V}^{N_m}}{\mathbb{E}[\hat{V}(s_0, s_{1:t})]} \le 1 + \varepsilon + \frac{C_{\mathrm{act}}}{N_m \epsilon} \log \frac{4}{\delta},$$

并平衡误差后，对于任意 $N_m \ge 8C_{\mathrm{act}} \log \frac{4}{\delta}$

$$1 - 2\sqrt{2}\sqrt{\frac{C_{\mathrm{act}} \log \frac{4}{\delta}}{N_m}} \le \frac{\tilde{V}^{N_m}}{\mathbb{E}_{s_t \sim \pi_{\mathrm{ref}}(\cdot|s_{1:t-1}, s_0)}[\hat{V}(s_0, s_{1:t})]} \le 1 + 2\sqrt{2}\sqrt{\frac{C_{\mathrm{act}} \log \frac{4}{\delta}}{N_m}}$$

□

算法 5 朴素提议序贯蒙特卡洛（使用 $V(s_0) := 1$，从 η_1 开始，终止步骤 $t = T+1$ 使用 $M_{T+1} = \mathrm{Id}$） 1: 输入：提示 s_0，时域 T，粒子数 N，基础模型 π_{ref}，评分 V。 2: 约定：$V(s_0) := 1$ 且 $V(S_{1:0}) := V(s_0) = 1$。 3: 初始化 ($t = 1$)：对 $i = 1, \ldots, N$，独立同分布地采样 $S_1^i \sim \eta_1(\cdot) = \pi_{\mathrm{ref}}(\cdot \mid s_0)$。 4: 对 $t = 2, 3, \ldots, T$ 执行 5: 权重：$w_{t-1}^j \leftarrow V(S_{1:t-1}^j)\, V(S_{1:t-2}^j)$，$\bar{w}_{t-1}^j \leftarrow w_{t-1}^j / \sum_{\ell=1}^N w_{t-1}^\ell$。 6: 重采样：对 $i = 1, \ldots, N$，独立同分布地从 $\mathrm{Cat}(\bar{w}_{t-1}^{1:N})$ 中采样 $A_{t-1}^i \sim$。 7: 传播（朴素提议分布）：采样 $S_t^i \sim \pi_{\mathrm{ref}}(\cdot \mid s_0, S_1^{A_{t-1}^i:t-1})$，设置 $S_{1:t}^i \leftarrow (S_1^{A_{t-1}^i:t-1}, S_t^i)$。 8: 结束循环 9: 终止步骤 ($t = T+1, M_{T+1} = \mathrm{Id}$)：$w_T^j \leftarrow V(S_{1:T}^j)\, V(S_{1:T-1}^j)$，$\bar{w}_T^j \leftarrow w_T^j / \sum_{\ell=1}^N w_T^\ell$。 10: 重采样：对 $i = 1, \ldots, N$，独立同分布地从 $\mathrm{Cat}(\bar{w}_T^{1:N})$ 中采样 $A_T^i \sim$。 11: 恒等传播：对所有 $i = 1, \ldots, N$，设置 $S_{1:T}^i \leftarrow S_1^{A_T^i:T}$。 12: 从 $\mathrm{Unif}(\{1, \ldots, N\})$ 中抽取 $I \sim$ 并输出 $S_{1:T}^I$。

算法 6 最优提议序贯蒙特卡洛（拒绝采样提议 + 蒙特卡洛重加权；终止步骤 $t = T{+}1$ 与 $M_{T+1} = \mathrm{Id}$）

输入： 提示 s_0、时域 T、#粒子 N、蒙特卡洛预算 N_m、基础模型 π_{ref}、评分 V、拒绝采样阈值 M、拒绝采样失败概率 δ。 **约定：** $V(s_0) := 1$ 和 $V(S_{1:0}) := 1$。

通过拒绝采样初始化 ($t = 1$)： 对每个 $i = 1, \ldots, N$，设定

$$S_1^i \leftarrow \textsc{RejectionSampling}\Big(g_1(\cdot) = V(\cdot),\ \mu = \pi_{\text{ref}}(\cdot \mid s_0),\ M,\ \delta/O(NT)\Big),$$

其中 拒绝采样 是算法3。

对于 $t = 2、3、\ldots、T$，**执行**

对于 $j = 1, \ldots, N$ **执行**

Z_{t-1} **在前缀** $S^j_{1:t-2}$ **上的蒙特卡洛估计：**

$$\widehat{Z}_{t-1}(S^j_{1:t-2}) \leftarrow \frac{1}{N_m}\sum_{m=1}^{N_m} V\big(S^j_{1:t-2}, \tilde{s}^{(j,m)}_{t-1}\big), \quad \tilde{s}^{(j,m)}_{t-1} \sim \pi_{\text{ref}}(\cdot \mid s_0, S^j_{1:t-2}) \text{ i.i.d.}$$

权重（前缀回移一位）：

$$w^j_{t-1} \leftarrow G_{t-1}(S^j_{1:t-1}) := \frac{\widehat{Z}_{t-1}(S^j_{1:t-2})}{V(S^j_{1:t-2})}.$$

结束循环

归一化 $\bar{w}^j_{t-1} \leftarrow w^j_{t-1}/\sum_{\ell=1}^N w^\ell_{t-1}$。

重采样： $A^i_{t-1} \sim$ 分类分布$(\bar{w}^{1:N}_{t-1})$ 对 $i = 1, \ldots, N$ 独立同分布。

通过 RS 传播（最优提议分布）： 对每个 $i = 1, \ldots, N$，设置

$$S^i_t \leftarrow \textsc{RejectionSampling}\Big(g_t(\cdot) = V(S^{A^i_{t-1}}_{1:t-1}, \cdot),\ \mu = \pi_{\text{ref}}(\cdot \mid s_0, S^{A^i_{t-1}}_{1:t-1}),\ M,\ \delta/O(NT)\Big),$$

并设置 $S^i_{1:t} \leftarrow (S_1^{A^i_{t-1}:\,t-1}, S^i_t)$。

结束循环

终止步骤（$t = T{+}1$, $M_{T+1} = \mathrm{Id}$）：

对于 $j = 1, \ldots, N$ **执行**

对前缀 S^j_1 **上** Z_T **的蒙特卡洛估计**：$T-1$**：**

$$\widehat{Z}_T(S^j_{1:T-1}) \leftarrow \frac{1}{N_m}\sum_{m=1}^{N_m} V\big(S^j_{1:T-1}, \tilde{s}^{(j,m)}_T\big), \quad \tilde{s}^{(j,m)}_T \sim \pi_{\text{ref}}(\cdot \mid s_0, S^j_{1:T-1}) \text{ i.i.d.}$$

终端权重：

$$w^j_T \leftarrow G_T(S^j_{1:T}) := \frac{\widehat{Z}_T(S^j_{1:T-1})}{V(S^j_{1:T-1})}.$$

结束循环将 $\bar{w}^j_T \leftarrow w^j_T/\sum_{\ell=1}^N w^\ell_T$ 归一化。 **重采样：** $A^i_T \sim$ 分类分布$(\bar{w}^{1:N}_T)$，对 $i = 1、\ldots\ldots、N$ 独立同分布地抽样。

恒等传播： 对所有 $i = 1、\ldots\ldots、N$，设置 $S^i_{1:T} \leftarrow S_1^{A^i_T:T}$。 从 $\{1、\ldots\ldots、N\}$ 中抽取 $I \sim$ 均匀分布，并**输出** $S^I_{1:T}$。

算法 7 SP-gSMC M-采用 Metropolis–Hastings 校正的重采样：详细版

输入： 时域 T、Metropolis–Hastings 步数 H、重采样批次大小 M、奖励模型 $\hat{V}$、基础模型 π_{ref}、初始提示 s_0、失败概率 δ。
初始化已接受轨迹 $s_{1:T}^{\text{acc}} \leftarrow \varnothing$ 及其权重 $w^{\text{acc}} \leftarrow 1$。
对于 $h = 1, \ldots, H$ **do**
  初始化提议轨迹容器 $s_1^{[h]:T} \leftarrow \varnothing$ 和权重 $w^{[h]} \leftarrow 1$.
  对于 $t = 1, \ldots, T$ **do**
    Draw M 候选项 $\{s_t^{(h,j)}\}_{j=1}^{M} \overset{iid}{\sim} \pi_{\text{ref}}(\cdot \mid s_0, s_{1\,t-1}^{[h]:})$.
    计算奖励 $v_t^{(h,j)} := \hat{V}(s_0, (s_1^{[h]:t-1}, s_t^{(h,j)}))$ 对于 $j = 1, \ldots, M$.
    构建经验归一化因子 $\bar{Z}_t^{[h]} := \frac{1}{M}\sum_{j=1}^{M} v_t^{(h,j)}$。并设置 $s_t^{[h]} \leftarrow s^{(h,A_t)\,t}$。
    重采样一个动作索引 $A_t \sim$ 分类分布 $\left(\left\{\frac{v_t^{(h,j)}}{\sum_{\ell=1}^{M} v_t^{(h,\ell)}}\right\}_{j=1}^{M}\right)$
    更新提议分布权重 $w^{[h]} \leftarrow w^{[h]} \cdot \frac{\hat{V}(s_0, s_{1\,t}^{[h]:})}{\bar{Z}_t^{[h]}}$.
  结束循环
  如果 $h = 1$ **则**
    设定 $s_{1:T}^{\text{acc}} \leftarrow s_{1:T}^{[h]}$ 和 $w^{\text{acc}} \leftarrow w^{[h]}$。
    继续执行
  结束 if
  从均匀分布中抽取 $\alpha \sim [0,1]$。
  如果 $\alpha < 1 \wedge \frac{w^{\text{acc}} \hat{V}(s_{1\,T}^{[h]:})}{w^{[h]} \hat{V}(s_{1:T}^{\text{acc}})}$ **then**
    令 $s_{1:T}^{\text{acc}} \leftarrow s_{1:T}^{[h]}$ 且 $w^{\text{acc}} \leftarrow w^{[h]}$。
  结束如果
结束循环
输出： $s_{1:T}^{\text{acc}}$.

算法 8 带 Metropolis-Hastings 校正的 SP-gSMC

1: 输入：总时域 T、粒子数 N、Metropolis Hasting 步数 H、奖励模型 $\hat{V}$、预训练基础模型 π_{ref}、初始提示 s_0、失败事件概率 δ。 2: 初始化当前接受轨迹容器 $s_{1:T}^{\text{acc}} \leftarrow \varnothing$ 3: 对 h $= 1, \cdots, H$ 执行 4: 初始化累积拒绝权重 $w \leftarrow 1$, $w' \leftarrow 1$ 5: 对 $t = 1, \cdots, T$ 执行 6: $s_t^{[h]} \leftarrow$ 调用拒绝采样（算法 3）$(\hat{V}(s_0, (s_{1:t-1}, \cdot)), \pi_{\text{ref}}(\cdot|s_0, s_{1:t-1}), O(L) > C_{\text{act}}, O(\delta/T\log(\delta^{-1})))$ 我们指定 $s_{1:0} := \varnothing$。 7: $w \leftarrow w \cdot \hat{V}(s_{1\,t}^{[h]:}) / \hat{\mathbb{E}}_{s_t' \sim \pi_{\text{ref}}(\cdot|s_0, s_1^{[h]:t-1})}[\hat{V}(s_0, (s_1^{[h]:t-1}, s_t'))]$ $\hat{\mathbb{E}}$ 是算法 4 的经验期望。 8: $w' \leftarrow w' \cdot \hat{V}(s_{1:t}^{\text{acc}}) / \hat{\mathbb{E}}_{s_t' \sim \pi_{\text{ref}}(\cdot|x_0, s_{1:t-1}^{\text{acc}})}[\hat{V}(s_0, (s_{1:t-1}^{\text{acc}}, s_t'))]$ 如果 $s_{1:T}^{\text{acc}} = \varnothing$, 我们定义 $\hat{Q}(\varnothing) \equiv 1$ 9: 结束循环 10: 如果 h $= 1$ 则 11: 继续循环 12: 结束条件 13: 如果 $\alpha \sim$ 均匀分布 $[0,1], \alpha < \frac{w' \hat{V}(s_{1:T}^{\text{acc}}) \cdot \hat{V}(s_{1\,T}^{[h]:})}{w}$ 则 14: 设定 $s_{1:T}^{\text{acc}} \leftarrow s_1^{[h]:T}$ 15: 结束条件 16: 结束循环 17: 输出： $s_{1:T}^{\text{acc}}$

C 第 4 节中的证明

定理 4.1的证明。 记作 $T = 3m$。 我们假设 L 是整数；该结论很容易推广到任何大于等于 1 的实数。固定任意随机化算法 A，并将其视为定义在确定性算法 a上的 $R \in \Delta(\mathsf{A})$。以访问的深度为-$2m$ 前缀数量来衡量时间。假设 R 的支持集中的每个 a 在每次运行中访问的深度为-$2m$ 前缀不超过 Q 个。

困难输入族。 设 $B := L$，并令 π_{unif}为叶节点 $[B]^T$上的均匀参考分布。对每个 $u \in [B]^{2m}$，定义 $\hat{} V_u$，使得 $\hat{} V_u(\varnothing) = 1$，并

$$\frac{\hat{V}_u(s_{1:t})}{\hat{V}_u(s_{1:t-1})} = \begin{cases} 1, & 1 \le t \le 2m, \\ L, & 2m < t \le 3m \text{ and } s_{1:2m} = u, \\ 1/L, & 2m < t \le 3m \text{ and } s_{1:2m} \ne u. \end{cases} \tag{6}$$

令 $I_u := B$， π_{unif}， $\hat{} V_u$，并将诱导出的叶节点上的目标分布记为 p_{I_u}。

对每个 $u \in [B]^{2m}$，定义

$$A_u := \{s_{1:T} \in [B]^T : s_{1:2m} = u\}.$$

在（6）的条件下， A_u 中的每个叶节点的未归一化权重均与 L^m成正比，而 A_u^c 中的每个叶节点的未归一化权重均与 L^{-m}成正比；因此 A_u 恰好是 在 p_{I_u}下概率最大的叶节点

$$p_{I_u}(A_u) = \frac{1}{2 - L^{-2m}} > \frac{1}{2}. \tag{7}$$

$\hat{}$ $\hat{}$ $\hat{}$ $\hat{}$令 $U := S_{1:2m}$。 那么 $\{U = u\}$ 恰好是事件 $\{S_{1:T} \in A_u\}$。

步骤 1（全变差控制迫使采样到最大叶节点的概率为常数）。根据假设，$\|q_{A,\ I_u} - p_{I_u}\|_{\text{TV}} \le 1/3$ 对每个$u \in [B]^{2m}$。对事件 A_u 应用全变差的定义，并利用（7），可得

$$q_{A,I_u}(A_u) \ \ge \ p_{I_u}(A_u) - \frac{1}{3} \ > \ \frac{1}{6}. \tag{8}$$

等价地，对每个 $u \in [B]^{2m}$,

$$\mathbb{E}_{\mathsf{a}\sim R}\, \mathbb{P}_{o\sim\mathbb{P}_{\pi_{\text{unif}}}}\big(\widehat{U} = u \mid \mathsf{a},\ I_u\big) \ > \ \frac{1}{6}. \tag{9}$$

步骤 2（通过显式并集界得到 Yao / 不等式形式）。定义输入分布 D 如下： $B = L$ 的概率为 1， $\pi_{\text{ref}} = \pi_{\text{unif}}$ = 的概率为 1，而$\hat{V} = \hat{}$ V_U，其中 $U \sim \text{Unif}([B]^{2m})$。那么 D $\in \Delta(\mathrm{I})$，其中 $\mathrm{I} : \{I_u : u \in [B]^{2m}\}$。

固定任意一个确定性a $\in$A，它访问的深度-$2m$前缀不超过 Q 个。令 $\mathrm{U}_o(\mathsf{a})[1], \ldots, \mathrm{U}_o(\mathsf{a})[Q] \in [B]^{2m}$ 表示使用预言机 o 运行 a 时访问的深度-$2m$ 前缀列表（依赖预言机的）。根据禁止猜测约束，事件 $\{U = U\}$ 意味着访问列表必须包含 U:

$$\{\widehat{U} = U\} \ \subseteq \ \bigcup_{i=1}^{Q} \{U = \mathcal{U}_o(\mathsf{a})[i]\}.$$

因此，利用 U 与 o 在 D 下的独立性,

$$\begin{aligned} \mathbb{E}_{U\sim\text{Unif}}\Big[\mathbb{P}_{o\sim\mathbb{P}_{\pi_{\text{unif}}}}\big(\widehat{U} = U \mid \mathsf{a}, I_U\big)\Big] &\le \mathbb{E}_U\Big[\mathbb{P}_o\Big(\bigcup_{i=1}^{Q}\{U = \mathcal{U}_o(\mathsf{a})[i]\}\Big)\Big] \\ &\le \mathbb{E}_U\Big[\sum_{i=1}^{Q} \mathbb{P}_o\big(U = \mathcal{U}_o(\mathsf{a})[i]\big)\Big] \\ &= \sum_{i=1}^{Q} \mathbb{E}_o\Big[\mathbb{P}_U\big(U = \mathcal{U}_o(\mathsf{a})[i] \mid o\big)\Big] \\ &= \sum_{i=1}^{Q} \mathbb{E}_o\Big[\frac{1}{B^{2m}}\Big] \ = \ \frac{Q}{B^{2m}}. \end{aligned} \tag{10}$$

现在对a $\sim R$ 取平均，并利用线性性:

$$\mathbb{E}_{I\sim\mathcal{D}}\, \mathbb{E}_{\mathsf{a}\sim R}\, \mathbb{P}_{o\sim\mathbb{P}_{\pi_{\text{unif}}}}\big(\widehat{U} = U \mid \mathsf{a}, I\big) \ \le \ \frac{Q}{B^{2m}}. \tag{11}$$

最后，s 由于对于任意可测的 f，$\inf_{I\in\mathcal{I}} f(I) \le \mathbb{E}_{I\sim\mathcal{D}}[f(I)]$，我们得到不等式形式的姚氏步骤

$$\inf_{I\in\mathcal{I}} \mathbb{E}_{\mathsf{a}\sim R}\,\mathbb{P}_{o\sim\mathbb{P}_{\pi_{\text{unif}}}}\big(\widehat{U}=u(I)\mid \mathsf{a}, I\big) \;\le\; \frac{Q}{B^{2m}}, \tag{12}$$

其中 $u(I)$表示满足 $I=I_u$ 的唯一索引 u。

步骤3（合并）。根据（9）及定义$\mathcal{I}:=\{I_u:\ u\in[B]^{2m}\}$，我们有

$$\inf_{I\in\mathcal{I}} \mathbb{E}_{\mathsf{a}\sim R}\,\mathbb{P}_{o\sim\mathbb{P}_{\pi_{\text{unif}}}}\big(\widehat{U}=u(I)\mid \mathsf{a}, I\big) \;>\; \frac{1}{6}. \tag{13}$$

结合（12）和（13）可得 $Q/B^{2m}\ge 1/6$，因此 $Q\ge B^{2m}/6$。回顾 $B=L$ 和 $2m=2T/3$，我们得出最坏情况复杂度为 $\Omega\left(L^{2T/3}\right)$。□

推论 *4.2* 的证明。通过实例化一个满足假设 4.1、3.1 和 3.2 的困难实例族，我们将问题归约到定理。

令 $B:=1+\varepsilon$，并令 π_{unif} 为 $[B]^T$上的均匀参考分布。对于每个 $u\in[B]^{2m}$，按照与（6）相同的构造定义 $\hat{}V_u$，但将 L 在各处替换为 $1+\varepsilon$:

$$\frac{\hat{V}_u(s_{1:t})}{\hat{V}_u(s_{1:t-1})} = \begin{cases} 1, & 1\le t\le 2m,\\ 1+\varepsilon, & 2m<t\le 3m \text{ and } s_{1:2m}=u,\\ (1+\varepsilon)^{-1}, & 2m<t\le 3m \text{ and } s_{1:2m}\neq u. \end{cases} \tag{14}$$

（与之前一样，可以验证每个 $I_u=\langle B,\pi_{\text{unif}},\hat{}V_u\rangle$ 都满足假设 3.1 和 3.2。）

现在原样应用定理 4.1 的证明（将 L 替换为 $1+\varepsilon$，并将 $\mathbb{P}_{\pi_{\text{unif}}}$ 替换为由 $\pi_{\text{ref}}=\pi_{\text{unif}}$诱导的预言机测度）：全变差条件 $\|q_{A,I_u}-p_{I_u}\|_{\text{TV}}\le 1/3$ 迫使 $q_{A,I_u}(A_u)>1/6$，而相同的并集界／姚氏论证给出了最坏情况下访问的深度为 $2m$ 的前缀数量下界$Q\ge \frac{1}{6}B^{2m}$。由于 $2m=2T/3$，这便得到所述的 $\Omega((1+\varepsilon)^{2T/3})$ 下界。□

L引理 C.1。设 $(\Omega、\mathcal{F}、\mu)$ 为一个测度空间。设 $\pi、\phi、\psi:\ \Omega\to[0,\infty)$ 为可测 *wit* h

$$0<Z_\phi:=\int \pi(x)\phi(x)\,d\mu(x)<\infty, \qquad 0<Z_\psi:=\int \pi(x)\psi(x)\,d\mu(x)<\infty.$$

定义概率测度 $P、Q$ 如下:

$$P(dx)=\frac{\pi(x)\phi(x)}{Z_\phi}\,d\mu(x), \qquad Q(dx)=\frac{\pi(x)\psi(x)}{Z_\psi}\,d\mu(x).$$

假设对于某个 $\varepsilon\in(0,1)$,

$$\operatorname*{ess\,sup}_{x\in\Omega}\max\left\{\frac{\phi(x)}{\psi(x)},\frac{\psi(x)}{\phi(x)}\right\}\le 1+\varepsilon.$$

Then

$$\|P-Q\|_{\text{TV}}\le\frac{\varepsilon}{1-\varepsilon}.$$

引理 *C.1*的证明。令

$$Z_\phi=\int \pi\phi\,d\mu, \qquad Z_\psi=\int \pi\psi\,d\mu.$$

一方面,

$$\frac{dQ}{d\mu}(x)=\frac{\pi(x)\psi(x)}{Z_\psi},$$

因此

$$\frac{dP}{dQ}(x)=\frac{\frac{dP}{d\mu}(x)}{\frac{dQ}{d\mu}(x)}=\frac{\pi(x)\phi(x)/Z_\phi}{\pi(x)\psi(x)/Z_\psi}=\frac{\phi(x)}{\psi(x)}\cdot\frac{Z_\psi}{Z_\phi}.$$

另一方面,

$$Z_\phi=\int \pi(x)\phi(x)\,d\mu(x)=\int\frac{\phi(x)}{\psi(x)}\,\pi(x)\psi(x)\,d\mu(x)=Z_\psi\,\mathbb{E}_Q\left[\frac{\phi}{\psi}\right],$$

so

$$\frac{Z_\psi}{Z_\phi}=\frac{1}{\mathbb{E}_Q(\phi/\psi)}.$$

Thus

$$\frac{dP}{dQ}(x) = \frac{\phi(x)/\psi(x)}{\mathbb{E}_Q(\phi/\psi)}.$$

定义

$$a(x) := \frac{\phi(x)}{\psi(x)}.$$

根据假设

$$\max\left\{\frac{\phi(x)}{\psi(x)}, \frac{\psi(x)}{\phi(x)}\right\} \le 1+\varepsilon$$

我们得到

$$\frac{1}{1+\varepsilon} \le a(x) \le 1+\varepsilon \quad \text{a.e.},$$

因此

$$|a(x)-1| \le \varepsilon \quad \text{a.e.}$$

因此

$$|\,\mathbb{E}_Q(a)-1\,| \le \mathbb{E}_Q|a-1| \le \varepsilon, \qquad \mathbb{E}_Q(a) \ge 1-\varepsilon > 0.$$

Now

$$\frac{dP}{dQ}(x) = \frac{a(x)}{\mathbb{E}_Q(a)},$$

so

$$\|P-Q\|_{\mathrm{TV}} = \frac{1}{2}\int \left|\frac{dP}{dQ}-1\right| dQ = \frac{1}{2}\,\mathbb{E}_Q\left|\frac{a}{\mathbb{E}_Q(a)}-1\right| = \frac{1}{2}\,\mathbb{E}_Q\left|\frac{a-\mathbb{E}_Q(a)}{\mathbb{E}_Q(a)}\right|.$$

利用 $\mathbb{E}_Q(a) \ge 1-\varepsilon$ 和三角不等式，

$$\mathbb{E}_Q\left|\frac{a-\mathbb{E}_Q(a)}{\mathbb{E}_Q(a)}\right| \le \frac{1}{1-\varepsilon}\,\mathbb{E}_Q|a-\mathbb{E}_Q(a)| \le \frac{1}{1-\varepsilon}\,\mathbb{E}_Q\big(|a-1|+|\mathbb{E}_Q(a)-1|\big) \le \frac{1}{1-\varepsilon}\,(\varepsilon+\varepsilon) = \frac{2\varepsilon}{1-\varepsilon}.$$

因此

$$\|P-Q\|_{\mathrm{TV}} \le \frac{1}{2}\cdot\frac{2\varepsilon}{1-\varepsilon} = \frac{\varepsilon}{1-\varepsilon},$$

这就是所需的界。 □

定理的证明 *4.3*。 对每个 $t \ge 0$，定义前缀上的中间分布

$$q_t(s_{1:t} \mid s_0) := \frac{\pi_{\mathrm{ref}}(s_{1:t} \mid s_0)\,\mathbb{E}_{s_{t+1}\sim\pi_{\mathrm{ref}}(\cdot|s_0,s_{1:t})}\big[\hat{V}(s_0,s_{1:t+1})\big]}{\mathbb{E}_{s_{1:t+1}\sim\pi_{\mathrm{ref}}(\cdot|s_0)}\big[\hat{V}(s_0,s_{1:t+1})\big]}. \tag{15}$$

根据构造，

$$\tilde{\pi}_{t+1}(s_{1:t+1} \mid s_0) = q_t(s_{1:t} \mid s_0)\,\hat{\pi}(s_{t+1} \mid s_0, s_{1:t}). \tag{16}$$

此外，根据序贯采样器的定义，

$$\hat{\pi}_{t+1}(s_{1:t+1} \mid s_0) = \hat{\pi}_t(s_{1:t} \mid s_0)\,\hat{\pi}(s_{t+1} \mid s_0, s_{1:t}). \tag{17}$$

利用（16）–（17），以及三角不等式，

$$\begin{aligned}\|\tilde{\pi}_{t+1}-\hat{\pi}_{t+1}\|_{\mathrm{TV}} &= \big\|q_t\,\hat{\pi}(\cdot \mid s_{1:t}) - \hat{\pi}_t\,\hat{\pi}(\cdot \mid s_{1:t})\big\|_{\mathrm{TV}} \\ &\le \big\|\tilde{\pi}_t\,\hat{\pi}(\cdot \mid s_{1:t}) - \hat{\pi}_t\,\hat{\pi}(\cdot \mid s_{1:t})\big\|_{\mathrm{TV}} + \big\|q_t\,\hat{\pi}(\cdot \mid s_{1:t}) - \tilde{\pi}_t\,\hat{\pi}(\cdot \mid s_{1:t})\big\|_{\mathrm{TV}}.\end{aligned}$$

对于这两项， 因为对它们应用相同的马尔可夫核不会增加全变差（数据处理不等式 ），

$$\big\|\tilde{\pi}_t\,\hat{\pi}(\cdot \mid s_{1:t}) - \hat{\pi}_t\,\hat{\pi}(\cdot \mid s_{1:t})\big\|_{\mathrm{TV}} \le \|\tilde{\pi}_t-\hat{\pi}_t\|_{\mathrm{TV}}, \qquad \big\|q_t\,\hat{\pi}(\cdot \mid s_{1:t}) - \tilde{\pi}_t\,\hat{\pi}(\cdot \mid s_{1:t})\big\|_{\mathrm{TV}} \le \|q_t-\tilde{\pi}_t\|_{\mathrm{TV}}.$$

根据引理 C.1 和假设 3.2，我们有 $\|q_t-\tilde{\pi}_t\|_{\mathrm{TV}} \le 2\varepsilon$。因此，

$$\|\tilde{\pi}_{t+1}-\hat{\pi}_{t+1}\|_{\mathrm{TV}} \le \|\tilde{\pi}_t-\hat{\pi}_t\|_{\mathrm{TV}} + 2\varepsilon.$$

以基本情况 $\|\tilde{\pi}_0-\hat{\pi}_0\|_{\mathrm{TV}} = 0$ 为起点，对 t 作望远镜求和，可得 $\|\tilde{\pi}_t-\hat{\pi}_t\|_{\mathrm{TV}} \le 2t\varepsilon$。 □

D 最优提议分布序贯蒙特卡洛

提议分布 π_{p} 的选择：朴素与最优。如 [10], 所述，抽取样本所依据的提议分布 π_{p} 的不同选择，可能导致归一化因子估计的方差不同。我们认为，采样情形与此完全相同。最简单、最直接的选择是朴素提议分布，即令 $\pi_{\mathrm{p}} = \pi_{\mathrm{ref}}$，也就是从预训练模型自身的输出分布中采样。在此情形下，$G_t(s_{1:t}) = \hat{V}(s_0, s_{1:t})\,\hat{V}(s_0, s_{1:t-1})$。另一种选择，即最小化经验归一化因子方差的最优提议分布 [10] 为 $\pi_{\mathrm{p}}(s_{t+1}|s_0, s_{1:t}) \propto \pi_{\mathrm{ref}}(s_{t+1}|s_0, s_{1:t})\,\hat{V}(s_0, s_{1:t+1})$，此时 $G_t(s_{1:t}) = \mathbb{E}_{s_t \sim \pi_{\mathrm{ref}}(\cdot|s_0, s_{1:t-1})}[\hat{V}(s_0, s_{1:t})]\,\hat{V}(s_0, s_{1:t-1})$。然而，很难得到精确的最优提议分布，因此我们使用拒绝采样进行近似，如算法 3 所示。具体而言，在附录 B 中，我们在算法 6 中针对最优提议分布实例化了算法1。

粒子复杂度。具体而言，对于最优提议分布，势函数（增量权重）$G_t(s_{1:t}) = \mathbb{E}_{s_t \sim \pi_{\mathrm{ref}}(\cdot|s_0, s_{1:t-1})}[\hat{V}(s_0, s_{1:t})]\,\hat{V}(s_0, s_{1:t-1})$包含对下一步奖励的条件期望。在我们的设定中，我们考虑使用一种实用的蒙特卡罗方法来估计该值，这会使计算复杂度额外增加一个 N_m 因子。也就是说，每当需要访问 $\mathbb{E}_{s_t \sim \pi_{\mathrm{ref}}(\cdot|s_0, s_{1:t-1})}[\hat{V}(s_0, s_{1:t})]$，时，我们就使用 $\frac{1}{N_m}\sum_{i=1}^{N_m} \hat{V}(s_0, (s_{1:t-1}, S_t))$ 作为估计，其中 $\{S_i\}_{i=1}^{N_m}$ 是从任何其他随机性来源中独立抽取的独立同分布样本。采用最优提议分布的序贯蒙特卡洛的详细算法见算法 6。相应地，刻画最优提议分布序贯蒙特卡洛粒子复杂度的定理如下。

定理 D.1 (粒子复杂度)。 固定时域 $T \geq 2$ 和目标精度 $\delta_{\mathrm{TV}} \in (0, 1)$。若假设 *3.1* 和*3.2* 在 L、$\varepsilon > 0$ 下成立，令 $\bar{\eta}_{T+1}^N := \mathbb{E}[\eta_{T+1}^N]$ 表示序贯蒙特卡洛采样器在终端重采样步骤 $T+1$ （使用$M_{T+1} = \mathrm{Id}$）之后的输出分布。那么，在最优提议分布下，以下充分条件可保证 $\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \leq \delta_{\mathrm{TV}}$。也就是说，粒子数 $N \geq \frac{((1+\varepsilon)^4-1)}{2\,\delta_{\mathrm{TV}}}\, T^2\, 1+\varepsilon^{\,6T}()$。

被省略的证明可见于附录E。

注记 D.2。 定理 D.1 展示了最优提议分布的一项显著优势：其粒子复杂度与L无关，并且可能显著小于朴素提议分布的粒子复杂度。此外，粒子复杂度会随着贝尔曼误差ε 的减小而降低；具体而言，当贝尔曼误差 $\varepsilon = 0$时，仅需一个粒子即可实现精确采样——也就是说，采用最优提议分布的序贯蒙特卡洛采样器退化为动作级采样器（SP-gSMC）。然而，以下命题 D.4表明，这种粒子复杂度方面的效率是有代价的——即增量权重的估计较为困难（尤其是在句子级设定下）。因此，正如我们稍后将在注记 E.8中提到的，在设计提议分布时，需要权衡不同来源的开销。

注 D.3。将定理 5.1 与定理 4.3 进行比较可以发现，具体就最优提议分布而言，在$\varepsilon = O(\delta_{\mathrm{TV}}/T)$机制下，序贯蒙特卡洛采样器的粒子复杂度为 $O\left(\frac{\varepsilon T^2}{\delta_{\mathrm{TV}}}\right) = O(T)$，而最简单的单粒子引导序贯蒙特卡洛采样器仅使用 1 个粒子。检视该证明可以发现，额外的 T 复杂度来自使用 N 个粒子所产生的局部方差，并被费曼-卡茨半群的稳定性系数放大，该系数与 T成比例。另一方面，尽管单粒子引导序贯蒙特卡洛采样器本质上是针对目标费曼-卡茨流的序贯蒙特卡洛采样器，且方差较大（相较于多粒子情形），但由于它只有一个粒子，并不存在实际的重新加权步骤，因此它可以灵活地对应于任何其他共享同一马尔可夫转移核 M_t的费曼-卡茨流，差异仅体现在 G_t 以及随后的局部方差上。在 G_t的所有选择中，常数势函数（例如 $G_t = 1$ ）使单粒子引导序贯蒙特卡洛采样器作为序贯蒙特卡洛采样器时的方差为 0，代价是产生 $O(\varepsilon T)$偏差。因此，上述观察表明，在使用 $o(T)$ 个粒子时，最优的序贯蒙特卡洛采样器未必优于仅使用单个粒子的采样器。

时间复杂度。从定理5.1 到推论5.2 的推导很直接，即通过乘以一个 T 代表时域的因子即可。然而，由于估计条件期望下一步奖励会引入误差，使用最优提议分布的序贯蒙特卡罗的复杂度并不能直接 T 等于粒子复杂度的 T 倍，因此需要在下述命题D.4中进行更仔细的讨论。

命题 D.4（使用最优提议分布的序贯蒙特卡罗采样器的时间复杂度）。在假设*3.1* 和*3.2*下，并考虑 $\varepsilon = O(1/T)$和 $\delta_{\mathrm{TV}} = O(1)$成立的情形。对于任意 $0 < \delta \lesssim \delta_{\mathrm{TV}}$，使用时间 $\tilde{O}\left(\frac{\varepsilon L T^5}{\delta_{\mathrm{TV}}^3} \log(\delta^{-1})\right)$，存在一个良好事件$\mathcal{E}$，使得 $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$；在以该事件为条件时，使用最优提议分布的序贯蒙特卡罗采样器可达到 δ_{TV} 精度，即$\|\bar{\pi}_T - \hat{\pi}_T|_{\mathcal{E}}\|_{\mathrm{TV}} \leq \delta_{\mathrm{TV}}$。

我们将证明推迟到附录E。

备注 D.5（关于采用最优提议分布的序贯蒙特卡洛采样器的低效率）。 如备注 D.2所述，采用最优提议分布的序贯蒙特卡洛采样器的缺点是，计算增量权重的成本过高——需要通过蒙特卡洛估计来计算经验期望，以达到 $O(\delta_{\mathrm{TV}}/T)$ 精度。不过，在词元级设定中，情况可能有所不同，因为期望值可以在 $O(|\mathcal{S}|)$ 的时间内计算出来。词元级设定下时间复杂度的计算是命题 D.1的一个相当简单直接的推论，但不在本文讨论范围之内。

E 节和附录5中的证明D

粒子记号。 在证明中，我们采用 [32], 所使用的记号，其中 $\xi_t^{(N,i)}$ 表示时刻 i的粒子池中的第 N 个粒子 t。粒子的演化与第 2.2 节中完全相同。

Dobrushin（遍历）系数与张量积。 设 $M:(E,\mathcal{E}) \to (F,\mathcal{F})$ 是一个有界积分算子，且具有常质量，即 $M(1)(x) = M(1)(y)$ 对所有 $x, y \in E$. 根据 [32], 定义 *Dobrushin* 系数

$$\beta(M) := \sup_{f \in \mathrm{Osc}_1(F)} \mathrm{osc}(Mf), \qquad \mathrm{osc}(f) := \sup f - \inf f.$$

特别地，对于马尔可夫核 P （因此 $P(1) = 1$）， $\beta(P)$ 也称为*Dobrushin* 遍历系数。对于定义在 (E,E) 上的有限符号测度 μ, ν，定义在 $(E \times E, \mathrm{E} \otimes \mathrm{E})$ 上的张量积测度 $\mu \otimes \nu$ 为

$$(\mu \otimes \nu)(h) := \iint h(x,y)\, \mu(dx)\, \nu(dy),$$

对于 $E \times E$ 上任意有界可测的h 。对于定义在 E 上的可测函数 g, h， 记 $(g \otimes h)(x,y) := g(x)h(y)$。

引理 E.1 （FK 一阶展开、余项与稳定性（ [32] 的附录 A.3））。 考虑一个具有如下马尔可夫核的费曼–卡茨模型 $(M_k)_{k\geq 1}$ 以及势函数 $(G_k)_{k\geq 0}$。定义

$$Q_k(f)(x) := G_{k-1}(x)\, M_k(f)(x), \qquad Q_{p,n} := Q_{p+1} \cdots Q_n, \qquad Q_{n,n} := \mathrm{Id},$$

以及归一化 *FK* 流

$$\Phi_{p,n}(\mu)(f) := \frac{\mu(Q_{p,n}(f))}{\mu(Q_{p,n}(1))}.$$

Let

$$P_{p,n}(f)(x) := \frac{Q_{p,n}(f)(x)}{Q_{p,n}(1)(x)}, \qquad G_{p,n,\eta}(x) := \frac{Q_{p,n}(1)(x)}{\eta(Q_{p,n}(1))}, \qquad q_{p,n} := \sup_{x,y \in E_p} \frac{Q_{p,n}(1)(x)}{Q_{p,n}(1)(y)}.$$

（***i***）一阶（线性）算子。 对于任意 $\eta \in \mathcal{P}(E_p)$， 定义线性算子 $D_\eta \Phi_{p,\,n}$ 为

$$D_\eta \Phi_{p,n}(f)(x) := G_{p,n,\eta}(x)\, P_{p,n}\big(f - \Phi_{p,n}(\eta)(f)\big)(x). \tag{18}$$

它是映射 $\mu \mapsto \Phi_{p,\,n}(\mu)$ 在 η 处的一阶（*Gâteaux/Fréchet*）线性化：对于任意有限符号测度 ν， 若满足 $\nu(1) = 0$,

$$\nu\big(D_\eta \Phi_{p,n}(f)\big) = \frac{d}{d\epsilon} \Phi_{p,n}(\eta + \epsilon\nu)(f)\Big|_{\epsilon=0}.$$

（***ii***）一阶分解 +显式二阶余项。 对于任意 μ、 $\eta \in \mathcal{P}(E_p)$， 以及定义在 E_n 上的有界可测 f,

$$\Phi_{p,n}(\mu)(f) - \Phi_{p,n}(\eta)(f) = (\mu - \eta)\big(D_\eta \Phi_{p,n}(f)\big) + R^{\Phi_{p,n}}(\mu,\eta)(f), \tag{19}$$

其中余项具有显式张量形式

$$R^{\Phi_{p,n}}(\mu,\eta)(f) := -\frac{1}{\mu(G_{p,n,\eta})}\, (\mu - \eta)^{\otimes 2}\Big(G_{p,n,\eta} \otimes D_\eta \Phi_{p,n}(f)\Big). \tag{20}$$

(iii) Dobrushin 稳定性。对任意 $f \in$ 振幅$_1(E_n)$,

$$\beta(D\Phi_{p,n}) := \sup_{\eta \in \mathcal{P}(E_p)} \beta\big(D_\eta \Phi_{p,n}\big) \leq 2\, q_{p,n}\, \beta(P_{p,n}) \leq 2\, q_{p,n}. \tag{21}$$

证明。 固定 $p \leq n, \mu,\ \eta \in \mathrm{P}(E_p)$ 以及 E_n 上的有界可测 f。为简便起见，记

$$g := Q_{p,n}(1), \qquad h := Q_{p,n}(f), \qquad d := \eta(g), \qquad G := G_{p,n,\eta} = \frac{g}{d}, \qquad P := P_{p,n} = \frac{Q_{p,n}(\cdot)}{Q_{p,n}(1)} = \frac{\cdot}{g}.$$

注意， $\eta(G) = 1$和 $\mu(G) = \mu(g)/d > 0$。另设 $\phi := \Phi_{p,\,n}(\eta)(\,f) = \eta($小时$)/d = \eta(G\,P(\,f))$。

步骤 1：精确商恒等式。根据定义，

$$\Phi_{p,n}(\mu)(f) - \Phi_{p,n}(\eta)(f) = \frac{\mu(h)}{\mu(g)} - \frac{\eta(h)}{\eta(g)} = \frac{\mu(h)d - \eta(h)\mu(g)}{\mu(g)d}.$$

利用$h = g\,P(f)$与 $\eta(h) = d\,\phi$，我们得到

$$\mu(h)d - \eta(h)\mu(g) = d\,\mu(gP(f)) - \mu(g)\,d\,\phi = d\,\mu\big(g(P(f) - \phi)\big) = d^2\,\mu\big(G\,P(f - \phi)\big).$$

类似地，

$$\eta(h)d - \eta(h)\eta(g) = d\,\eta(gP(f)) - d^2\phi = 0,$$

所以 $d^2\,\mu(GP(f - \phi)) = d^2(\mu - \eta)(GP(f - \phi))$。因此

$$\Phi_{p,n}(\mu)(f) - \Phi_{p,n}(\eta)(f) = \frac{d^2}{\mu(g)d}\,(\mu - \eta)\big(G\,P(f - \phi)\big) = \frac{1}{\mu(G)}\,(\mu - \eta)\big(D_\eta\Phi_{p,n}(f)\big),$$

其中我们使用了 定义 $D_\eta\Phi_{p,n}(f) = G\,P(f - \Phi_{p,n}(\eta)(f)) = G\,P(f - \phi)$。这证明了（精确）恒等 y

$$\Phi_{p,n}(\mu)(f) - \Phi_{p,n}(\eta)(f) = \frac{1}{\mu(G)}\,(\mu - \eta)\big(D_\eta\Phi_{p,n}(f)\big). \tag{22}$$

步骤 2：一阶分解 +显式余项。由于 $\eta(G) = 1$，我们有

$$\frac{1}{\mu(G)} - 1 = \frac{1 - \mu(G)}{\mu(G)} = -\frac{(\mu - \eta)(G)}{\mu(G)}.$$

将其代入（22）可得

$$\Phi_{p,n}(\mu)(f) - \Phi_{p,n}(\eta)(f) = (\mu - \eta)\big(D_\eta\Phi_{p,n}(f)\big) - \frac{(\mu - \eta)(G)}{\mu(G)}\,(\mu - \eta)\big(D_\eta\Phi_{p,n}(f)\big).$$

根据张量积约定（$\mu - \eta$）$^{\otimes 2}$（$G \otimes u$） =（$\mu - \eta$）（G）（$\mu - \eta$）（u），最后一项恰为

$$-\frac{1}{\mu(G)}(\mu - \eta)^{\otimes 2}\Big(G \otimes D_\eta\Phi_{p,n}(f)\Big),$$

这就是（20）。这证明了（19）–（20）。

步骤 3：Dobrushin 稳定性。令 $f \in$振幅$_1(E_n)$，并保持$\phi = \Phi_{p,n}(\eta)(f)$。由于 P是一个马尔可夫算子， $P(f)$的取值位于 f的凸包中，因此 $\operatorname{osc}(P(f)) \le \beta(P)\operatorname{osc}(f) \le \beta(P)$，并且对于任意常数 c， $\|P(f) - c\|_\infty \le$振幅$(P(f))$。利用$\phi = \eta(GP(f))$，我们可以对任意 $x \in E_p$写出：

$$D_\eta\Phi_{p,n}(f)(x) = G(x)\Big(P(f)(x) - \eta(GP(f))\Big).$$

因此，

$$\big\|D_\eta\Phi_{p,n}(f)\big\|_\infty \le \|G\|_\infty\,\big\|P(f) - \eta(GP(f))\big\|_\infty \le \|G\|_\infty\operatorname{osc}(P(f)) \le q_{p,n}\,\beta(P),$$

因为 $\|G\|_\infty = \sup_x g(x)/\eta(g) \le \sup_x g(x)/\inf_y g(y) = q_{p,\,n}$。
接下来，对于任意 $x,\ x' \in E_p$，

$$\begin{aligned}&\big|D_\eta\Phi_{p,n}(f)(x) - D_\eta\Phi_{p,n}(f)(x')\big| \\ &\quad\le |G(x) - G(x')|\,\big|P(f)(x) - \eta(GP(f))\big| + \|G\|_\infty\,|P(f)(x) - P(f)(x')|.\end{aligned}$$

对 x， x'取上确界 并利用 $\|P(f) - \eta(GP(f))\|_\infty \le$振幅$(P(f)) \le \beta(P)$， 振幅$(P(f)) \le \beta(P)$， $\|G\|_\infty \le q_{p,\,n}$， 以及

$$\operatorname{osc}(G) = \sup_{x,x'}\Big|\frac{g(x)}{\eta(g)} - \frac{g(x')}{\eta(g)}\Big| = \frac{\sup g - \inf g}{\eta(g)} \le \frac{\sup g - \inf g}{\inf g} \le q_{p,n},$$

我们得到

$$\operatorname{osc}\big(D_\eta\Phi_{p,n}(f)\big) \le \operatorname{osc}(G)\,\beta(P) + \|G\|_\infty\,\beta(P) \le 2\,q_{p,n}\,\beta(P).$$

由于这对所有 $f \in$振幅$_1(E_n)$ 都成立，因此可得 $\beta(D_\eta\Phi_{p,\,n}) \le 2q_{p,\,n}\beta(P)$， 对 η 取上确界即可得到 (21)。最后，由于 P是马尔可夫核， $\beta(P) \le 1$，因此 $\beta(D\Phi_{p,\,n}) \le 2q_{p,\,n}$。 □

引理 E.2（续 FK 比率控制与一阶稳定性）。在假设 *3.1* 和假设 *3.2*下，考虑：

（*a*）朴素提议分布。 $M_t(s_{1:t-1},\ ds_{1:t}) = \pi_{\text{ref}}(s_t|s_0,\ s_{1:t-1}),\ G_t(s_{1:t}) = \hat{V}(s_0,\ s_{1:t})\hat{V}(s_0,\ s_{1:t-1})$。

（b）最优提议分布。 $M_t(s_{1:t-1}, ds_{1:t}) = \pi_{\text{ref}}(s_t|s_0, s_{1:t-1})\,\hat{V}(s_0, s_{1:t})\,\mathbb{E}_{s_t\sim\pi_{ref}(\cdot|s_0,s_{1:t-1})}[\hat{V}(s_0, s_{1:t})],\ G_t(s_{1:t}) = \mathbb{E}_{s_t\sim\pi_{ref}(\cdot|s_0,s_{1:t-1})}[\hat{V}(s_0, s_{1:t})]\hat{V}(s_0, s_{1:t-1})$。

那么，对于任意 $n \geq m$,

1. 情形 *(a)*：$q_{m,\ n} \leq L^2(1+\varepsilon)^{2(n-m-1)}$ 对于 $n > m$, 且 $q_{m,\ m} = 1$。*2.* 情形 *(b)*：$q_{m,\ n} \leq (1+\varepsilon)^{2(n-m)}$ 对于 $n > m$, 且 $q_{m,\ m} = 1$。*3.* 两种情形下：$\beta(D\Phi_{m,\ n}) \leq 2\,q_{m,\ n}$。

证明。对于$q_{m,\ n}$,我们证明朴素提议分布的情形，而最优提议分布的情形可由类似的论证得出 t.
当 $m = n$ 时，根据定义 $Q_{m,\ n} = \text{Id}$，因此 $q_{m,\ n} = \sup_{x,\ y} Q_{m,\ n}(1)(x)\,Q_{m,\ n}(1)(y) = 1$。

对 $Q_{m,\ n}(1)(x)$ 的直观解释。量 $Q_{m,\ n}(1)(x)$ 可以理解为从当前前缀/状态 $x \in E_m$开始的剩余前瞻 归一化因子：它是未来增量势函数乘积的条件期望（在从时间 m 到 $n-1$ 的提议动力学下），

$$Q_{m,n}(1)(x) \;=\; \mathbb{E}\left[\prod_{t=m}^{n-1} G_t(S_{1:t}) \;\middle|\; S_{1:m} = x\right].$$

其在不同 x 之间的变异性可以通过两种效应来理解。首先，未来贡献 $\mathbb{E}[\prod_{t=m+1}^{n-1} G_t \mid S_{1:m} = x]$ 会受到条件期望的“平滑”；在贝尔曼误差较小的假设下，相应的单步前瞻比率会一致地接近 1，因此这一未来项几乎处处接近常数，从而几乎与 x 无关。其次，当前增量势函数 $G_m(x)$ 是 $\sigma(S_{1:m})$- 可测的，因此不会被平滑 —— 其波动强烈取决于提议分布的选择（在朴素提议分布下通常更大，此时 $G_m = \hat{V}(s_0,\ s_{1:m})/\hat{V}(s_0,\ s_{1:m-1})$；而在最优提议分布下则更小，此时 G_m 是一个条件期望比率）。

当 $n > m$时，我们使用归纳法证明 $L^{-1}\cdot(1+\varepsilon)^{-(n-m-1)} \leq Q_{m,\ n}(1)(x) \leq L\cdot(1+\varepsilon)^{n-m-1}$。回顾对于所有 $x \in S_{1:m}$,

$$Q_{m,n}(1)(x) = \mathbb{E}\left[\prod_{p=m}^{n-1} G_p(X_p)\middle| X_m = x\right].$$

我们首先证明 $Q_{m,\ m+1}(1)(x) \in [L^{-1}, L]$, 这显然成立。接着假设 $Q_{m,\ n}(1)(x) \in [L^{-1}\cdot(1+\varepsilon)^{-(n-m-1)},\ L\cdot(1+\varepsilon)^{n-m-1}]$, 我们需要证明 $Q_{m,\ n+1}(1)(x) \in [L^{-1}\cdot(1+\varepsilon)^{-(n-m)},\ L\cdot(1+\varepsilon)^{n-m}]$。这是因为

$$\begin{aligned}
Q_{m,n+1}(1)(x) &= \mathbb{E}\left[\prod_{p=m}^{n} G_p(X_p)\middle| X_m = x\right] = \mathbb{E}\left[\mathbb{E}[G_n(X_n)|X_{n-1}]\prod_{p=m}^{n-1} G_p(X_p)\middle| X_m = x\right] \\
&\leq \sup \mathbb{E}[G_n(X_n)|X_{n-1}]\cdot\mathbb{E}\left[\prod_{p=m}^{n-1} G_p(X_p)\middle| X_m = x\right] \\
&= \sup\left\{\frac{\mathbb{E}_{s_t\sim\pi_{\text{ref}}(\cdot|s_0,s_{1:t-1})}[\hat{V}(s_0,s_{1:t})]}{\hat{V}(s_0,s_{1:t-1})}\right\}\cdot Q_{m,n}(1)(x) \\
&\leq (1+\varepsilon)\cdot L\cdot(1+\varepsilon)^{n-m-1} = L\cdot(1+\varepsilon)^{n-m}.
\end{aligned}$$

另一个方向也类似,

$$\begin{aligned}
Q_{m,n+1}(1)(x) &\geq \inf \mathbb{E}[G_n(X_n)|X_{n-1}]\cdot\mathbb{E}\left[\prod_{p=m}^{n-1} G_p(X_p)\middle| X_m = x\right] \\
&= \inf\left\{\frac{\mathbb{E}_{s_t\sim\pi_{\text{ref}}(\cdot|s_0,s_{1:t-1})}[\hat{V}(s_0,s_{1:t})]}{\hat{V}(s_0,s_{1:t-1})}\right\}\cdot Q_{m,n}(1)(x) \\
&\geq (1+\varepsilon)^{-1}\cdot L^{-1}\cdot(1+\varepsilon)^{-(n-m-1)} = L^{-1}\cdot(1+\varepsilon)^{-(n-m)}.
\end{aligned}$$

这样便完成了归纳法。于是我们得出 $q_{m,n} = \sup_{x,y} Q_{m,n}(1)(x)\, Q_{m,n}(1)(y) \le L^2 \cdot (1+\varepsilon)^{2(n-m-1)}$ 对于 $n > m$。

最后，根据引理 E.1和（21），

$$\beta(D\Phi_{m,n}) \le 2\, q_{m,n}\, \beta(P_{m,n}) \le 2\, q_{m,n},$$

由于 $P_{m,n}$ 是一个马尔可夫核，因此 $\beta(P_{m,n}) \le 1$。

□

引理 E.3（随机望远镜恒等式）。对于任意 $n \ge 1$ 以及定义在 E_n 上的任意有界可测测试函数 f，

$$[\eta_n^N - \eta_n](f) = \sum_{p=1}^{n} \Big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\Phi_p(\eta_{p-1}^N))\Big)(f). \tag{23}$$

因此，

$$[\bar{\eta}_n^N - \eta_n](f) = \sum_{p=1}^{n} \mathbb{E}\Big[\big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\Phi_p(\eta_{p-1}^N))\big)(f)\Big], \qquad \bar{\eta}_n^N := \mathbb{E}[\eta_n^N]. \tag{24}$$

证明。恒等式（23）纯粹是代数性的：对于每个 $p \ge 1$,

$$\Phi_{p-1,n} = \Phi_{p,n} \circ \Phi_p,$$

因此

$$\Phi_{p,n}(\Phi_p(\eta_{p-1}^N)) = \Phi_{p-1,n}(\eta_{p-1}^N).$$

将 $\Phi_{p,n}(\eta_p^N) - \Phi_{p-1,n}(\eta_{p-1}^N)$ 对 $p = 1, \ldots, n$ 求和，可经望远镜求和得到 $\eta_n^N(f) - \eta_n(f)$，因为 $\Phi_{n,n} = \mathrm{Id}$ 与 $\eta_0^N = \eta_0$。取期望即得（24）。□

引理 E.4（McKean 核下的单步条件协方差）。固定 $p \ge 1$，并令 $\mathcal{F}_{p-1}^N$ 为粒子系统直至时刻 $p-1$ 所生成的 σ–域。假设在给定 $\mathcal{F}_{p-1}^N$ 的条件下，粒子经由 *McKean* 核 K_{p,η_{p-1}^N} 独立更新，即

$$\xi_p^{(N,i)} \,\big|\, \mathcal{F}_{p-1}^N \;\sim\; K_{p,\eta_{p-1}^N}(\xi_{p-1}^{(N,i)}, \cdot), \qquad i = 1, \ldots, N,$$

并且关于 i条件独立。定义单步预测测度

$$\nu_p^N := \Phi_p(\eta_{p-1}^N) = \eta_{p-1}^N K_{p,\eta_{p-1}^N} \in \mathcal{P}(E_p),$$

其中第二个等式由 *McKean* 一致性（1）得出。于是，对任意有界可测 u、v（定义在 E_p 上），

$$\mathbb{E}\Big([\eta_p^N - \nu_p^N](u) \,\Big|\, \mathcal{F}_{p-1}^N\Big) = 0, \tag{25}$$

$$\mathbb{E}\Big([\eta_p^N - \nu_p^N](u)\,[\eta_p^N - \nu_p^N](v) \,\Big|\, \mathcal{F}_{p-1}^N\Big) = \frac{1}{N}\,\eta_{p-1}^N\Big(K_{p,\eta_{p-1}^N}\big((u - K_{p,\eta_{p-1}^N}u)\,(v - K_{p,\eta_{p-1}^N}v)\big)\Big). \tag{26}$$

特别地，

$$\mathbb{E}\Big([\eta_p^N - \nu_p^N](u)^2 \,\Big|\, \mathcal{F}_{p-1}^N\Big) = \frac{1}{N}\,\eta_{p-1}^N\Big(K_{p,\eta_{p-1}^N}\big((u - K_{p,\eta_{p-1}^N}u)^2\big)\Big). \tag{27}$$

此外，

$$\Big|\mathbb{E}\big([\eta_p^N - \nu_p^N](u)\,[\eta_p^N - \nu_p^N](v) \mid \mathcal{F}_{p-1}^N\big)\Big| \le \frac{1}{4N}\,\mathrm{osc}(u)\,\mathrm{osc}(v). \tag{28}$$

证明。在给定 F_{p-1}^N 的条件下，我们有

$$\eta_p^N(u) = \frac{1}{N}\sum_{i=1}^{N} u(\xi_p^{(N,i)}), \qquad \nu_p^N(u) = \eta_{p-1}^N K_{p,\eta_{p-1}^N}(u) = \frac{1}{N}\sum_{i=1}^{N} K_{p,\eta_{p-1}^N} u(\xi_{p-1}^{(N,i)}).$$

因此

$$[\eta_p^N - \nu_p^N](u) = \frac{1}{N}\sum_{i=1}^{N}\Big(u(\xi_p^{(N,i)}) - K_{p,\eta_{p-1}^N} u(\xi_{p-1}^{(N,i)})\Big).$$

取条件期望，并利用 $\mathbb{E}[u(\xi_{(N,\,i)p}) \mid \mathcal{F}_{p-1}^N] = K_{p,\,\eta_{p-1}^N} u(\xi_{p-1}^{(N,i)})$即可得到（25）。

对于第二个恒等式，展开乘积，并利用 i之间的条件独立性：由于每一项的条件均值均为 0，所有交叉项都消失，因此得到

$$\begin{aligned}
&\mathbb{E}\Big([\eta_p^N - \nu_p^N](u)\,[\eta_p^N - \nu_p^N](v)\,\Big|\,\mathcal{F}_{p-1}^N\Big) \\
&\quad= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\big(u(\xi_p^{(N,i)}) - K_{p,\eta_{p-1}^N}u(\xi_{p-1}^{(N,i)})\big)\big(v(\xi_p^{(N,i)}) - K_{p,\eta_{p-1}^N}v(\xi_{p-1}^{(N,i)})\big)\,\Big|\,\mathcal{F}_{p-1}^N\Big] \\
&\quad= \frac{1}{N^2}\sum_{i=1}^{N}K_{p,\eta_{p-1}^N}\big((u - K_{p,\eta_{p-1}^N}u)(v - K_{p,\eta_{p-1}^N}v)\big)\big(\xi_{p-1}^{(N,i)}\big),
\end{aligned}$$

其为（26）。取 $u = v$ 得到（27）。

最后，对于任意概率测度 κ，

$$\big|\mathrm{Cov}_\kappa(u,v)\big| \le \sqrt{\mathrm{Var}_\kappa(u)\,\mathrm{Var}_\kappa(v)} \le \frac{1}{4}\,\mathrm{osc}(u)\,\mathrm{osc}(v),$$

因此，对于每个 x,

$$\big|K_{p,\eta}((u - K_{p,\eta}u)(v - K_{p,\eta}v))(x)\big| \le \frac{1}{4}\,\mathrm{osc}(u)\,\mathrm{osc}(v).$$

对 η_{p-1}^N 取平均并乘以$1/N$ 得到（28）。 □

引理 E.5 （局部偏差仅来自二阶余项）。 在与引理*E.4*相同的条件下。固定 $n \ge 1, p \in \{1,\dots,n\}$，以及 $f \in$振幅$_1$（E_n）。令 $\nu_p^N = \Phi_p(\eta_{p-1}^N)$ 如上，并记 $\beta_{p,n} := \beta(D\Phi_{p,n})$。则

$$\Big|\mathbb{E}\Big[\big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\nu_p^N)\big)(f)\Big]\Big| \le \frac{1}{4N}\,(q_{p,n}^2 - 1)\,\beta_{p,n}. \tag{29}$$

特别地，如果 $q_{p,n} = 1$，那么这个局部偏差项为 0。

证明。 对 $\Phi_{p,n}$ 在 $\eta = \nu_p^N$ 和 $\mu = \eta_p^N$ 处应用一阶展开（引理 E.1）：

$$\big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\nu_p^N)\big)(f) = [\eta_p^N - \nu_p^N]\big(D_{\nu_p^N}\Phi_{p,n}(f)\big) + R^{\Phi_{p,n}}(\eta_p^N, \nu_p^N)(f).$$

在给定 $\mathcal{F}_{p-1}^N$ 的条件下取条件期望，并利用（25）消去线性项，因此

$$\mathbb{E}\Big[\big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\nu_p^N)\big)(f)\,\Big|\,\mathcal{F}_{p-1}^N\Big] = \mathbb{E}\big[R^{\Phi_{p,n}}(\eta_p^N, \nu_p^N)(f) \mid \mathcal{F}_{p-1}^N\big].$$

现在使用显式的 Feynman–Kac 余项形式（20）：

$$R^{\Phi_{p,n}}(\eta_p^N, \nu_p^N)(f) = -\frac{1}{\eta_p^N(G_{p,n,\nu_p^N})}\,[\eta_p^N - \nu_p^N](G_{p,n,\nu_p^N})\,[\eta_p^N - \nu_p^N]\big(D_{p,n,\nu_p^N}(f)\big)$$

其中 $D_{p,n,\eta}$是 $D_\eta\Phi_{p,n}$的简写。根据引理 E.1中的比值界，$G_{p,n,\nu_p^N}(x) \in [q_{p,n}^{-1}, q_{p,n}]$，因此

$$\frac{1}{\eta_p^N(G_{p,n,\nu_p^N})} \le q_{p,n}, \qquad \mathrm{osc}(G_{p,n,\nu_p^N}) \le q_{p,n} - q_{p,n}^{-1}.$$

因此，在给定 F_{p-1}^N 的条件下，

$$\begin{aligned}
\Big|\mathbb{E}\big[R^{\Phi_{p,n}}(\eta_p^N, \nu_p^N)(f) \mid \mathcal{F}_{p-1}^N\big]\Big| &\le q_{p,n}\,\mathbb{E}\Big[\big|[\eta_p^N - \nu_p^N](G_{p,n,\nu_p^N})\,[\eta_p^N - \nu_p^N](D_{p,n,\nu_p^N}(f))\big|\,\Big|\,\mathcal{F}_{p-1}^N\Big] \\
&\le q_{p,n}\Big(\mathbb{E}\big[[\eta_p^N - \nu_p^N](G_{p,n,\nu_p^N})^2 \mid \mathcal{F}_{p-1}^N\big]\Big)^{1/2}\Big(\mathbb{E}\big[[\eta_p^N - \nu_p^N](D_{p,n,\nu_p^N}(f))^2 \mid \mathcal{F}_{p-1}^N\big]\Big)^{1/2} \\
&\le q_{p,n}\,\frac{1}{4N}\,\mathrm{osc}(G_{p,n,\nu_p^N})\,\mathrm{osc}(D_{p,n,\nu_p^N}(f)),
\end{aligned}$$

其中最后一步使用（28）与 $u = v$。利用 $q_{p,n}$振幅$(G_{p,n,\nu_p^N}) \le q_{p,n}(q_{p,n} - q_{p}^{-1},_n) = q_{p,n}^2 - 1$可得

$$\Big|\mathbb{E}\big[R^{\Phi_{p,n}}(\eta_p^N, \nu_p^N)(f) \mid \mathcal{F}_{p-1}^N\big]\Big| \le \frac{1}{4N}\,(q_{p,n}^2 - 1)\,\beta_{p,n}.$$

对 $\mathcal{F}_{p-1}^N$ 取期望可得（29）。 □

定理 E.6（序贯蒙特卡洛采样 全变差偏差界）。 对于与 *FK* 模型相关联的序贯蒙特卡洛采样器，任意 $N \geq 1$ 和任意 $n \geq 1$,

$$\|\bar{\eta}_n^N - \eta_n\|_{\mathrm{TV}} \leq \frac{1}{4N} \sum_{p=1}^{n} (q_{p,n}^2 - 1)\, \beta(D\Phi_{p,n}). \tag{30}$$

特别地，如果 $q_{p,\,n} = 1$ 对所有 $p \leq n$ 均成立（例如势函数完全平坦），那么 $\bar{\eta}_n^N = \eta_n$。

定理 *E.6* 的证明。 任取 $f \in$ 振幅$_1$（E_n）。根据偏差望远镜求和（24），

$$[\bar{\eta}_n^N - \eta_n](f) = \sum_{p=1}^{n} \mathbb{E}\Big[\big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\Phi_p(\eta_{p-1}^N))\big)(f)\Big].$$

由于 $\Phi_p(\eta_{p-1}^N) = \nu_p^N$，每一项均可由引理 E.5加以界定，因此

$$|[\bar{\eta}_n^N - \eta_n](f)| \leq \frac{1}{4N} \sum_{p=1}^{n} (q_{p,n}^2 - 1)\, \beta(D\Phi_{p,n}).$$

对 $f \in$ 振幅$_1$（E_n）取上确界可得（30）。 □

定理*5.1* 和定理*D.1*的证明。从局部总变差偏差界（定理E.6）开始:

$$\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \leq \frac{1}{4N} \sum_{p=1}^{T+1} (q_{p,T+1}^2 - 1)\, \beta(D\Phi_{p,T+1}). \tag{31}$$

由引理E.1，$\beta(D\Phi_{p,\,T+1}) \leq 2q_{p,\,T+1}\beta(P_{p,\,T+1}) \leq 2q_{p,\,T+1}$，因此由（31）可得

$$\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \leq \frac{1}{2N} \sum_{p=1}^{T+1} q_{p,T+1}\,(q_{p,T+1}^2 - 1). \tag{32}$$

注意，最后一项 $p = T+1$ 会消失，因为 $q_{T+1,\,T+1} = 1$。

(i) 朴素命题 osal. 由引理 E.2，对于 $p \leq T$, $q_{p,\,T+1} \leq L^2(1+\varepsilon)^{2(T-p)}$。利用 $q(q^2-1) \leq q^3$ 对于 $q \geq 1$,

$$\sum_{p=1}^{T} q_{p,T+1}(q_{p,T+1}^2 - 1) \leq \sum_{p=1}^{T} q_{p,T+1}^3 \leq L^6 \sum_{p=1}^{T} (1+\varepsilon)^{6(T-p)} = L^6 \sum_{j=0}^{T-1} a^j, \quad a := (1+\varepsilon)^6 \geq 1.$$

利用初等界 $\sum_{j=0}^{m-1} a^j \leq m\, a^{m-1}$ 对于 $a \geq 1$（其中 $m = T$），可得

$$\sum_{p=1}^{T} q_{p,T+1}(q_{p,T+1}^2 - 1) \leq L^6\, T\, (1+\varepsilon)^{6(T-1)}.$$

代入（32）可得

$$\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \leq \frac{L^6\, T\, (1+\varepsilon)^{6(T-1)}}{2N},$$

因此 $N \geq \frac{L^6\, T\, (1+\varepsilon^{\,6(T-1)})}{2\,\delta_{\mathrm{TV}}}$保证了 $\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \leq \delta_{\mathrm{TV}}$。

(ii) 最优提议分布。由引理 E.2，对于 $p \leq T$, $q_{p,T+1} \leq (1+\varepsilon)^{2(T+1-p)}$。令 $k := T+1-p \in \{1, \ldots, T\}$。于是

$$q_{p,T+1}(q_{p,T+1}^2 - 1) \leq (1+\varepsilon)^{2k}\big((1+\varepsilon)^{4k} - 1\big) \leq k\big((1+\varepsilon)^4 - 1\big)(1+\varepsilon)^{6k},$$

其中我们对满足 $b = (1+\varepsilon)^4$ 的 $b \geq 1$ 使用了 $b^k - 1 \leq k(b-1)b^k$。因此

$$\sum_{p=1}^{T} q_{p,T+1}(q_{p,T+1}^2 - 1) \leq \big((1+\varepsilon)^4 - 1\big) \sum_{k=1}^{T} k\, a^k, \qquad a := (1+\varepsilon)^6 \geq 1.$$

利用粗略界 $\sum_{k=1}^{m} k\, a^k \leq m^2 a^m$（针对 $a \geq 1$，此处 $m = T$），我们得到

$$\sum_{p=1}^{T} q_{p,T+1}(q_{p,T+1}^2 - 1) \leq \big((1+\varepsilon)^4 - 1\big)\, T^2\, (1+\varepsilon)^{6T}.$$

代入（32）可得

$$\|\bar{\eta}^N_{T+1} - \eta_{T+1}\|_{\mathrm{TV}} \le \frac{\left((1+\varepsilon)^4 - 1\right) T^2 \, (1+\varepsilon)^{6T}}{2N},$$

因此 $N \ge \frac{((1+\varepsilon)^4-1)\quad T^2\ 1+\varepsilon\ ^{6T}\,()}{2\,\delta_{\mathrm{TV}}}$保证 $\|\bar{\eta}^N_{T+1} - \eta_{T+1}\|_{\mathrm{TV}} \le \delta_{\mathrm{TV}}$。 □

命题 E.7（任意提议分布的粒子复杂度）。 固定时域 $T \ge 2$ 和目标精度 $\delta_{\mathrm{TV}} \in (0,\ 1)$。在假设 *3.1* 和 *3.2*下，令 $\bar{\eta}^N_{T+1}$:= $\mathbb{E}[\eta^N_{T+1}]$ 为序贯蒙特卡洛采样器的输出分布（经过终端重采样步骤T+1，其中 $M_{T+1} = \mathrm{Id}$）。此外，我们假设

$$\max_{t\in[T]} \max_{s_{1:t}\in\mathcal{S}_{1:t}} \max \left\{ \frac{\pi_{\mathrm{ref}}(s_t \mid s_{1:t-1})}{\pi_{\mathrm{p}}(s_t \mid s_{1:t-1})} \cdot \frac{\hat{V}(s_0, s_{1:t})}{\hat{V}(s_0, s_{1:t-1})}, \frac{\pi_{\mathrm{p}}(s_t \mid s_{1:t-1})}{\pi_{\mathrm{ref}}(s_t \mid s_{1:t-1})} \cdot \frac{\hat{V}(s_0, s_{1:t-1})}{\hat{V}(s_0, s_{1:t})} \right\} \le L_{\mathrm{p}}.$$

那么，保证 $\|\bar{\eta}^N_{T+1} - \eta_{T+1}\|_{\mathrm{TV}} \le \delta_{\mathrm{TV}}$ 的充分条件是

$$N \ge \frac{L_{\mathrm{p}}^6 T (1+\varepsilon)^{6(T-1)}}{2\delta_{\mathrm{TV}}}.$$

证明。 证明遵循与定理5.1类似的模式。 □

注 E.8。命题 E.7 表明，选择提议分布 π_{p} 可以平衡粒子复杂度所带来的开销与执行提议分布本身的开销。设计良好的提议分布可能具有适中的 L_{p}，但可能难以从中采样，或者其对应的增量权重 $G_t = {}^{\pi_{\mathrm{ref}}(s_t|s_{1:t-1})}\,\pi_{\mathrm{p}}(s_t|s_{1:t-1}) \cdot \hat{}v(s_0,s_{1:t})\ \hat{}v(s_0, s_{1:t-1})$ 可能难以计算到所需精度。从这个意义~~上说，朴素提议分布~~（推论 5.2）和最优提议分布（命题 D.4）代表了两个极端。朴素提议分布极易采样，且无需进一步计算即可利用奖励模型直接得到增量权重，但粒子复杂度在最坏情况下可能按 $O(L^6)$ 增长。另一方面，最优提议分布的粒子复杂度可能极低，按 $O(\varepsilon T)$ 增长；当贝尔曼误差足够小时，甚至可能退化为只需一个粒子。但估计增量权重的成本相对较高。因此，粒子数与提议分布的难度（包括增量权重的计算）之间存在重要的权衡。

我们还要指出，这里考虑的是句子级推理算法。在词元级场景中，可以获取后续词元的显式概率，因此权衡也可能有所不同。不过，粒子复杂度与定理 5.1 和命题 E.7 中的完全相同。

命题 *D.4* 的证明。 在此证明中，每当我们写下 $\mathbb{E}[X|\mathcal{F}、\ E]$ 对于某个随机变量 X、σ-代数 $\mathcal{F}$和事件E时，我们指的是 $\mathbb{E}[X|\sigma(\mathcal{F} \cup \{E\})]\mathbb{I}_E$，因为我们通常不关心 E之外发生的事情。接着不难看出，由于

$$\mathbb{E}[X|\sigma(\mathcal{F} \cup \{E\})] = \mathbb{E}[X\mathbb{I}_E|\mathcal{F}]\mathbb{I}_E/\mathbb{P}(E|\mathcal{F}) + \mathbb{E}[X\mathbb{I}_{E^c}|\mathcal{F}]\mathbb{I}_{E^c}/\mathbb{P}(E^c|\mathcal{F}),$$

我们有

$$\mathbb{E}[X|\mathcal{F}, E] = \mathbb{E}[X\mathbb{I}_E|\mathcal{F}]\mathbb{I}_E/\mathbb{P}(E|\mathcal{F}).$$

设置与滤过。 我们的 FK 时域为 $n := T + 1$，其中终止步骤 T+1 使用 $M_{T+1} = \mathrm{Id}$ 并应用最后一个势函数G_T（不会生成新词元）。蒙特卡洛归一化常数估计和拒绝采样提议仅发生在词元步骤 $t = 1, \ldots, T$。

给定 N 个粒子，它们在算法中总共运行 T 个词元生成步骤。在每个此类步骤中，我们使用蒙特卡洛估计 N 个条件期望，每个期望都通过对 N_m 个样本取平均来估计。对于每个 $t \in [T], i \in [N]$，该估计表示为一个随机变量

$$\hat{Z}_{t,i} := \frac{1}{N_m} \sum_{m=1}^{N_m} \hat{V}\big(s_0, (S_{1:t-1}, \tilde{S}_t^{[m]})\big),$$

其中，每个 $\tilde{S}_t^{[m]}$ 都在给定 $\mathrm{F}_{t-1}^{N\ i-1}$,的条件下独立抽取自 π_{ref}，F表示由之前所有粒子（截至时间 $t-1$）生成的σ-代数，其中包括使用拒绝采样获得最优提议分布，以及蒙特卡洛估计随机变量。由索引标记的σ-代数族在样本空间 Ω上构成自然滤过：

$$(\mathcal{F}_1^{N,0}, \mathcal{F}_1^{N,1}, \mathcal{F}_1^{N,2}, \ldots, \mathcal{F}_1^{N,N}, \mathcal{F}_2^{N,0}, \ldots, \mathcal{F}_2^{N,N}, \ldots, \mathcal{F}_T^{N,0}, \ldots, \mathcal{F}_T^{N,N}).$$

理想情况下，每个 $\hat{}Z_{t,\ i}$ 都是对.....的良好估计

$$Z_{t,i} := \mathbb{E}_{s_t\sim\pi_{\mathrm{ref}}(\cdot|s_0,S_{1:t-1})}[\hat{V}(s_0, (S_{1:t-1}, s_t))] = \mathbb{E}[\hat{Z}_{t,i}|\mathcal{F}_{t-1}^{N,i-1}].$$

蒙特卡洛估计的良好事件。现在我们为任意 $\delta > 0$ 定义失败事件，并且

$$\xi := 2\sqrt{2}\sqrt{\frac{C_{\text{act}} \log \frac{4}{\delta}}{N_m}}$$

as

$$\mathcal{B}_{t,i} := \left\{ \left| \frac{\hat{Z}_{t,i}}{Z_{t,i}} - 1 \right| > \xi \right\}.$$

此处 C_{act} 在引理 B.5中定义。
由于 $\hat{Z}_{t,i}, Z_{t,i}$ 都是 $\mathrm{F}_{t-1}^{N\,i}$,-可测的，因此我们有 $\mathrm{B}_{t,i} \in \mathrm{F}_{t-1}^{N\,i}$,。
然后我们使用引理B.5的条件版本，得到

$$\mathbb{P}(\mathcal{B}_{t,i}^{c} | \mathcal{F}_{t-1}^{N,i-1}) \geq 1 - \delta \qquad a.s.$$

那么显然

$$\mathbb{P}\left(\bigcup_{t \leq T} \bigcup_{i \leq N} \mathcal{B}_{t,i} \right) \leq \sum_{\substack{t \leq T \\ i \leq N}} \mathbb{E}\Big[\mathbb{E}[\mathbb{I}_{\mathcal{B}_{t,i}} | \mathcal{F}_{t-1}^{N,i-1}] \Big] < NT\delta.$$

拒绝采样（最优提议分布）的良好事件。随后，由于我们的最优提议分布是通过拒绝采样（算法 3）采样得到的，我们需要定义坏事件，使得采样失败。考虑随机变量 I_t，$_i \in \{0, 1\}$，用于表示从时间步 $t-1$ 到 t 时第 i 个粒子的提议是否成功；也就是说，令 E_{acc} 表示定理 B.1 中 (t, i) 处的良好事件，$I_{t,i} := \mathbb{I}_{\mathcal{E}_{acc}}$，则 $\mathrm{B}_{t,i}' := \{I_{t,i} = 0\}$。将 (t, i) 处的采样结果记为随机变量 $W_{t,i}$。那么定理 B.1 的条件版本表明，当至多使用 $N_s := 4C_{\text{act}}\log(\delta^{-1})$ 个提议-拒绝步骤时，$\mathbb{P}(\mathrm{B}_{t,i}' | \mathrm{F}_{t-1}^{N\,N},) < \delta$。因此

$$\mathbb{P}\left(\bigcup_{t \leq T} \bigcup_{i \leq N} \mathcal{B}_{t,i}' \right) \leq \sum_{\substack{t \leq T \\ i \leq N}} \mathbb{E}\Big[\mathbb{E}[\mathbb{I}_{\mathcal{B}_{t,i}'} | \mathcal{F}_{t-1}^{N,N}] \Big] < NT\delta.$$

注意，$\sigma(I_t,1, W_t,1)$、$\sigma(I_t,2, W_t,2)$、$\cdots$、$\sigma(I_{t,N}, W_{t,N})$ 关于 $\mathrm{F}_{t-1}^{N\,N}$ 条件独立。并且 $\mathrm{F}_{t}^{N\,0}$ $=,\sigma\Big(\mathrm{F}_{t-1}^{N\,N} \cup \Big(\bigcup_{i=1}^{N} \sigma, (I_{t,i}, W_{t,i})\Big)\Big)$。

现在，我们为估计条件期望定义每一步的良好事件 $\mathrm{E}_t := (\bigcup_{i=1}^{N} \mathrm{B}_{t,i})^c$（对于 $t \leq T$），并将总体良好事件定义为其交集 $\mathrm{E} := \bigcap_{t \leq T} \mathrm{E}_t$、$\mathrm{E}_{\leq t} := \bigcap_{t' \leq t} \mathrm{E}_{t'}$、$\mathrm{E}_{>t} := \bigcap_{t < t' \leq T} \mathrm{E}_{t'}$。我们还有 $\mathbb{P}(\mathrm{E}_t) \geq 1 - N\delta$、$\mathbb{P}(\mathrm{E}_{\leq t}) \geq 1 - NT\delta$ 以及 $\mathbb{P}(\mathrm{E}_{>t}) \geq 1 - NT\delta$。对于拒绝采样，每一步的良好事件为 $\mathrm{E}_t' := \bigcap_{i=1}^{N} \mathrm{E}_{t,i}'$，其中 $\mathrm{E}_{t,i}' := (\mathrm{B}_{t,i}')^c$（对于 $t \leq T$），以及 $\mathrm{E}' := \bigcap_{t \leq T} \mathrm{E}_t'$、$\mathrm{E}_{\leq t}' := \bigcap_{t' \leq t} \mathrm{E}_{t'}'$、$\mathrm{E}_{>t}' := \bigcap_{t < t' \leq T} \mathrm{E}_{t'}'$。我们还有 $\mathbb{P}(\mathrm{E}_t') \geq 1 - N\delta$、$\mathbb{P}(\mathrm{E}_{\leq t}') \geq 1 - NT\delta$ 以及 $\mathbb{P}(\mathrm{E}_{>t}') \geq 1 - NT\delta$。

望远镜求和（包含终止步骤 $n = T + 1$）。为应用我们先前关于望远镜求和的证明，对于任意测试函数 $f \in$ 振幅$_1$，我们定义随机变量

$$Y_p := \Big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\Phi_p(\eta_{p-1}^N)) \Big)(f), \qquad n := T + 1.$$

注意，Y_p 是作用于测试函数 f的两个随机概率测度之差，因此 $\|Y_p\|_\infty \leq \mathrm{osc}(f) = 1$。对于终端 FK 步骤 $p = T + 1$，不存在拒绝采样；我们可以令 $\mathrm{E}_{T+1}' := \Omega$ 和 $\mathrm{E}_{T+1} := \Omega$，从而使下文的记号保持统一。

于是我们有

$$\begin{aligned} \mathbb{E}[Y_p | \mathcal{E} \cap \mathcal{E}'] = \mathbb{E}[Y_p \mathbb{I}_{\mathcal{E}} \mathbb{I}_{\mathcal{E}'}] / \mathbb{P}(\mathcal{E} \cap \mathcal{E}') &= \mathbb{E}[Y_p \mathbb{I}_{\mathcal{E}_{\leq p}} \mathbb{I}_{\mathcal{E}'_{<p}} \mathbb{I}_{\mathcal{E}'_p} \mathbb{I}_{\mathcal{E}'_{>p}} \mathbb{I}_{\mathcal{E}_{>p}}] / \mathbb{P}(\mathcal{E} \cap \mathcal{E}') \\ &= \mathbb{E}\Big[\mathbb{I}_{\mathcal{E}_{\leq p}} \mathbb{I}_{\mathcal{E}'_{<p}} \mathbb{E}\big[Y_p \mathbb{I}_{\mathcal{E}'_p} \mathbb{I}_{\mathcal{E}'_{>p}} \mathbb{I}_{\mathcal{E}_{>p}} \big| \mathcal{F}_{p-1}^{N,N} \big] \Big] / \mathbb{P}(\mathcal{E} \cap \mathcal{E}'), \end{aligned}$$

最后一步是因为 $\mathbb{I}_{\mathcal{E}_{\leq p}} \mathbb{I}_{\mathcal{E}'_{<p}}$ 是 $\mathrm{F}_{p-1}^{N\,N}$,-可测的。

接下来我们关注 $\left|\mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}_{>p}} \mathbb{I}_{\mathcal{E}'_p} \mathbb{I}_{\mathcal{E}'_{>p}} | \mathrm{F}_{p-1}^{N,N}\right]\right|$，它可以按如下方式进行界定

$$\begin{aligned}
\left|\mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}_{>p}} \mathbb{I}_{\mathcal{E}'_p} \mathbb{I}_{\mathcal{E}'_{>p}} | \mathcal{F}_{p-1}^{N,N}\right]\right| &= \left|\mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}'_p} | \mathcal{F}_{p-1}^{N,N}\right] - \mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}'_p} \mathbb{I}_{\mathcal{E}^c_{>p} \cup (\mathcal{E}'_{>p})^c} | \mathcal{F}_{p-1}^{N,N}\right]\right| \\
&\leq \left|\mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}'_p} | \mathcal{F}_{p-1}^{N,N}\right]\right| + \underbrace{\|Y_p \mathbb{I}_{\mathcal{E}'_p}\|_\infty}_{\leq 1} \cdot \underbrace{\mathbb{P}(\mathcal{E}^c_{>p} \cup (\mathcal{E}'_{>p})^c | \mathcal{F}_{p-1}^{N,N})}_{< 2NT\delta} \\
&< \left|\mathbb{E}\left[Y_p | \mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p\right]\right| + \underbrace{\left|\mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}'_p} | \mathcal{F}_{p-1}^{N,N}\right]\right|}_{\leq 1} \mathbb{I}_{(\mathcal{E}'_p)^c} + 2NT\delta.
\end{aligned}$$

其中我们利用了如下事实：$\mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}'_p} | \mathrm{F}_{p-1}^{N,N}\right] \mathbb{I}_{\mathcal{E}'_p} = \mathbb{E}[Y_p | \mathrm{F}_{p-1}^{N,N}, \mathrm{E}'_p] \cdot \mathbb{P}(\mathrm{E}'_p | \mathrm{F}_{p-1}^{N,N})$。

因此我们写成

$$\begin{aligned}
\mathbb{E}[Y_p | \mathcal{E} \cap \mathcal{E}'] &= \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_{\leq p}} \mathbb{I}_{\mathcal{E}'_{<p}} \mathbb{E}\left[Y_p \mathbb{I}_{\mathcal{E}'_p} \mathbb{I}_{\mathcal{E}'_{>p}} \mathbb{I}_{\mathcal{E}_{>p}} | \mathcal{F}_{p-1}^{N,N}\right]\right] / \mathbb{P}(\mathcal{E} \cap \mathcal{E}') \\
&< \mathbb{E}\left[\mathbb{I}_{\mathcal{E}_{\leq p}} \mathbb{I}_{\mathcal{E}'_{<p}} \left(\left|\mathbb{E}[Y_p | \mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p]\right| + \mathbb{I}_{(\mathcal{E}'_p)^c} + 2NT\delta\right)\right] / \mathbb{P}(\mathcal{E} \cap \mathcal{E}') \\
&\leq \frac{1}{1 - 2NT\delta} \left(\sup_{\mathcal{E}_{\leq p}} \left|\mathbb{E}\left[Y_p \mid \mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p\right]\right| + N\delta + 2NT\delta\right).
\end{aligned}$$

（a）与扰动 FK 映射的联系。 令 Ψ_G 表示玻尔兹曼–吉布斯变换

$$\Psi_G(\mu)(dx) = \frac{G(x)\mu(dx)}{\mu(G)}.$$

在 $\omega \in \mathrm{E}_{\leq p}$ 下，对于（随机）归一化常数，我们有

$$(1 - \xi) Z_{p,i} \leq \hat{Z}_{p,i} \leq (1 + \xi) Z_{p,i}, \qquad \forall i \in [N].$$

特别地，所诱导的势函数比满足一致界

$$1 - \xi \;\leq\; \frac{G'_p}{G_p} \;\leq\; 1 + \xi \qquad \Rightarrow \qquad \left|\frac{G'_p}{G_p} - 1\right| \leq \xi.$$

经过标准计算可得，对于任意概率测度 μ 以及任意 $f \in$ 振幅$_1$，

$$\left|(\Psi_{G'_p}(\mu) - \Psi_{G_p}(\mu))(f)\right| = \left|\frac{\mu(G'_p f)}{\mu(G'_p)} - \frac{\mu(G_p f)}{\mu(G_p)}\right| \leq \frac{2\xi}{1 - \xi},$$

因此

$$\|\Psi_{G'_p}(\mu) - \Psi_{G_p}(\mu)\|_{\mathrm{TV}} \;\leq\; \frac{2\xi}{1 - \xi}. \tag{33}$$

由于马尔可夫核的变异操作具有全变差收缩性（且 $M_{T+1} = \mathrm{Id}$ 是一种特殊情况），相应的一步 FK 映射满足

$$\|\Phi'_p(\mu) - \Phi_p(\mu)\|_{\mathrm{TV}} \;\leq\; \frac{2\xi}{1 - \xi}, \qquad \forall \mu. \tag{34}$$

（b）局部（线性）偏差界。 回顾 $u = D_{\nu_p^N} \Phi_{p,\ n}(f)$ 和 $\nu_p^N = \Phi_p(\eta_{p-1}^N)$。根据 Φ'_p 的定义以及上述条件期望恒等式，

$$\begin{aligned}
\mathbb{E}\Big([\eta_p^N - \nu_p^N](u) \,\Big|\, \mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p\Big) &= \frac{1}{N} \sum_{i=1}^{N} \left(K'_{p,\eta_{p-1}^N} u(\xi_{p-1}^{(N,i)}) - K_{p,\eta_{p-1}^N} u(\xi_{p-1}^{(N,i)})\right) \\
&= \eta_{p-1}^N \Big((K'_{p,\eta_{p-1}^N} - K_{p,\eta_{p-1}^N}) u\Big) \\
&= \left(\Phi'_p(\eta_{p-1}^N) - \Phi_p(\eta_{p-1}^N)\right)(u).
\end{aligned}$$

对于任意有界可测的 g，利用中心化技巧 $g = g_0 + c$，结合 $\|g_0\|_\infty \le \frac{1}{2}$振幅$(g)$ 和 $(\mu-\nu)(g) = (\mu-\nu)(g_0)$，我们有$|(\mu-\nu)(g)| \le \frac{1}{2}$振幅$(g)\|\mu-\nu\|_{\mathrm{TV}}$。将其应用于 $g = u$并结合（34），可得，对于每个 $\omega \in \mathrm{E}_{\le p}$，

$$\begin{aligned}\left|\mathbb{E}\Big([\eta_p^N - \nu_p^N](u)\,\big|\,\mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p\Big)\right| &\le \frac{1}{2}\,\mathrm{osc}(u)\,\|\Phi'_p(\eta_{p-1}^N) - \Phi_p(\eta_{p-1}^N)\|_{\mathrm{TV}} \\ &\le \frac{1}{2}\,\beta(D\Phi_{p,n}) \cdot \frac{2\xi}{1-\xi} = \frac{\xi}{1-\xi}\,\beta(D\Phi_{p,n}),\end{aligned} \tag{35}$$

其中我们利用了 $\mathrm{osc}(u) \le \beta(D\Phi_{p,n})\mathrm{osc}(f) \le \beta(D\Phi_{p,n})$，因为 $f \in$振幅$_1$。

(c) 二阶余项。根据定理 E.6 中相同的论证，在给定 $\mathrm{F}_{p-1}^{N,N}$,E'_p 的条件下，

$$\sup_{f \in \mathrm{Osc}_1} \left|\mathbb{E}\left[R^{\Phi_{p,n}}(\eta_p^N, \nu_p^N)(f)\,\middle|\,\mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p\right]\right| \le \frac{1}{4N}\,(q_{p,n}^2 - 1)\,\beta(D\Phi_{p,n}). \tag{36}$$

(d) 合并 (b)(c) 得到 $\sup_{\mathcal{E}_{\le p}} |\mathbb{E}[Y_p\,|\,\mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p]|$。合并 (35) 和 (36，对每个 $\omega \in \mathcal{E}_{\le p}$，

$$\left|\mathbb{E}\left[\big(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\nu_p^N)\big)(f)\,\middle|\,\mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p\right]\right| \le \frac{\xi}{1-\xi}\,\beta(D\Phi_{p,n}) + \frac{1}{4N}\,(q_{p,n}^2 - 1)\,\beta(D\Phi_{p,n}).$$

因此，

$$\sup_{\mathcal{E}_{\le p}} \left|\mathbb{E}\big[Y_p \mid \mathcal{F}_{p-1}^{N,N}, \mathcal{E}'_p\big]\right| \le \frac{\xi}{1-\xi}\,\beta(D\Phi_{p,n}) + \frac{1}{4N}\,(q_{p,n}^2 - 1)\,\beta(D\Phi_{p,n}). \tag{37}$$

(e) 对 p (t 求和逐项相消）以控制 E 下的全变差。取上确界$_{f \in \text{振幅}_1}$并对 $p \le n = T+1$ 求和，得到 s

$$\begin{aligned}\big\|\tilde{\pi}_T - \hat{\pi}_T \mid \mathcal{E}\big\|_{\mathrm{TV}} &= \sup_{f \in \mathrm{Osc}_1} \left|\sum_{p=1}^{T+1} \mathbb{E}[Y_p \mid \mathcal{E}]\right| \le \sum_{p=1}^{T+1} \sup_{f \in \mathrm{Osc}_1} \big|\mathbb{E}[Y_p \mid \mathcal{E}]\big| \\ &\le \sum_{p=1}^{T+1} \left[\frac{1}{1-2NT\delta}\left(\frac{\xi}{1-\xi}\,\beta(D\Phi_{p,T+1}) + \frac{1}{4N}\,(q_{p,T+1}^2 - 1)\,\beta(D\Phi_{p,T+1})\right) + \frac{2NT\delta + N\delta}{1-2NT\delta}\right].\end{aligned} \tag{38}$$

(f) 代入精确的 FK 半群稳定性常数。根据假设 3.2，精确的 *FK* 半群满足通常的界（终止时域为 $T+1$）：

$$q_{p,T+1} \;\le\; (1+\varepsilon)^{2(T+1-p)}, \qquad \beta(D\Phi_{p,T+1}) \;\le\; 2q_{p,T+1} \;\le\; 2(1+\varepsilon)^{2(T+1-p)}.$$

因此，

$$\sum_{p=1}^{T+1} \beta(D\Phi_{p,T+1}) \le 2\sum_{k=0}^{T} (1+\varepsilon)^{2k} \le 2(T+1)(1+\varepsilon)^{2T}, \tag{39}$$

$$\sum_{p=1}^{T+1} (q_{p,T+1}^2 - 1)\beta(D\Phi_{p,T+1}) \le 2\sum_{p=1}^{T+1} q_{p,T+1}(q_{p,T+1}^2 - 1) \;\lesssim\; \varepsilon\,(T+1)^2\,(1+\varepsilon)^{6T}, \tag{40}$$

其中（40）是与我们先前最优提议分析中使用的、带有k 因子的几何级数界相同。在$\varepsilon = O\,(1/T)$ 且 $\delta_{\mathrm{TV}} = O\,(1)$ 的情况下，指数因子可并入常数中。

(g) 选择参数 N、N_m 并实例化置信度 δ。首先，将上文使用的 δ 实例化为 $\delta \leftarrow \delta/(4N(T+1))$，使得 $\mathbb{P}(\mathrm{E} \cap \mathrm{E}') \ge 1-\delta$。接着，选择 $\xi \le 1/2$ 并设

$$\xi \asymp \frac{\delta_{\mathrm{TV}}}{T+1}, \qquad N \asymp \frac{\varepsilon(T+1)^2}{\delta_{\mathrm{TV}}}.$$

采用这些选择，利用（38）–（40）且 $2NT\delta$可忽略不计，式（38）右侧为 $\le \delta_{\mathrm{TV}}$。最后，回顾一下 $\xi = 2\sqrt{2}\sqrt{C_{\mathrm{act}}\log \frac{4}{\delta}}$。因此，只需取

$$N_m \asymp \frac{C_{\mathrm{act}}\,(T+1)^2}{\delta_{\mathrm{TV}}^2}\,\log\left(\frac{N(T+1)}{\delta}\right).$$

最后回顾一下，我们完成最优提议分布的拒绝采样花费了 $N_s := 4C_{\text{act}}\log(\delta^{-1})$ 时间。值得注意的是，与 N_m 相比，这可以忽略不计，因为拒绝采样只需将归一化因子估计到常数相对误差，而后者则需要高达 $O(\delta_{\text{TV}}/T)$。

（h）时间复杂度。 奖励模型评估的总次数（或蒙特卡洛归一化常数估计中的参考核调用次数）为 $N\cdot T\cdot N_m$ 阶（终止步骤 $T{+}1$ 仅会增加 $O(N)$ 的重采样成本），因此

$$\text{Time} \asymp NT\,(N_m + N_s) = \tilde{O}\left(\frac{\varepsilon(T+1)^2}{\delta_{\text{TV}}}\cdot T\cdot\frac{C_{\text{act}}(T+1)^2}{\delta_{\text{TV}}^2}\log(\delta^{-1})\right) = \tilde{O}\left(\frac{\varepsilon\, C_{\text{act}}\, T^5\log(\delta^{-1})}{\delta_{\text{TV}}^3}\right).$$

此外，由于 $C_{\text{act}} \le L(1+\varepsilon)$，在 $\varepsilon = O(1/T)$ 的范围内，我们可以写成

$$\text{Time} = \tilde{O}\left(\frac{\varepsilon\, L\, T^5\log(\delta^{-1})}{\delta_{\text{TV}}^3}\right),$$

由此即可证明该结果。 □

使用任意 McKean 核的序贯蒙特卡洛。 以下命题给出了定理E.6的更一般形式，区分了不同的 McKean 核 $K_{p,\ \eta}$。尽管不同的核可能对应于同一个费曼-卡茨流，即只要它们都满足（1），定理 E.6 仅对所有这些核给出一个上界，而命题 E.9则刻画了它们之间的区别。

命题 E.9。 *考虑一个使用 McKean 转移核的序贯蒙特卡洛采样器 $K_{p,\ \eta}$ 它满足 （1），对应于一个 FK 模型；那么对于任意 $N \ge 1$ 和任意 $n \ge 1$，令 $\nu :\ = \Phi_p(\eta_{p-1}^N)$，我们有*

$$\|\bar{\eta}_n^N - \eta_n\|_{\text{TV}} \le \frac{1}{2N}\sum_{p=1}^{n}\left(q_{p,n}\,\beta(D\Phi_{p,n})\cdot\mathbb{E}\left[\sqrt{\eta_{p-1}^N\left(K_{p,\eta_{p-1}^N}\left((G_{p,n,\nu} - K_{p,\eta_{p-1}^N}G_{p,n,\nu})^2\right)\right)}\right]\right). \tag{41}$$

命题的证明*E.9.* 固定 $n \ge 1$与 $N \ge 1$，并令 $f \in$ 振幅$_1(E_n)$。根据偏差望远镜求和（24），

$$[\bar{\eta}_n^N - \eta_n](f) = \sum_{p=1}^{n}\mathbb{E}\left[\left(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\Phi_p(\eta_{p-1}^N))\right)(f)\right].$$

对每个 p，令 $\nu := \Phi_p(\ \eta_{p-1}^N) = \nu_p^N$。将一阶展开（引理 E.1）应用于 $\Phi_{p,\ n}$ 在 $(\eta,\ \mu) = (\nu,\ \eta_p^N)$ 给出 s

$$\left(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\nu)\right)(f) = [\eta_p^N - \nu]\left(D_{p,n,\nu}(f)\right) + R^{\Phi_{p,n}}(\eta_p^N,\nu)(f).$$

取给定 $\mathcal{F}_{p-1}^N$ 的条件期望，并利用引理 E.4–(25) 消去线性项，因此

$$\mathbb{E}\left[\left(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\nu)\right)(f)\,\middle|\,\mathcal{F}_{p-1}^N\right] = \mathbb{E}\left[R^{\Phi_{p,n}}(\eta_p^N,\nu)(f)\mid\mathcal{F}_{p-1}^N\right].$$

利用显式的 FK 余项形式（20），

$$R^{\Phi_{p,n}}(\eta_p^N,\nu)(f) = -\frac{1}{\eta_p^N(G_{p,n,\nu})}\,[\eta_p^N - \nu](G_{p,n,\nu})\,[\eta_p^N - \nu]\left(D_{p,n,\nu}(f)\right),$$

其中再次，$D_{\ p,\ n,\ \eta}$是 $D_\eta\Phi_{p,\ n}$的简写。根据 $q_{p,\ n}$的定义，我们有 $1/\eta_p^N(G_{p,\ n,\ \nu}) \le q_{p,\ n}$，因此 ce

$$\begin{aligned}\left|\mathbb{E}\left[R^{\Phi_{p,n}}(\eta_p^N,\nu)(f)\mid\mathcal{F}_{p-1}^N\right]\right| &\le q_{p,n}\,\mathbb{E}\left[\left|[\eta_p^N - \nu](G_{p,n,\nu})\right|\left|[\eta_p^N - \nu]\left(D_{p,n,\nu}(f)\right)\right|\,\middle|\,\mathcal{F}_{p-1}^N\right]\\ &\le q_{p,n}\left(\mathbb{E}\left[[\eta_p^N - \nu](G_{p,n,\nu})^2\mid\mathcal{F}_{p-1}^N\right]\right)^{1/2}\left(\mathbb{E}\left[[\eta_p^N - \nu](D_{p,n,\nu}(f))^2\mid\mathcal{F}_{p-1}^N\right]\right)^{1/2}.\end{aligned}$$

根据引理 E.4–(27) 应用于 $u = G_{p,\ n,\ \nu}$，

$$\mathbb{E}\left[[\eta_p^N - \nu](G_{p,n,\nu})^2\mid\mathcal{F}_{p-1}^N\right] = \frac{1}{N}\,\eta_{p-1}^N\left(K_{p,\eta_{p-1}^N}\left((G_{p,n,\nu} - K_{p,\eta_{p-1}^N}G_{p,n,\nu})^2\right)\right).$$

对于 D 项，引理 E.4和初等界 $\eta K((h\ - Kh)^2) \le \frac{1}{4}$振幅$(h)^2$给出 -

$$\mathbb{E}\left[[\eta_p^N - \nu](D_{p,n,\nu}(f))^2\mid\mathcal{F}_{p-1}^N\right] = \frac{1}{N}\eta_{p-1}^N K_{p,\eta_{p-1}^N}\left((D_{p,n,\nu}(f) - K_{p,\eta_{p-1}^N}D_{p,n,\nu}(f))^2\right) \le \frac{1}{4N}\,\text{osc}(D_{p,n,\nu}(f))^2.$$

利用 $\operatorname{osc}(D_{p,\,n,\,\nu}(f)) \le \beta(D\Phi_{p,\,n})\operatorname{osc}(f) \le \beta(D\Phi_{p,\,n})$，对于 $f \in$振幅$_1(E_n)$，可得

$$\left(\mathbb{E}\left[[\eta_p^N - \nu](D_{p,n,\nu}(f))^2 \mid \mathcal{F}_{p-1}^N\right]\right)^{1/2} \le \frac{1}{2\sqrt{N}}\,\beta(D\Phi_{p,n}).$$

结合以上各式可得，对于任意 $f \in$振幅$_1(E_n)$，

$$\left|\mathbb{E}\left[\left(\Phi_{p,n}(\eta_p^N) - \Phi_{p,n}(\nu)\right)(f)\right]\right| \le \frac{1}{2N}\,q_{p,n}\,\beta(D\Phi_{p,n})\,\mathbb{E}\left[\sqrt{\eta_{p-1}^N\left(K_{p,\eta_{p-1}^N}((G_{p,n,\nu} - K_{p,\eta_{p-1}^N}G_{p,n,\nu})^2)\right)}\right].$$

对 $p = 1, \dots, n$ 求和，并对 $f \in$振幅$_1(E_n)$ 取上确界（41）。 □

推论 E.10（McKean 核示例：分层（分位数）转移）。考虑<款 id='7'>朴素提议分布 *FK* 模型，并取参数范围 $\varepsilon = O(1/T)$。回顾一下，我们附加一个终止步骤 $T{+}1$，其中 $M_{T+1} = \mathrm{Id}$，使目标为 η_{T+1}。令 $\bar{\eta}_{T+1}^N := \mathbb{E}[\eta_{T+1}^N]$ 表示 *SMC* 输出分布。

将 $Q_{p,T+1}(1)$ 的代理量定义为任意 $p \in [T]$ 对应的一步比率：

$$\widetilde{Q}_{p,T+1}(1)(s_{1:p}) \;:=\; \frac{\hat{V}(s_0, s_{1:p})}{\hat{V}(s_0, s_{1:p-1})},$$

并令 $\nu_p^N := \Phi_p(\eta_{p-1}^N)$。对于每个 p，根据 $\widetilde{Q}_{p,T+1}(1)$ 的 ν_p^N 分位数，将 E_p 划分为 $(E_{p,i})_{i=1}^N$，使得 $\nu_p^N(E_{p,i}) = 1/N$，且 $E_{p,i}$ 按 $Q_{p,T+1}(1)$ 递增排序。定义索引分层转移

$$K^{\mathrm{q}}_{p,\eta_{p-1}^N}(X_{p-1}^i, dy) := \nu_p^N(dy \mid \mathsf{E}_{p,i}), \qquad i = 1, \dots, N.$$

那么，$\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \le \delta_{\mathrm{TV}}$ 的一个充分条件是

$$N \;\gtrsim\; \max\left\{\left(\frac{L^6 T}{\delta_{\mathrm{TV}}}\right)^{2/3},\; \frac{\varepsilon L^6 T^2}{\delta_{\mathrm{TV}}}\right\}.$$

推论 *E.10* 的证明。令 $n := T + 1$ 为 FK 时域（其中 $M_{T+1} = \mathrm{Id}$）。我们沿用命题 E.9 中使用的核敏感偏差分解，但注意到 K^{q} 在粒子之间是 分层的，而非独立同分布的。关键在于，分层仍然保持相同的 条件无偏性 （McKean 一致性），因此线性项与独立同分布情形中一样严格为零；剩余贡献则由各层 局部条件方差 的平均值控制。

步骤 1（FK 常数 在朴素提议分布下，时域 $n = T + 1$）。根据引理 E.2，对于朴素提议分布和 $p \le T$，

$$q_{p,n} \le L^2(1+\varepsilon)^{2(n-p-1)} = L^2(1+\varepsilon)^{2(T-p)} \qquad \text{and} \qquad \beta(D\Phi_{p,n}) \le 2L^2(1+\varepsilon)^{2(T-p)}.$$

因此

$$q_{p,n}\beta(D\Phi_{p,n}) \le 2L^4(1+\varepsilon)^{4(T-p)} = O(L^4) \quad \text{under } \varepsilon = O(1/T),$$

从而 $\sum_{p=1}^T q_{p,\,n}\beta(D\Phi_{p,\,n}) = O(L^4 T)$。

步骤 2（分层更新：条件无偏性与定义良好的方差泛函）。固定 p 并写作 $\nu := \nu_p^N = \Phi_p(\eta_{p-1}^N)$ 和 $\mathsf{E}_i := \mathsf{E}_{p,\,i}$。在分层转移下，给定 F_{p-1}^N，新粒子相互独立，且

$$X_p^i \sim \nu(\cdot \mid \mathsf{E}_i), \qquad i = 1, \dots, N.$$

由于 $\nu(\mathsf{E}_i) = 1/N$，对于任意有界可测的 h，

$$\frac{1}{N}\sum_{i=1}^N \nu(\cdot \mid \mathsf{E}_i) = \nu(\cdot),$$

因此，分层更新是条件无偏的：

$$\mathbb{E}[\eta_p^N(h) \mid \mathcal{F}_{p-1}^N] = \frac{1}{N}\sum_{i=1}^N \nu(h \mid \mathsf{E}_i) = \nu(h).$$

因此，命题 E.9 中使用的相同一阶（线性）项抵消仍然成立。

不同之处在于二阶矩。定义分层方差泛函

$$\mathcal{V}^{\mathrm{q}}_{p,\nu}(h) \;:=\; \frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}_{\nu(\cdot|\mathsf{E}_i)}(h). \tag{42}$$

然后，利用给定 F^N_{p-1} 时 i 之间的独立性，

$$\mathbb{E}\Big[(\eta^N_p(h)-\nu(h))^2 \mid \mathcal{F}^N_{p-1}\Big] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}_{\nu(\cdot|\mathsf{E}_i)}(h) = \frac{1}{N}\cdot\frac{1}{N}\mathcal{V}^{\mathrm{q}}_{p,\nu}(h).$$

此外，对于每个分层，$\mathrm{Var}_{\nu(\cdot|\mathsf{E}_i)}(h) \le \frac{1}{4}$振幅$_{\mathsf{E}_i}(h)^2$,，因此

$$\sqrt{\mathcal{V}^{\mathrm{q}}_{p,\nu}(h)} \le \frac{1}{2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{osc}_{\mathsf{E}_i}(h)^2} \;\le\; \frac{1}{2\sqrt{N}}\sum_{i=1}^{N}\mathrm{osc}_{\mathsf{E}_i}(h), \tag{43}$$

其中最后一步用到了 $\sqrt{\sum a_i^2} \le \sum |a_i|$。

~**步骤 3（归一化以及使用 $Q_{p,n}(1)$-分位数所导致的不匹配）**。回想一下 $G_{p,n,\nu}(x) = Q_{p,n}(1)(x)/\nu(Q_{p,n}(1))$，并且分位数在归一化下保持不变。振荡按如下方式缩放

$$\mathrm{osc}_{\mathsf{E}_i}(G_{p,n,\nu}) = \frac{1}{\nu(Q_{p,n}(1))}\,\mathrm{osc}_{\mathsf{E}_i}(Q_{p,n}(1)).$$

在朴素提议分布下，引理 E.2意味着在 $\varepsilon = O(1/T)$ 的范围内，$1/\nu(Q_{p,n}(1)) = O(L)$（吸收 $(1+\varepsilon)^{O(T)}$）。
接下来，如引理 E.2 的证明所示，将单步贝尔曼误差控制沿剩余时域串联起来，即可得到逐点乘法近似

$$Q_{p,n}(1)(x) = \widetilde{Q}_{p,n}(1)(x)\cdot(1\pm O(\varepsilon T)), \qquad \forall x\in E_p.$$

令 $\alpha := O(\varepsilon T)$。由此可得每个分层上的确定性振荡界：

$$\mathrm{osc}_{\mathsf{E}_i}(Q_{p,n}(1)) \le (1+\alpha)\,\mathrm{osc}_{\mathsf{E}_i}(\widetilde{Q}_{p,n}(1)) + O(\alpha)\,\|\widetilde{Q}_{p,n}(1)\|_\infty.$$

~对 i 求和，并利用按 Q_p 排序的层. $_n(1)$,

$$\sum_{i=1}^{N}\mathrm{osc}_{\mathsf{E}_i}(\widetilde{Q}_{p,n}(1)) \le \mathrm{osc}(\widetilde{Q}_{p,n}(1)) = O(L), \qquad \|\widetilde{Q}_{p,n}(1)\|_\infty = O(L),$$

根据假设3.1。因此

$$\sum_{i=1}^{N}\mathrm{osc}_{\mathsf{E}_i}(Q_{p,n}(1)) = O(L) + O(\alpha)\,O(L)\,N,$$

因此，利用 $1/\nu(Q_{p,n}(1)) = O(L)$,

$$\sum_{i=1}^{N}\mathrm{osc}_{\mathsf{E}_i}(G_{p,n,\nu}) = O(L^2) + O(\alpha)\,O(L^2)\,N.$$

将 $h = G_{p,n,\nu}$ 代入（43）得

$$\sqrt{\mathcal{V}^{\mathrm{q}}_{p,\nu}(G_{p,n,\nu})} \;\le\; O\!\left(\frac{L^2}{\sqrt{N}}\right) + O(\alpha)\,O(L^2).$$

步骤 4（合并）。 将上述展示式代入核敏感偏差界（其前因子为 $1/N$），并利用步骤 1 得到

$$\|\bar{\eta}^N_{T+1} - \eta_{T+1}\|_{\mathrm{TV}} \le O\!\left(\sum_{p=1}^{T}\frac{L^4}{N}\left(\frac{L^2}{\sqrt{N}} + \alpha L^2\right)\right) = O\!\left(\frac{L^6 T}{N^{3/2}}\right) + O\!\left(\frac{L^6\alpha T}{N}\right).$$

由于 $\alpha = O(\varepsilon T)$，且我们处于 $\varepsilon = O(1/T)$ 的情形，因此所述充分条件成立。 □

备注 E.11（从独立同分布 McKean 转移到分层重采样）。 命题 E.9 针对一个（可能依赖于 η 的）马尔可夫核 $K_{p,\ \eta}$ 该核在时刻 p 对所有粒子均以相同方式应用，因此在给定 $\mathcal{F}_{p-1}^{N}$ 的条件下，后代 $\{X_p^i\}_{i=1}^N$ 独立同分布地来自同一分布 $\Phi_p(\eta_{p-1}^N) = \eta_{p-1}^N K_{p,\ \eta_{p-1}^N}$。在推论 E.10中，“分位数/分层”更新则改为采样

$$X_p^i \sim \nu_p^N(\cdot \mid \mathsf{E}_{p,i}), \qquad i = 1, \dots, N,$$

其中 $\{\mathsf{E}_{p,\ i}\}_{i=1}^N$ 是一个可测划分，满足 $\nu_p^N(\mathsf{E}_{p,\ i}) = 1/N$。此更新不能表示为对所有粒子一致地应用单个核；然而，它仍满足相同的 McKean 一致性（条件无偏性）：对于任意有界可测h,

$$\mathbb{E}\left[\eta_p^N(h) \mid \mathcal{F}_{p-1}^N\right] = \frac{1}{N}\sum_{i=1}^{N} \nu_p^N(h \mid \mathsf{E}_{p,i}) = \nu_p^N(h) = \Phi_p(\eta_{p-1}^N)(h).$$

因此，命题E.9所依据的望远镜求和／一阶分解中的一阶（线性）项，在分层重采样中同样原样抵消。唯一的修改在于二阶矩控制：独立同分布情形下的条件方差代理量 $\eta_{p-1}^N K_{p,\ \eta_{p-1}^N}((h - K_{p,\ \eta_{p-1}^N} h)^2)$ 应替换为分层方差泛函

$$\mathcal{V}_{p,\nu}^{\mathrm{q}}(h) := \frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}_{\nu(\cdot \mid \mathsf{E}_{p,i})}(h), \qquad \nu := \nu_p^N,$$

以及确定性振荡界 $\sqrt{\mathcal{V}_{p,\ \nu}^{\mathrm{q}}(h)} \le \frac{1}{2\sqrt{N}} \sum_{i=1}^{N}$ 振幅$_{\mathsf{E}_{p,\ i}}(h)$。作此替换后，命题E.9 的其余步骤均可照常进行，从而得到推论E.10。

分层重采样已在粒子滤波文献中得到广泛研究；例如，参见 [28]。本文考虑的更新可以视为一种特殊的分层规则，该规则 适用于指定函数的分位数（在我们的情形中，即代理前瞻剩余价值），使每个分层承载相等的质量，同时将分辨率集中于该函数变化最大的区域。

F 节中的证明6

引理 F.1（均匀全变差收缩 （独立 Metropolis–Hastings）。 设 (E,E) 为一个可测空间。设 π_{ref} 为 E 上的概率测度，并设 G ， $R:\ \mathsf{E}\to(0,\infty)$ 为可测的，且

$$0 < Z_G := \pi_{\mathrm{ref}}(G) < \infty, \qquad 0 < Z_R := \pi_{\mathrm{ref}}(R) < \infty.$$

定义目标分布和提议分布

$$\tilde{\pi}(dx) := \frac{G(x)}{Z_G}\,\pi_{\mathrm{ref}}(dx), \qquad q(dx) := \frac{R(x)}{Z_R}\,\pi_{\mathrm{ref}}(dx).$$

考虑以 $\tilde{\pi}$ 为目标分布、以 q 为提议分布的独立 *Metropolis–Hastings* 核 K， 即

$$K(x,dy) = \alpha(x,y)\,q(dy) + \Big(1 - \int \alpha(x,z)\,q(dz)\Big)\delta_x(dy), \qquad \alpha(x,y) := 1 \wedge \frac{\tilde{\pi}(dy)\,q(dx)}{\tilde{\pi}(dx)\,q(dy)} = 1 \wedge \frac{G(y)R(x)}{G(x)R(y)}.$$

假设存在 $b \ge 1$ 使得

$$\operatorname*{ess\,sup}_{x\in\mathsf{E}} \max\left\{\frac{G(x)}{R(x)}, \frac{R(x)}{G(x)}\right\} \ \le\ b.$$

那么：

(i)（小化）对于每个 $x \in \mathsf{E}$,

$$K(x,\cdot) \ \ge\ \frac{1}{b^2}\,\tilde{\pi}(\cdot).$$

(ii)（一致总变差收缩）对于任意概率测度 μ、ν （定义在(E，$\mathcal{E}$) 上）以及任意整数 $m \ge 1$,

$$\|\mu K^m - \nu K^m\|_{\mathrm{TV}} \ \le\ \Big(1 - \frac{1}{b^2}\Big)^m \|\mu - \nu\|_{\mathrm{TV}}.$$

特别地，*Dobrushin* 系数满足 $\beta(K) \le 1 - \frac{1}{b^2}$。 —

证明。 写出 $\tilde{\pi}$ 关于 q 的 Radon–Nikodym 导数：

$$w(x) := \frac{d\tilde{\pi}}{dq}(x) = \frac{G(x)/Z_G}{R(x)/Z_R} = \frac{Z_R}{Z_G}\cdot\frac{G(x)}{R(x)}.$$

根据假设， $\frac{G}{R} \le b$ q-几乎处处成立。此外，由于 $\frac{G}{R} \in [1/b,\ b]$ π_{ref}-几乎处处成立，我们有 $Z_G = \pi_{\mathrm{ref}}(G)$ — $\in [\frac{1}{b}Z_R, bZ_R]$，因此 $\frac{Z_R}{Z_G} \le b$。所以 $w(x) \le b^2$ q-几乎处处成立。记 $M :=$ 本质上确界$_q\, w \le b^2$。

对于独立 Metropolis–Hastings， $\alpha(x,y) = \underline{1 \wedge \frac{w}{(y)\,w(x)}}$。利用 $w(x) \le M$ 以及 $1 \wedge u \ge u/M$ （对于 $u \ge 0$），可得

$$\alpha(x,y) = 1 \wedge \frac{w(y)}{w(x)} \ \ge\ \frac{w(y)}{M}.$$

因此，对于任意可测的 A,

$$K(x,A) \ \ge\ \int_A \alpha(x,y)\,q(dy) \ \ge\ \frac{1}{M}\int_A w(y)\,q(dy) \ =\ \frac{1}{M}\tilde{\pi}(A) \ \ge\ \frac{1}{b^2}\tilde{\pi}(A),$$

这就证明了 (i)。

对于 (ii)， (i) 蕴含通常的 Doeblin 分解：对于某个马尔可夫核 K'， $K(x,\cdot) = \frac{1}{b^2}\tilde{\pi}(\cdot) + (1 - \frac{1}{b^2})K'(x,\cdot)$。因此，对于任意满足 $\Delta(\mathsf{E}) = 0$ 的符号测度 Δ,

$$\|\Delta K\|_{\mathrm{TV}} \le \Big(1 - \frac{1}{b^2}\Big)\|\Delta\|_{\mathrm{TV}},$$

并不断迭代即可得到（ii）。 □

定理*6.1*的证明。我们 考虑重要性重采样的路径空间，其中对于每个$h \in [H]$，随机变量 e

$$Y_h := ((S_1^{(h,1)}, S_1^{(h,2)}, \cdots, S_1^{(h,M)}, A_1^h), (S_2^{(h,1)}, S_2^{(h,2)}, \cdots, S_2^{(h,M)}, A_2^h), \cdots, (S_T^{(h,1)}, S_T^{(h,2)}, \cdots, S_T^{(h,M)}, A_T^h))$$

表示第h个 Metropolis-Hastings 提议，其中 $A_t^h \in [M]$ 是重采样状态在 M 个候选项中的索引。我们还记 $S_{1\,t}^{[h]} := (S^{(h,A_1^h)}{}_1, S^{(h,A_2^h)}{}_2, \cdots, S^{(h,A_t^h)}{}_t)$。然后计算接受率 $\alpha(X_{h-1}, Y_h)$，并将 $U_h \sim$ 均匀分布$(0,1)$与接受率进行比较，以决定是否接受 Y_h。若接受，则 $X_h = Y_h$；否则 $X_h = X_{h-1}$。

我们使用 $\bar{Z}_t^{[h]} := \frac{1}{M}\sum_{j=1}^{M} \hat{V}(s_0, (S_1^{[h]}{}_{:t-1}, S^{(h,j)}{}_t))$ 作为给定前缀 $S_1^{[h]}{}_{:t}$ 时 M 个候选项奖励的平均值。

要应用 Metropolis-Hastings 算法，通常需要计算生成整个 Y_h as

$$\mathbb{P}(Y_h) = \prod_{t=1}^{T}\left(\frac{\hat{V}(s_0, (S_{1:t-1}^{[h]}, S_t^{(h,A_t^h)}))}{M \cdot \bar{Z}_t^{[h]}} \prod_{j=1}^{M} \pi_{\text{ref}}(S_t^{(h,j)} | S_{1:t-1}^{[h]})\right).$$

相应地，目标分布选为

$$\mathbb{P}(Y_h') \propto \hat{V}(S_0, S_{1:T}^{[h]}) \prod_{t=1}^{T}\left(\frac{1}{M}\prod_{j=1}^{M} \pi_{\text{ref}}(S_t^{(h,j)} | S_{1:t-1}^{[h]})\right)$$

其中，$S_1^{[h]}{}_{:T}$的边缘分布（或可视为随机变量 $Y_h \mapsto S_1^{[h]}{}_{:T}$ 的确定性变换）为$\mathbb{P}(S_1^{[h]}{}_{:T}) \propto \pi_{\text{ref}}(S_1^{[h]}{}_{:T} | s_0) \cdot \hat{V}(S_0, S_1^{[h]}{}_{:T})$，这就是我们所需的 $\tilde{\pi}_T$。

现在，我们将每个G 处的良好集合定义为$h \in [H]$，即

$$Y_h \in \mathsf{G} \iff \forall t \in [T], \quad \left|\frac{\bar{Z}_t^{[h]}}{\mathbb{E}_{s_t \sim \pi_{\text{ref}}(\cdot | s_{1:t-1}^{[h]})}[\hat{V}(s_0, (s_{1:t-1}^{[h]}, s_t))]} - 1\right| \le \xi.$$

并将良好事件 E 定义为

$$\mathcal{E} := \bigcap_{h=2}^{H+1} \{Y_h \in \mathsf{G}\} \cap \{X_1 \in \mathsf{G}\}.$$

现在我们声称，在以良好事件 E 为条件时，$X_1, X_2, \cdots, X_{H+1}$ 仍构成马尔可夫链。为证明这一点，首先定义滤过

$$\mathcal{F}_h := \sigma(X_1, (Y_2.U_2), (Y_3, U_3), \cdots, (Y_h, U_h)).$$

由于每个 Y_h 都服从我们记为 $Q(\cdot)$ 的独立同分布，因此我们考虑 Y_h 在良好集合 G 内的条件分布，记作 $Q(\cdot|\mathsf{G})$。容易看出，在良好事件 E 条件下，$\{Y_h\}_{h=1}^{H+1}$仍然服从 Q_{G} 所指定的独立同分布：$= Q(\cdot|\mathsf{G})$，并且与 $U_1, U_2, \cdots$ 相互独立。我们还将目标分布记作 $\mathbb{P}(Y_{H+1}') \sim \Pi(\cdot)$，其条件版本为 $\Pi_{\mathsf{G}} := \Pi(\cdot|\mathsf{G})$。我们定义条件 MH 核

$$K_{\mathsf{G}}(x, dy) := Q_{\mathsf{G}}(dy)\alpha(x, y) + \left(1 - \int Q_{\mathsf{G}}(du)\alpha(x, u)\right)\delta_x(dy).$$

我们的目标是证明，对于任意有界可测的 f，

$$\mathbb{E}[f(X_{h+1}) | \mathcal{E}, \mathcal{F}_h] = K_{\mathsf{G}} f(X_h) \quad \text{almost surely on } \mathcal{E}.$$

回顾一下（根据 MH 更新的定义）

$$X_{h+1} = \begin{cases} Y_{h+1}, & U_{h+1} \le \alpha(X_h, Y_{h+1}), \\ X_h, & U_{h+1} > \alpha(X_h, Y_{h+1}). \end{cases}$$

等价地，

$$f(X_{h+1}) = f(Y_{h+1})\,\mathbb{I}\{U_{h+1} \le \alpha(X_h, Y_{h+1})\} + f(X_h)\,\mathbb{I}\{U_{h+1} > \alpha(X_h, Y_{h+1})\}.$$

现在我们计算给定 $\sigma(\mathsf{F}_h \cup \{\mathsf{E}\})$ 时的条件期望。在 E 上，我们有 $Y_{h+1} \in \mathsf{G}$，并且（根据 E 的构造）第 $h+1$ 步的提议随机性与 F_h 相互独立，且其条件分布为 Q_{G}；此外，$U_{h+1} \sim$ 均匀分布于 $\mathsf{Unif}(0,1)$，并且与 (F_h, Y_{h+1}) 相互独立（仍是在 E 条件下）。另请注意，X_h 是 F_h-可测的。

因此，在 E 上，

$$\begin{aligned}
\mathbb{E}[f(Y_{h+1})\,\mathbb{I}\{U_{h+1} \le \alpha(X_h, Y_{h+1})\} \mid \mathcal{E}, \mathcal{F}_h] &= \mathbb{E}[\mathbb{E}[f(Y_{h+1})\,\mathbb{I}\{U_{h+1} \le \alpha(X_h, Y_{h+1})\} \mid \mathcal{E}, \mathcal{F}_h, Y_{h+1}] \mid \mathcal{E}, \mathcal{F}_h] \\
&= \mathbb{E}[f(Y_{h+1})\,\mathbb{P}(U_{h+1} \le \alpha(X_h, Y_{h+1}) | \mathcal{E}, \mathcal{F}_h, Y_{h+1}) \mid \mathcal{E}, \mathcal{F}_h] \\
&= \mathbb{E}[f(Y_{h+1})\,\alpha(X_h, Y_{h+1}) \mid \mathcal{E}, \mathcal{F}_h] \\
&= \int Q_{\mathsf{G}}(dy)\, f(y)\, \alpha(X_h, y).
\end{aligned}$$

类似地，在 E 上，

$$\begin{aligned}\mathbb{E}[\mathbb{I}\{U_{h+1} > \alpha(X_h, Y_{h+1})\} \mid \mathcal{E}, \mathcal{F}_h] &= 1 - \mathbb{E}[\mathbb{I}\{U_{h+1} \le \alpha(X_h, Y_{h+1})\} \mid \mathcal{E}, \mathcal{F}_h] \\ &= 1 - \int Q_{\mathsf{G}}(dy)\, \alpha(X_h, y).\end{aligned}$$

由于 $f(X_h)$ 是 F_h-可测的，结合上述两个式子可得，在 E 上，

$$\begin{aligned}\mathbb{E}[f(X_{h+1}) \mid \mathcal{E}, \mathcal{F}_h] &= \int Q_{\mathsf{G}}(dy)\, f(y)\, \alpha(X_h, y) + f(X_h)\Big(1 - \int Q_{\mathsf{G}}(dy)\, \alpha(X_h, y)\Big) \\ &= K_{\mathsf{G}} f(X_h).\end{aligned}$$

这证明了

$$\mathbb{E}[f(X_{h+1}) \mid \mathcal{E}, \mathcal{F}_h] = K_{\mathsf{G}} f(X_h) \qquad \text{a.s. on } \mathcal{E},$$

如所需。

因此，条件分布满足

$$\mathrm{Law}(X_{H+1})|_{\mathcal{E}} = \mathrm{Law}(X_1)|_{\mathcal{E}} K_{\mathsf{G}}^{H}.$$

此外，我们证明 Π_{G} 是 K_{G} 的平稳分布。回顾一下，$Q_{\mathsf{G}} := Q(\cdot \mid \mathsf{G})$ 和 $\Pi_{\mathsf{G}} := \Pi(\cdot \mid \mathsf{G})$ 的支撑集均包含于 G。由于 $\Pi \ll Q$ 位于路径空间上，我们也有 $\Pi_{\mathsf{G}} \ll Q_{\mathsf{G}}$ 位于 G 上。定义 Radon–Nikodym 导数

$$r(x) := \frac{d\Pi_{\mathsf{G}}}{dQ_{\mathsf{G}}}(x), \qquad x \in \mathsf{G}.$$

注意，r 等于 $w := d\Pi/dQ$，在G上至多相差一个正常数，因此

$$\alpha(x, y) = 1 \wedge \frac{w(y)}{w(x)} = 1 \wedge \frac{r(y)}{r(x)}, \qquad x, y \in \mathsf{G}.$$

我们首先验证非对角部分的细致平衡。对任意 $x,\ y \in \mathsf{G}$,

$$\begin{aligned}\Pi_{\mathsf{G}}(dx)\, Q_{\mathsf{G}}(dy)\, \alpha(x, y) &= r(x)\, Q_{\mathsf{G}}(dx)\, Q_{\mathsf{G}}(dy)\, \Big(1 \wedge \frac{r(y)}{r(x)}\Big) \\ &= Q_{\mathsf{G}}(dx)\, Q_{\mathsf{G}}(dy)\, \min\{r(x), r(y)\} \\ &= r(y)\, Q_{\mathsf{G}}(dy)\, Q_{\mathsf{G}}(dx)\, \Big(1 \wedge \frac{r(x)}{r(y)}\Big) \\ &= \Pi_{\mathsf{G}}(dy)\, Q_{\mathsf{G}}(dx)\, \alpha(y, x).\end{aligned}$$

因此，对任意可测的 $B \subseteq \mathsf{G}$,

$$\begin{aligned}(\Pi_{\mathsf{G}} K_{\mathsf{G}})(B) &= \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dx) \int_{\mathsf{G}} Q_{\mathsf{G}}(dy)\, \alpha(x, y)\, \mathbb{I}_B(y) + \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dx) \Big(1 - \int_{\mathsf{G}} Q_{\mathsf{G}}(du)\, \alpha(x, u)\Big) \mathbb{I}_B(x) \\ &=: I_1 + I_2.\end{aligned}$$

我们利用上述细致平衡恒等式和Fubini-Tonelli定理来化简 I_1：

$$\begin{aligned}I_1 &= \int_{\mathsf{G}} \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dx)\, Q_{\mathsf{G}}(dy)\, \alpha(x, y)\, \mathbb{I}_B(y) = \int_{\mathsf{G}} \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dy)\, Q_{\mathsf{G}}(dx)\, \alpha(y, x)\, \mathbb{I}_B(y) \\ &= \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dy)\, \mathbb{I}_B(y) \int_{\mathsf{G}} Q_{\mathsf{G}}(dx)\, \alpha(y, x) = \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dx)\, \mathbb{I}_B(x) \int_{\mathsf{G}} Q_{\mathsf{G}}(du)\, \alpha(x, u).\end{aligned}$$

将其代入 I_2 可得

$$\begin{aligned}(\Pi_{\mathsf{G}} K_{\mathsf{G}})(B) &= \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dx)\, \mathbb{I}_B(x) \int_{\mathsf{G}} Q_{\mathsf{G}}(du)\, \alpha(x, u) + \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dx) \Big(1 - \int_{\mathsf{G}} Q_{\mathsf{G}}(du)\, \alpha(x, u)\Big) \mathbb{I}_B(x) \\ &= \int_{\mathsf{G}} \Pi_{\mathsf{G}}(dx)\, \mathbb{I}_B(x) = \Pi_{\mathsf{G}}(B).\end{aligned}$$

因此 $\Pi_{\mathsf{G}} K_{\mathsf{G}} = \Pi_{\mathsf{G}}$，即 Π_{G} 是 K_{G}的平稳分布。

因此，标准的Dobrushin压缩陈述（引理F.1）给出

$$\|\mathrm{Law}(X_{H+1})|_{\mathcal{E}} - \Pi_{\mathsf{G}}\|_{\mathrm{TV}} \le (1 - b^{-2})^{H} \|\mathrm{Law}(X_1)|_{\mathcal{E}} - \Pi_{\mathsf{G}}\|_{\mathrm{TV}},$$

其中 $b := \left((1+\varepsilon)\frac{1+\xi}{1-\xi}\right)^{T-1}$。因此

$$\begin{aligned}\|\mathrm{Law}(X_{H+1})|_{\mathcal{E}} - \Pi\|_{\mathrm{TV}} &\le \|\mathrm{Law}(X_{H+1})|_{\mathcal{E}} - \Pi_{\mathsf{G}}\|_{\mathrm{TV}} + \|\Pi - \Pi_{\mathsf{G}}\|_{\mathrm{TV}} \\ &\le (1-b^{-2})^{H}\|\mathrm{Law}(X_1)|_{\mathcal{E}} - \Pi_{\mathsf{G}}\|_{\mathrm{TV}} + \|\Pi - \Pi_{\mathsf{G}}\|_{\mathrm{TV}} \\ &\le 2(1-b^{-2})^{H} + 2\mathbb{P}(Y \notin \mathsf{G}).\end{aligned}$$

注意，$S_{1\,T}^{[H+1]}$:是通过确定性变换（边缘化）$g\colon X_{H+1} \mapsto S_{1\,T}^{[H+1]}$:得到的，并且 Π 的边缘化版本恰好是 $\tilde{\pi}_T$，利用数据处理不等式可得

$$\|\mathrm{Law}(S_{1:T}^{[H+1]})|_{\mathcal{E}} - \tilde{\pi}_T\|_{\mathrm{TV}} \le \|\mathrm{Law}(X_{H+1})|_{\mathcal{E}} - \Pi\|_{\mathrm{TV}}.$$

因此，在 $\varepsilon = O\ 1\ T$ $\xi = O\ 1\ T$ $b = O\ 1$ $1 - b^{-2\ H} \le \delta_{\mathrm{TV}}\ 4$ $\mathbb{P}\ Y \notin$ $\le$(/) 与 (/) 的情形下，我们有 $\delta_{\mathrm{TV}}\ 4$ $H = O$ $\delta_{\mathrm{TV}}^{-1}$ $M = O\ C_{\mathrm{act}}T^2$ δ^{-1} C_{act} 5 $\mathbb{P}$ $\ge 1 - HT\delta$ ()。因此，令 ( ) / 与 (G) /，可得 (log( )) 与 (log( )) （其中定义见引理 B. ）。由于 (E)，将 δ 替换为 $\delta/(HT)$，并利用 $C_{\mathrm{act}} \le L(1+\varepsilon)$，可得总时间复杂度为 $MTH = \tilde{O}(LT^3 \log(\delta^{-1}) \log(\delta_{\mathrm{TV}}^{-1}))$。□

G 基于拒绝的 SP-gSMC 与 Metropolis-Hastings 校正

拒绝采样 MH。 作为第 6 节中重采样池 MH 的替代方案，可以使用拒绝采样从所需的局部倾斜中获得（条件）精确样本。然而，外层 MH 步骤的接受概率现在依赖于估计的归一化常数，由此产生的估计误差以根本不同的方式进入分析。正如我们所见，虽然内部采样步骤在良好事件上是精确的，但总体复杂度在 δ_{TV} 中变得更差。

引理 *G.1*（算法 *8* 的精确 *MH* 收缩）。考虑理想化设置下的算法 *8*，其中接受率中出现的所有条件期望均被精确计算（即不存在归一化因子估计误差）。假设假设 *3.2* 在参数 ε 下成立。令 K 为作用于 $S_{1:T}$ 上、用于外部 *MH* 循环的独立 *Metropolis–Hastings* 核，其目标分布为 $\tilde{\pi}_T(\,ds_{1:T} \mid s_0) \propto \pi_{\mathrm{ref}}(ds_{1:T} \mid s_0)\,\hat{V}(s_0, s_{1:T})$，提议分布则为内循环单次运行所产生的（朴素引导）轨迹分布。那么，对于 $S_{1:T}$ 上的任意概率测度 μ, ν 以及任意整数 $m \geq 1$，

$$\|\mu K^m - \nu K^m\|_{\mathrm{TV}} \;\leq\; c^m\,\|\mu - \nu\|_{\mathrm{TV}}, \qquad c \;\leq\; 1 - \frac{1}{(1+\varepsilon)^{2T}}.$$

证明。令 $\pi_{\mathrm{ref}}(\cdot \mid s_0)$ 为由 $\pi_{\mathrm{ref}}(\cdot \mid s_0, s_{1:t-1})$ 在 $\mathsf{S}_{1:T}$ 上诱导的基路径测度。对于 $t = 1, \ldots, T$，定义精确的单步归一化常数

$$Z_t(s_0, s_{1:t-1}) := \mathbb{E}_{s_t \sim \pi_{\mathrm{ref}}(\cdot \mid s_0, s_{1:t-1})}\big[\hat{V}(s_0, (s_{1:t-1}, s_t))\big], \qquad s_{1:0} := \varnothing.$$

提议分布 内循环生成的完整轨迹上的提议分布与 $\pi_{\mathrm{ref}}(s_{1:T} \mid s_0)\; R(s_{1:T})$ 成正比，其中 th

$$R(s_{1:T}) := \prod_{t=1}^{T} \frac{\hat{V}(s_0, s_{1:t})}{Z_t(s_0, s_{1:t-1})},$$

而目标分布与 $\pi_{\mathrm{ref}}(s_{1:T} \mid s_0)\; G(s_{1:T})$ 成正比，其中

$$G(s_{1:T}) := \frac{\hat{V}(s_0, s_{1:T})}{Z_1(s_0)}.$$

因此

$$\frac{R(s_{1:T})}{G(s_{1:T})} = \prod_{t=1}^{T-1} \frac{\hat{V}(s_0, s_{1:t})}{Z_{t+1}(s_0, s_{1:t})}, \qquad \frac{G(s_{1:T})}{R(s_{1:T})} = \prod_{t=1}^{T-1} \frac{Z_{t+1}(s_0, s_{1:t})}{\hat{V}(s_0, s_{1:t})}.$$

根据假设 3.2，对于每个前缀 $s_{1:t}$，

$$\max\left\{ \frac{\hat{V}(s_0, s_{1:t})}{Z_{t+1}(s_0, s_{1:t})},\; \frac{Z_{t+1}(s_0, s_{1:t})}{\hat{V}(s_0, s_{1:t})} \right\} \leq 1 + \varepsilon,$$

因此在 $\mathsf{S}_{1:T}$ 上逐点成立，

$$\max\left\{ \tfrac{R}{G}, \tfrac{G}{R} \right\} \leq (1+\varepsilon)^T.$$

应用 引理 F.1 结合 $b = (1+\varepsilon)^T$ 可得全变差收缩，其收缩因子为 $c \leq 1 - 1/b^2 = 1 - (1+\varepsilon)^{-2T}$。 □

定理 G.2 (采用 Metropolis-Hastings 的基于拒绝的 SP-gSMC 的时间复杂度)。 在假设 *3.1* 和 *3.2* 成立的条件下，并考虑贝尔曼误差 $\varepsilon = O(1/T)$的情形。则对于任意 $0 < \delta \lesssim \delta_{\mathrm{TV}}$，使用 $\tilde{O}\left(\frac{LT^3 \log(\delta^{-1})}{\delta_{\mathrm{TV}}^2}\right)$ 的时间，存在一个良好事件 $\mathcal{E}$ 使得 $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$，在该条件下，算法 *8* 的采样分布全变差误差小于 δ_{TV}，即 $\|\tilde{\pi}_T - \hat{\pi}_T|_{\mathcal{E}}\|_{\mathrm{TV}} \leq \delta_{\mathrm{TV}}$。

定理 *G.2* 的证明。 在本证明中，记 $X_h \in \mathcal{S}_{1:T}$ 为 Metropolis-Hastings 第 h 步接受的提议。$Y_h \in \mathcal{S}_{1:T}$ 为拒绝/接受前的提议。在事件$\hat{\pi}_T|_{\mathcal{E}}$ 上的输出分布定义为 $\mathrm{Law}(X_{H+1})|_{\mathcal{E}}$。注意，与命题 D.4 的证明相同，拒绝采样提议所花费的时间相较于估计归一化常数所需的时间可忽略不计，因为前者只需估计到常数级相对误差，而后者（正如我们将在本证明中再次看到的）需要估计到 $O(\delta_{\mathrm{TV}}/T)$ 的相对误差。我们省略前者，复杂度阶不变。

固定 $f \in$ 振幅$_1$。定义精确的 MH 半群

$$f_h := K^{H-h+1} f, \qquad h = 1, \ldots, H+1,$$

因此 $f_1 = K^H f$ 且 $f_{H+1} = f$，并且 $f_h = K f_{h+1}$。那么振幅$(f_h) \leq 1$。

令 X_1 为第一个提议分布（默认接受）。对于每个 $h = 1, \ldots, H$，给定当前状态 X_h，抽取一个独立的提议分布Y_{h+1}，并计算带噪声的接受概率 $\hat{\alpha}_h = \hat{\alpha}_h(X_h, Y_{h+1})$；令 α_h 为真实的接受概率。然后抽取$U_h \sim$ 均匀分布$[0,1]$ 并设定

$$X_{h+1} = \begin{cases} Y_{h+1}, & U_h \leq \hat{\alpha}_h, \\ X_h, & U_h > \hat{\alpha}_h. \end{cases}$$

令 F_h 为 σ-代数，包含截至抽取 U_h 之前一刻已揭示的所有随机性（因此包含 X_h、Y_{h+1} 以及用于计算 $\hat{\alpha}_h$的所有蒙特卡罗随机性，但不包含 U_h 及任何未来的随机性）。

将全局良好事件写为 $\mathrm{E} = \mathrm{E}_{\le h} \cap \mathrm{E}_{>h}$，其中 $\mathrm{E}_{>h} := \cap_{j=h+1}^{H+1} \mathrm{E}_j$。考虑每个归一化因子均在 $1+\xi$ 相对准确度内被估计。则归一化因子的精度在 $\xi_{\mathrm{acc}} = \binom{1+\xi}{1-\xi}^{T} - 1$内。具体而言，引理 B.5 表明，当 $N_m \ge 4C_{\mathrm{act}} T^2 \log(T\delta^{-1})$ 时，事件 E_h:

$$= \left\{ |\xi_{\mathrm{acc}}| \le 16\sqrt{2T}\sqrt{\frac{C_{\mathrm{act}} \log(T\delta^{-1})}{N_m}} \right\}$$ 在以 F_{h-1} 为条件的情况下以大于 $1-\delta$ 的概率发生。因此 $\mathbb{P}(\mathrm{E}_h) \ge 1-\delta$。

我们有

$$\mathbb{E}\big[f_{h+1}(X_{h+1})\mathbb{I}_{\mathcal{E}}\big] = \mathbb{E}\big[f_{h+1}(X_{h+1})\big] - \mathbb{E}\big[f_{h+1}(X_{h+1})\mathbb{I}_{\mathcal{E}^c}\big].$$

对于每个$h = 1, \ldots\ldots, H$，定义辅助的“真接受”下一状态

$$\widetilde{X}_{h+1} = \begin{cases} Y_{h+1}, & U_h \le \alpha_h, \\ X_h, & U_h > \alpha_h. \end{cases}$$

在 F_h 的条件下（从而 U_h 是仅剩的随机性）。则

$$\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big] = \hat{\alpha}_h f_{h+1}(Y_{h+1}) + (1-\hat{\alpha}_h) f_{h+1}(X_h),$$

并且类似地

$$\mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big] = \alpha_h f_{h+1}(Y_{h+1}) + (1-\alpha_h) f_{h+1}(X_h),$$

and

$$\mathbb{E}[\mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big]] = \mathbb{E}[K f_{h+1}(X_h)] = \mathbb{E}[f_h(X_h)].$$

我们还有

$$\mathbb{E}[\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big]] = \mathbb{E}[f_{h+1}(X_{h+1})].$$

因此得到精确恒等式

$$\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big] - \mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big] = (\hat{\alpha}_h - \alpha_h)\big(f_{h+1}(Y_{h+1}) - f_{h+1}(X_h)\big). \tag{44}$$

我们只关注良好事件 $\mathrm{E}_{\le h}$ 下的差值，在该事件中 $|\hat{\alpha}_h - \alpha_h|$ 很小，因此我们也有

$$\begin{aligned} &\Big|\mathbb{E}\left[\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big] - \mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big]\right]\Big| \\ \le\ &\left|\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_h}\left(\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big] - \mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big]\right)\right]\right| \\ &\qquad + \left|\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_h^c}\left(\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big] - \mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big]\right)\right]\right| \\ \le\ &\sup_{\mathcal{E}_h} |(\hat{\alpha}_h - \alpha_h)|\big|\big(f_{h+1}(Y_{h+1}) - f_{h+1}(X_h)\big)\big|\mathbb{P}(\mathcal{E}_h) + \mathrm{osc}(f_{h+1})\mathbb{P}(\mathcal{E}_h^c). \end{aligned}$$

现在我们进行望远镜求和

$$\mathbb{E}\big[f(X_{H+1})\big] - \mathbb{E}\big[K^H(f)(X_1)\big] = \sum_{h=1}^{H} \mathbb{E}\left[\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big] - \mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big]\right].$$

Thus

$$\begin{aligned} &\mathbb{E}\big[f(X_{H+1})\mathbb{I}_{\mathcal{E}}\big]/\mathbb{P}(\mathcal{E}) - \mathbb{E}\big[K^H(f)(X_1)\big] \\ &= (\mathbb{E}\big[f(X_{H+1})\big] - \mathbb{E}\big[K^H(f)(X_1)\big])/\mathbb{P}(\mathcal{E}) - (\mathbb{E}\big[f(X_{H+1})\mathbb{I}_{\mathcal{E}^c}\big] - (1-\mathbb{P}(\mathcal{E}))\mathbb{E}\big[K^H(f)(X_1)\big])/\mathbb{P}(\mathcal{E}) \\ &= \frac{1}{\mathbb{P}(\mathcal{E})}\sum_{h=1}^{H} \mathbb{E}\left[\mathbb{E}_{U_h}\big[f_{h+1}(X_{h+1}) \mid \mathcal{F}_h\big] - \mathbb{E}_{U_h}\big[f_{h+1}(\widetilde{X}_{h+1}) \mid \mathcal{F}_h\big]\right] - \frac{\mathbb{P}(\mathcal{E}^c)}{\mathbb{P}(\mathcal{E})}(\mathbb{E}\big[f(X_{H+1})\mathbb{I}_{\mathcal{E}^c}\big]/\mathbb{P}(\mathcal{E}^c) - \mathbb{E}\big[K^H(f)(X_1)\big]). \end{aligned}$$

so

$$|\mathbb{E}\big[f(X_{H+1})\mathbb{I}_{\mathcal{E}}\big]/\mathbb{P}(\mathcal{E}) - \mathbb{E}\big[K^H(f)(X_1)\big]| \le \sum_{h=1}^{H}(2H\delta + \xi_{\mathrm{acc}})\mathrm{osc}(f_{h+1}) + 2H\delta\mathrm{osc}(f) \le (2H(H+1)\delta + H\xi_{\mathrm{acc}}).$$

其中我们利用了在良好事件 E_h 上 $|\hat{\alpha}_h - \alpha_h| \le \xi_{\mathrm{acc}}$ 这一事实。

因此我们有

$$\begin{aligned}\|\mathrm{Law}(X_{H+1})|_{\mathcal{E}} - \tilde{\pi}_T\|_{\mathrm{TV}} &\le \|\mathrm{Law}(X_{H+1})|_{\mathcal{E}} - \mathrm{Law}(X_1)K^H\|_{\mathrm{TV}} + \|\mathrm{Law}(X_1)K^H - \tilde{\pi}_T\|_{\mathrm{TV}} \\ &\le 2H(H+1)\delta + H\xi_{\mathrm{acc}} + c^H\|\mathrm{Law}(X_1) - \tilde{\pi}_T\|_{\mathrm{TV}}\end{aligned}$$

根据引理 G.1，精确的 MH 核满足一致全变差收缩，且 $c \le 1 - \frac{1}{(1+\varepsilon)^{2T}}$。

回顾定理 4.3 表明 $\|\mathrm{Law}(X_1) - \tilde{\pi}_T\|_{\mathrm{TV}} \le 2\varepsilon T$。我们关注 $\varepsilon = O(1/T)$ 情形，特别是当 $\varepsilon < \frac{1}{5T}$ 时，因此 $c^H \le (4\varepsilon T)^H$，并且

$$\|\mathrm{Law}(X_{H+1})|_{\mathcal{E}} - \tilde{\pi}_T\|_{\mathrm{TV}} \le 2H(H+1)\delta + H\xi_{\mathrm{acc}} + (4\varepsilon T)^H \cdot 2\varepsilon T.$$

取 $\frac{1}{2}\delta_{TV} = (4\varepsilon T)^H \cdot 2\varepsilon T$，可得 $H = \log(\frac{\delta_{\mathrm{TV}}}{4\varepsilon T})\log(4\varepsilon T) = O(\log(\delta_{\mathrm{TV}}^{-1}))$。接着取 $\delta \le \frac{\delta_{\mathrm{TV}}}{8H}(H+1) = O(\delta_{\mathrm{TV}})$，且 $\xi_{\mathrm{acc}} = \delta_{\mathrm{TV}}(4H)^{-1}$，即 $N_m = \tilde{O}(T^2 C_{\mathrm{act}}\log(\delta^{-1})\delta_{\mathrm{TV}}^2)$。代入 $\delta \leftarrow \frac{\delta}{H}$ 不会改变渐近速率。然后，在事件 E 内，时间复杂度为 $O(N_m T H) = \tilde{O}(T^3 L\log(\delta^{-1})\delta_{\mathrm{TV}}^2)$。

□

备注 G.3. **为何它比重采样池（第 6 节）更差。** 在这一构造中，MH 接受概率取决于通过蒙特卡洛得到的估计归一化常数。因此，估计误差不再仅仅影响（常数阶的）收缩系数；相反，它会在每个 MH 步骤中引入额外的加性 误差，并随着 MH 迭代次数的增加而 线性累积。因而，为了使最终的全变差误差低于 δ_{TV}，我们必须将相关归一化常数的相对误差控制在 $O(\delta_{\mathrm{TV}}/T)$，这会导致不可避免的 $\delta_{\mathrm{TV}}^{-2}$ 开销。

H 新假设：全局贝尔曼误差

假设 H.1（全局贝尔曼误差界）。 我们假设对于某个 ε_g，以下条件成立 ：

$$\max_{t\in[T]}\max_{s_{1:t}\in\mathcal{S}_{1:t}}\max\left\{\frac{\hat{V}(s_0,s_{1:t})}{\mathbb{E}_{s_{t+1:T}\sim\pi_{\text{ref}}(\cdot|s_0,s_{1:t})}[\hat{V}(s_0,s_{1:T})]},\frac{\mathbb{E}_{s_{t+1:T}\sim\pi_{\text{ref}}(\cdot|s_0,s_{1:t})}[\hat{V}(s_0,s_{1:T})]}{\hat{V}(s_0,s_{1:t})}\right\}\le 1+\varepsilon_g$$

在此假设下，我们得到一个新的序贯蒙特卡罗复杂度界，该界由全局贝尔曼误差刻画，而非先前的局部贝尔曼误差。我们还将在后文证明，该结果对任意提议分布 π_{P} 同样成立。

命题 H.2。 在与定理 *5.1*相同的条件下，将假设*3.2* 替换为假设 *H.1*，要使采样总变差误差低于 δ_{TV}，所需的粒子数为 $N \geq L^6T(1+\varepsilon_g{}^6)$。总时间复杂度为

$$O\Big(\frac{L^6(1+\varepsilon_g)^6T^2}{\delta_{\text{TV}}^2}\Big).$$

命题 *H.2* 的证明。 我们完全沿用定理5.1的证明策略，从局部总变差偏差界（定理 E.6）开始：

$$\|\bar{\eta}^N_{T+1}-\eta_{T+1}\|_{\text{TV}}\le\frac{1}{4N}\sum_{p=1}^{T+1}(q^2_{p,T+1}-1)\,\beta(D\Phi_{p,T+1}). \tag{45}$$

由引理 E.1，我们有 $\beta(D\Phi_{p,T+1})\le 2q_{p,T+1}\beta(P_{p,T+1})\le 2q_{p,T+1}$，因此

$$\|\bar{\eta}^N_{T+1}-\eta_{T+1}\|_{\text{TV}}\le\frac{1}{2N}\sum_{p=1}^{T+1}q_{p,T+1}\,(q^2_{p,T+1}-1). \tag{46}$$

最后一项 $p=T+1$ 消失了，因为 $q_{T+1,\ T+1}=1$。

步骤 1：在$q_{p,\ T+1}$满足全局贝尔曼误差的条件下的统一界。 我们考虑朴素提议分布，其中 $M_t(s_{1:t-1},\ ds_{1:t})=\pi_{\text{ref}}(s_t\mid s_0,\ s_{1:t-1})$，且 $G_t(s_{1:t})=\hat{}V(s_0,\ s_{1:t})/\hat{}V(s_0,\ s_{1:t-1})$。回顾 FK 半群的解释

$$Q_{p,T+1}(1)(x)=\mathbb{E}\left[\prod_{t=p}^{T}G_t(S_{1:t})\,\middle|\,S_{1:p}=x\right].$$

由于 G_t 是望远镜比率，乘积会消去所有交叉项：

$$\prod_{t=p}^{T}G_t(S_{1:t})=\frac{\hat{V}(s_0,S_{1:T})}{\hat{V}(s_0,S_{1:p-1})}.$$

因此，对于任意前缀/状态 $x=s_{1:p}\in E_p$，

$$Q_{p,T+1}(1)(x)=\frac{\mathbb{E}\big[\hat{V}(s_0,S_{1:T})\mid S_{1:p}=x\big]}{\hat{V}(s_0,s_{1:p-1})}. \tag{47}$$

现在在时间 $t=p$应用假设H.1：

$$\frac{1}{1+\varepsilon_g}\hat{V}(s_0,s_{1:p})\ \le\ \mathbb{E}\big[\hat{V}(s_0,S_{1:T})\mid S_{1:p}=s_{1:p}\big]\ \le\ (1+\varepsilon_g)\,\hat{V}(s_0,s_{1:p}).$$

代入（47）可得

$$\frac{1}{1+\varepsilon_g}\cdot\frac{\hat{V}(s_0,s_{1:p})}{\hat{V}(s_0,s_{1:p-1})}\ \le\ Q_{p,T+1}(1)(s_{1:p})\ \le\ (1+\varepsilon_g)\cdot\frac{\hat{V}(s_0,s_{1:p})}{\hat{V}(s_0,s_{1:p-1})}.$$

根据假设3.1，$\hat{}V(s_0,s_{1:p})/\hat{}V(s_0,s_{1:p-1})\in[1/L,\ L]$。因此对所有 $p\le T$以及所有 $x\in E_p$ ，

$$\frac{1}{L(1+\varepsilon_g)}\ \le\ Q_{p,T+1}(1)(x)\ \le\ L(1+\varepsilon_g).$$

因此，

$$q_{p,T+1}=\sup_{x,y\in E_p}\frac{Q_{p,T+1}(1)(x)}{Q_{p,T+1}(1)(y)}\ \le\ L^2(1+\varepsilon_g)^2,\qquad\forall p\le T. \tag{48}$$

步骤 2：代入局部总变差偏差界。 利用 $q(q^2-1) \le q^3$ 与 (48)，对所有 $p \le T$，我们得到：

$$q_{p,T+1}\,(q_{p,T+1}^2-1) \;\le\; q_{p,T+1}^3 \;\le\; L^6(1+\varepsilon_g)^6.$$

代入 (46) 得到

$$\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \le \frac{1}{2N}\sum_{p=1}^{T} L^6(1+\varepsilon_g)^6 = \frac{L^6\,T\,(1+\varepsilon_g)^6}{2N}.$$

因此，只需取 $N \ge \dfrac{L^6\,T\,(1+\varepsilon_g)^6}{2\,\delta_{\mathrm{TV}}}$ 即可保证 $\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \le \delta_{\mathrm{TV}}$。

最后，由于朴素提议序贯蒙特卡洛运行 T 个传播步骤（每步代价为 $O(N)$），再加上终止步骤的重加权/重采样步骤，总时间为 $O(NT)$；代入上述对 N 的选择，得到 $O\!\left(\dfrac{L^6\,(1+\varepsilon_g)^6\,T^2}{\delta_{\mathrm{TV}}}\right)$。 □

这些界的幂次上没有 T，因此无需将 ε_g 限制在 $O(1/T)$ 范围内，以避免指数复杂度。我们还指出，这并不违反我们先前的下界，因为该下界是在局部贝尔曼误差的假设下证明的。对于 [12]也同样如此，它所采用的假设与假设H.1完全相同。

假设 H.1 与假设 3.2 可以相互转换：若假设 H.1 在 $1+\varepsilon_g$下成立，则假设 3.2 在 $(1+\varepsilon) = (1+\varepsilon_g)^2$下成立。反之，若假设 3.2 在 $(1+\varepsilon)$下成立，则假设 H.1 在 $(1+\varepsilon_g) = (1+\varepsilon)^T$下成立。

对于任意提议分布 π_{p}（包括最优提议分布），粒子复杂度也有类似的结果。我们有以下结论

命题 H.3（任意提议分布的粒子复杂度）。 固定时域 $T \ge 2$ 和目标精度 $\delta_{\mathrm{TV}} \in (0,\,1)$。在假设 *3.1* 和 *H.1*下，设 $\bar{\eta}_{T+1}^N := \mathbb{E}[\eta_{T+1}^N]$ 为一个序贯蒙特卡洛采样器在终端重采样步骤之后的输出分布$T+1$，其中 $M_{T+1} =$ 为恒等映射。进一步假设

$$\max_{t\in[T]}\ \max_{s_{1:t}\in\mathcal{S}_{1:t}}\ \max\left\{\frac{\pi_{\mathrm{ref}}(s_t \mid s_{1:t-1})}{\pi_{\mathrm{p}}(s_t \mid s_{1:t-1})}\cdot\frac{\hat{V}(s_0,s_{1:t})}{\hat{V}(s_0,s_{1:t-1})},\ \frac{\pi_{\mathrm{p}}(s_t \mid s_{1:t-1})}{\pi_{\mathrm{ref}}(s_t \mid s_{1:t-1})}\cdot\frac{\hat{V}(s_0,s_{1:t-1})}{\hat{V}(s_0,s_{1:t})}\right\} \le L_{\mathrm{p}}.$$

那么，保证 $\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \le \delta_{\mathrm{TV}}$ 的一个充分条件是

$$N \ge \frac{L_{\mathrm{p}}^6 T(1+\varepsilon_g)^6}{2\delta_{\mathrm{TV}}}.$$

具体而言，如果 π_{p} 是最优提议分布，我们还有

$$N \;\ge\; \frac{\left((1+\varepsilon_g)^{12}-1\right)T\,(1+\varepsilon_g)^6}{2\,\delta_{\mathrm{TV}}}$$

是 $\|\bar{\eta}_{T+1}^N - \eta_{T+1}\|_{\mathrm{TV}} \le \delta_{\mathrm{TV}}$ 的一个充分条件。

证明。 证明与定理 5.1、定理 D.1 和命题 H.2的证明类似。 □