预印本卷：1–35，2026

Pol平均奖励多链马尔可夫决策过程中的icy梯度算法 s

Jongmin Lee[1] **和 Ernest K. Ryu**[2]

arXiv:2602.18003v1 [math.OC] 20 Feb 2026

摘要

尽管已有大量研究分析折扣累积奖励马尔可夫决策过程的策略梯度方法，但针对平均奖励马尔可夫决策过程的策略梯度方法的既有工作仍然有限，且大多数现有结果仅限于遍历或单链设置。在本工作中，我们首先基于常返状态与暂态状态分类的不变性，为平均奖励多链马尔可夫决策过程建立了策略梯度定理。在此基础上，我们开展了更精细的分析，并得到一系列收敛与样本复杂度结果，从而推进了对该设置的理解。特别地，我们证明所提出的 α-截断策略镜像上升算法相对于正策略可达到 ϵ-最优策略。

关键词：平均奖励多链马尔可夫决策过程，访问测度，常返状态与暂态状态，策略梯度定理，策略镜像上升，收敛与样本复杂度分析

1. 引言

平均奖励马尔可夫决策过程为建模序贯决策问题提供了基本框架，其目标是最大化长期稳态性能。在动态规划与强化学习文献中，基于价值的方法与基于策略的方法均已在平均奖励设定下得到广泛研究。然而令人惊讶的是，自具有开创性的策略梯度定理——其提出者为Sutton 等人（1999年）——以来，针对一般多链马尔可夫决策过程的策略梯度定理尚未建立。现有关于平均奖励准则下策略梯度方法的分析仅限于遍历或单链马尔可夫决策过程，多链马尔可夫决策过程中策略梯度方法的收敛性仍是一个未决问题。

贡献。在本文中，我们研究平均奖励多链马尔可夫决策过程中策略梯度方法的收敛性。利用状态的常返-暂态分类概念，我们通过引入新的访问测度，首先建立了多链马尔可夫决策过程的策略梯度定理；这些访问测度分别称为常返访问测度和暂态访问测度。在此框架基础上，我们分析了表格设置下 α-截断策略镜像上升的收敛性，并将分析进一步推广到生成模型设置。值得注意的是，针对一般多链马尔可夫决策过程，我们在表格设置和生成模型设置下均建立了关于正策略的 ϵ-最优性保证。

1.1. 先前研究

平均奖励马尔可夫决策过程。 平均奖励马尔可夫决策过程的设定最早在动态规划文献中由Howard（1960年）提出，而Blackwell （1962年）为其分析建立了理论框架。在强化学习（RL）中，平均奖励马尔可夫决策过程主要在转移矩阵和奖励未知的基于样本的设置中加以考虑（Mahadevan, 1996年；Dewanto 等人，2020年）。针对这一设置，人们提出了多种方法：基于模型的方法

1：首尔国立大学 <dlwhd2000@snu.ac.kr>，2：加州大学洛杉矶分校 <eryu@math.ucla.edu>

（Jin 和 Sidford，2021；Tuynman 等人，2024年；Zurek 和 Chen，2024年）、基于价值的方法（Wan 等人，2024年；Bravo 和 Cominetti，2024年；Chen，2025）以及策略梯度方法（Bai 等人，2024年；Murthy 和 Srikant，2023年；Kumar 等人，2024年）。具体而言，在生成模型下获得 ϵ-最优解的样本复杂度（Wang，2017；Zhang 和 Xie，2023年；Li 等人，2024年；Jin 等人，2024年；Lee 等人，2025）以及遗憾最小化框架（Burnetas 和 Katehakis，1997年；Jaksch 等人，2010；Zhang 和 Ji，2019年；Boone 和 Zhang，2024年）已在理论方面得到积极研究。

策略梯度方法。策略梯度方法（Williams，1992年；Sutton 等人，1999年；Konda 和 Tsitsiklis，1999年；Kakade，2001年）是基础性的强化学习算法，通常使用深度神经网络进行策略参数化（Schulman 等人，2015、2017）。与其实践成功相一致，策略梯度变体的收敛与样本复杂度已在多种设定下得到广泛研究（Shani等人，2020年；Mei 等人，2020年；Agarwal 等人，2021；Cen 等人，2022年；Xiao，2022年；Bhandari 和 Russo，2024年）。

对于平均回报马尔可夫决策过程，Sutton 等人 （1999年）；Marbach 和 Tsitsiklis （2001年）；Baxter 和 Bartlett （2000）在单链马尔可夫决策过程中建立了策略梯度定理。最近，Murthy 和 Srikant （2023年）在马尔可夫决策过程的不可约性假设下建立了自然策略梯度的全局收敛，并且 Kumar 等人 （2024年）给出了表格型遍历马尔可夫决策过程中投影策略梯度的全局收敛性分析。一般参数化策略梯度方法的遗憾保证由Bai 等人 （2024年）研究，随后，Ganesh 等人 （2025）通过方差缩减技术改进了样本复杂度。在单链马尔可夫决策过程中，Ganesh 和 Aggarwal（2025）分析了批量自然演员–评论家算法，并且Li 等人 （2024年）在假设混合时间的条件下，为生成式采样模型与马尔可夫采样模型建立了样本复杂度保证。

然而，在多链和弱连通马尔可夫决策过程中，目前尚无关于策略梯度方法的已有研究结果。据我们所知，策略梯度定理尚未建立，目前仅存在基于模型和基于价值的方法（Wei 等人，2020年；Zurek 和 Chen，2024年；Lee 和 Ryu，2025a）。

1.2. 预备知识与符号

平均回报马尔可夫决策过程。令 $\mathcal{M}(\mathcal{X})$ 表示集合 $\mathcal{X}$ 上概率分布构成的空间。记 $(\mathcal{S},\ \mathcal{A},\ P,\ r)$ 为一个无限时域无折扣马尔可夫决策过程，其有限状态空间为 $\mathcal{S}$，有限动作空间为 $\mathcal{A}$，转移矩阵为 $P\colon \mathcal{S}\times\mathcal{A}\to\mathcal{M}(\mathcal{S})$，有界奖励为 $r\colon \mathcal{S}\times\mathcal{A}\to[-R,R]$。对于策略，记$\pi\colon \mathcal{S}\to\mathcal{M}(\mathcal{A})$，

$$J^\pi(s)=\liminf_{H\to\infty}\frac{1}{H}\mathbb{E}_\pi\left[\sum_{h=0}^{H-1}r(s_h,a_h)\,\middle|\,s_0=s\right],$$

$$K^\pi(s,a)=\liminf_{H\to\infty}\frac{1}{H}\mathbb{E}_\pi\left[\sum_{h=0}^{H-1}r(s_h,a_h)\,\middle|\,s_0=s,a_0=a\right]$$

对于给定策略的平均回报，以及

$$V^\pi(s)=\lim_{H\to\infty}\frac{1}{H}\sum_{h=1}^{H}\mathbb{E}_\pi\left[\sum_{i=0}^{h-1}\left(r^\pi(s_i)-J^\pi(s_i)\right)\,\middle|\,s_0=s\right]$$

$$Q^\pi(s,a)=\lim_{H\to\infty}\frac{1}{H}\sum_{h=1}^{H}\mathbb{E}_\pi\left[\sum_{i=0}^{h-1}\left(r(s_i,a_i)-K^\pi(s_i,a_i)\right)\,\middle|\,s_0=s,a_0=a\right]$$

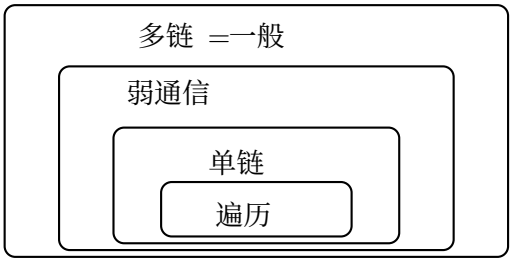

图 1：马尔可夫决策过程的分类。

对于状态和状态-动作相对价值（在非周期性假设下，定义中可以省略关于 H 的平均（Puterman, 2014，章节8.2））， 其中 $\mathbb{E}_\pi$ 表示由 P和 π 诱导的所有轨迹（s_0， a_0， s_1， a_1， ....., s_{H-1}， a_{H-1}）上的期望值，并将 r^π（s） $=\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}$ [r（s， a）]作为由策略 π诱导的奖励。给定 $\mu \in \mathcal{M}$（$\mathcal{S}$），定义 $J^\pi_\mu = \mathbb{E}_{s_0\sim\mu}[J^\pi_{s_0}]$。若 $J^\star = \max_\pi J^\pi$，则称 $J^\star$为最优平均奖励，并且最优平均奖励在多链马尔可夫决策过程中始终存在（Puterman，2014，章节9.1）。将 P^π（s， s'） $=$Prob（$s \to s' \,|\, a \sim\pi$（$\cdot\,|\,s$） $s' \sim P$（$\cdot\,|\,s$， a））记为由策略 π诱导的转移矩阵。我们记

$$P_\star = \lim_{H\to\infty} \frac{1}{H} \sum_{i=0}^{H-1} P^i$$

随机矩阵的 Ces`aro 极限 $P_\star$。（随机矩阵的 Ces`aro 极限总是存在（Puterman，2014，定理 A.6））。因此，根据定义，我们可以写成（Puterman，2014，第 8.2 节）

$$J^\pi = P^\pi_\star r^\pi, \qquad V^\pi = (I - P^\pi + P^\pi_\star)^{-1}(I - P^\pi_\star) r^\pi, \qquad K^\pi = PJ^\pi, \qquad Q^\pi = r + PV^\pi - K^\pi.$$

最后，对于每个策略 π，已知 V^π 和 J^π 满足以下贝尔曼方程（Puterman，2014，定理 8.2.6）：

$$P^\pi J^\pi = J^\pi, \qquad r^\pi + P^\pi V^\pi = J^\pi + V^\pi.$$

马尔可夫决策过程的分类。根据转移矩阵的结构，马尔可夫决策过程分类如下。（有关马尔可夫决策过程的基本概念定义，如不可约类、通信类、可达性等，请参阅Puterman (2014，附录 A.2）以及Br´emaud （2013，第 3 节。）

如果每个策略 π 所诱导的转移矩阵都只有一个常返类且是非周期的，则该马尔可夫决策过程是遍历的。如果每个确定性策略对应的转移矩阵都由一个不可约常返类以及可能为空的暂态状态集组成，则该马尔可夫决策过程是单链的。如果存在一个闭状态集，其中每个状态在某个确定性策略下都可从该集合中的任何其他状态到达，且另有一个可能为空的状态集，该集合中的状态在任何策略下都是暂态，则该马尔可夫决策过程是弱通信的。如果任何确定性策略对应的转移矩阵都包含一个或多个不可约常返类，则该马尔可夫决策过程是多链的。

如果马尔可夫决策过程是遍历的，则它是单链的；如果马尔可夫决策过程是单链的，则它是弱通信的；如果马尔可夫决策过程是弱通信的，则它是多链的。由于每个马尔可夫决策过程都是多链的，我们将多链和一般 这两个表达互换使用。

如果马尔可夫决策过程是单链的，那么对于每个策略 π，都存在唯一的平稳分布 $g^\pi \in \mathcal{M}(\mathcal{S})$满足$(g^\pi)^\top P^\pi = g^\pi$ (Puterman，2014，定理 8.3.2）。如果马尔可夫决策过程是遍历的，那么对于每个策略

π， g^π 严格为正（Levin 和 Peres，2017，命题 1.19）。如果马尔可夫决策过程是单链的，那么对于每个策略 π， J^π都是一个常数向量，即 $J^\pi = c\mathbf{1}$对于某个 $c \in \mathbb{R}$其中 $\mathbf{1} \in \mathbb{R}^n$是各分量均为1的向量。

在下一章节中，我们讨论多链马尔可夫决策过程中映射 $\pi \mapsto J^\pi$ 的连续性。由于 $|\mathcal{S}|$ 和 $|\mathcal{A}|$ 是有限的，我们可以将 π 和 J^π 视为有限维向量，即$\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 和 $J^\pi \in \mathbb{R}^{|\mathcal{S}|}$。因此， $\pi \mapsto J^\pi$ 的连续性可以解释为在通常的度量下，从 $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 到 $\mathbb{R}^{|\mathcal{S}|}$ 的映射的连续性。我们定义策略类

$$\Pi = \text{set of all policies} = \mathcal{M}(\mathcal{A})^{\mathcal{S}}, \qquad \Pi_+ = \{\pi \in \Pi \,|\, \pi(a \,|\, s) > 0 \text{ for all } s, a\},$$

使得 Π_+ 是 Π的（相对）内部。另将 $\mathcal{M}_+(\mathcal{X})$ 定义为集合 $\mathcal{X}$ 上具有全支撑的概率分布空间。

2. 策略梯度的常返-暂态理论

在本章节中，我们将马尔可夫链的常返-暂态理论应用于平均奖励多链马尔可夫决策过程，并引入常返访问测度和暂态访问测度的概念。基于这些概念，我们为平均奖励多链马尔可夫决策过程建立了策略梯度定理。

2.1. 状态的常返-暂态分类

定义 1 给定策略 $\pi \in \Pi$，若状态 $s \in \mathcal{S}$ 从 s 出发的返回时间以概率1为有限，则该状态为常返状态。否则， s为暂态。

等价地，若 n_s 是表示访问状态 s 、并从s出发的次数的随机变量，则 s 为常返状态当且仅当 $\mathbb{E}_\pi[n_s] = \sum_{k=0}^{\infty}(P^\pi)^k(s,\ s) = \infty$，否则它为暂态（Br´emaud，2013，定理 3.1.3）。

设 $\pi \in \Pi$。对于给定的 P^π，这些状态可分类为常返状态与暂态状态，且该马尔可夫链可规范地表示如下（Puterman，2014，附录 A.2）：

$$P^\pi = \begin{bmatrix} R_1^\pi & 0 & 0 & \cdot & \cdot & 0 \\ 0 & R_2^\pi & 0 & \cdot & \cdot & 0 \\ & & & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \cdot & R_m^\pi & 0 \\ S_1^\pi & S_2^\pi & \cdot & \cdot & S_m^\pi & T^\pi \end{bmatrix}, \qquad P_\star^\pi = \begin{bmatrix} R_{1,\star}^\pi & 0 & 0 & \cdot & \cdot & 0 \\ 0 & R_{2,\star}^\pi & 0 & \cdot & \cdot & 0 \\ 0 & & & \cdot & \cdot & 0 \\ 0 & 0 & 0 & \cdot & R_{m,\star}^\pi & 0 \\ S_{1,\star}^\pi & S_{2,\star}^\pi & \cdot & \cdot & S_{m,\star}^\pi & 0 \end{bmatrix},$$

其中 R_1^π，……， R_m^π、 T^π、 S_1^π，……， S_m^π 分别表示再生状态之间、暂态状态之间以及从暂态到再生状态的转移概率。（注意，多链马尔可夫决策过程可以有 $m \geq 1$ 个常返类，而单链马尔可夫决策过程必须只有一个常返类。） $P_\star^\pi$是 P^π 的 Ces`aro 极限，并且已知 R_i^π，$_\star = \mathbf{1}(g_i^\pi)^\top$，其中 g_i^π是概率矩阵 R_i^π的唯一平稳分布，且 S_i^π，$_\star = (I - T^\pi)^{-1}S_i^\pi R_i^\pi$，$_\star$对 $i = 1$，……， m成立（Puterman，2014，附录 A.4）。

对于任意 $\pi \in \Pi$，都存在此类规范表示，但状态的常返-暂态分类以及 P^π 相应规范分解的结构可能随 $\pi \in \Pi$ 而变化。然而，对于所有 $\pi \in \Pi_+$，常返-暂态分类保持不变。

事实 1 *(Lee 和 Ryu，2025年b，* 命题 *1)* 状态的常返-暂态分类不取决于 $\pi \in \Pi_+$ 的选择。更具体地说，状态集 $\mathcal{S}$ 分解为暂态状态 $\mathcal{T}$ 和 m 常返类 $\mathcal{R}_1$，……，$\mathcal{R}_m$ 其中某个 $m \in \mathbb{N}$，且该分解对于所有 $\pi \in \Pi_+$都相同。

我们进一步说明。对于任意 $\pi \in \Pi_+$，常返-暂态分类由转移核 P 决定，而非由 $\pi \in \Pi_+$ 的具体选择决定。虽然策略 $\pi \in \Pi \setminus \Pi_+$（即对某些动作赋予零概率的策略）可能导致不同的分类，但我们所考虑的算法（以及采用 Softmax 输出层的深度强化学习算法）是在 Π_+ 上进行搜索，而非在完整集合 Π 上搜索。

根据事实 1，因此，对于任意 $\pi \in \Pi_+$，我们固定常返类集合 $\bigcup_{i=1}^{m} \mathcal{R}_i \subset \mathcal{S}$和暂态集合 $\mathcal{T} \subset \mathcal{S}$，其中 m 表示常返类的数量。我们还记

$$\mathcal{R} := \bigcup_{i=1}^{m} \mathcal{R}_i$$

为所有常返状态构成的集合。

2.2. J^π 在 Π_+ 上的连续性

对于多链马尔可夫决策过程，平均奖励 $J^\pi(s)$ 可能是策略$\pi \in \Pi$ 的不连续函数，（正如（Schweitzer，1968年，第8节）中的反例所示。不过，我们确实能在 $\pi \in \Pi_+$ 上保证连续性。

引理 2 在多链马尔可夫决策过程中，对于任意固定的 $\mu \in \mathcal{M}(\mathcal{S})$，映射 $\pi \mapsto J^\pi$和 $\pi \mapsto J^\pi_\mu$（其中 $J^\pi_\mu = \mathbb{E}_{s\sim\mu}[J^\pi(s)]$）在 Π_+ 上连续。

换言之，我们在 Π 的（相对）内部具有连续性，而（Schweitzer，1968年，第 8 章）中所述的不连续性只可能出现在 Π的（相对）边界上。在我们的策略梯度方法中，搜索范围限制为策略 $\pi \in \Pi_+$。

接下来，定义

$$J^\star_{+,\mu} = \sup_{\pi\in\Pi_+} \mathbb{E}_{s\sim\mu}[J^\pi(s)], \qquad J^\star_\mu = \max_{\pi\in\Pi} \mathbb{E}_{s\sim\mu}[J^\pi(s)],$$

其中 μ是一个初始状态分布。根据定义，$J^\star_\mu \geq J^\star_{+,\,\mu}$。由于我们的策略梯度方法是在 Π_+上搜索，因此应将其视为针对 $J^\star_{+,\,\mu}$进行优化。因此，给定 $\mu \in \mathcal{M}(\mathcal{S})$，若策略 π满足 $J^\star_{+,\,\mu} - J^\pi_\mu \leq \epsilon$，则称其为相对于正策略的 ϵ-最优策略。

2.3. 访问度量

在先前研究中，平均奖励单链马尔可夫决策过程在策略 π 下的状态访问度量被定义为平稳分布，它与初始状态 s_0无关。相比之下，对于多链马尔可夫决策过程，平稳分布通常取决于起始状态，并且可能不唯一。为此，我们为多链设定引入了新的访问度量。

首先，定义关于策略 π 和起始状态 s_0 的<风格 id='1'>常返访问测度</风格>为

$$d^\pi_{s_0}(s) = \lim_{H\to\infty} \frac{1}{H} \sum_{h=0}^{H-1} \mathrm{Prob}(s_h = s \mid s_0; P^\pi) = e_{s_0}^\intercal P^\pi_\star e_s,$$

其中， e_s 是第 s 个单位向量。同样地，定义关于策略 π 和起始状态分布 μ 的<风格 id='3'>常返访问测度</风格>为

$$d_\mu^\pi = \mathbb{E}_{s_0\sim\mu}[d_{s_0}^\pi].$$

注意，对于所有暂态状态 s，我们有 $d_{s_0}^\pi(s) = d_\mu^\pi(s) = 0$。

接下来，依照（Lee 和 Ryu，2025年b），定义暂态矩阵：

$$\bar{T}^\pi = \begin{bmatrix} 0 & 0 \\ 0 & T^\pi \end{bmatrix}, \qquad \text{i.e.,} \qquad \bar{T}^\pi(s_1, s_2) = \begin{cases} P^\pi(s_1, s_2) & \text{if } s_1, s_2 \text{ are both transient} \\ 0 & \text{otherwise.} \end{cases}$$

事实 2 *(Berman and Plemmons，1994，引理 8.3.20)* T^π 的谱半径严格小于1。

根据事实2和经典的诺伊曼级数论证，我们有$(I - \bar{T}^\pi)^{-1} = \sum_{i=0}^{\infty}(\bar{T}^\pi)^i$。现在，我们将相对于策略暂态访问测度 π 和起始状态 s_0 定义如下：

$$\delta_{s_0}^\pi(s) = \sum_{h=0}^{\infty} \text{Prob}(s_h = s \mid s_0; \bar{T}^\pi) = e_{s_0}^\intercal (I - \bar{T}^\pi)^{-1} e_s,$$

以及 暂态访问测度相对于策略 π 和起始状态分布 μ a s

$$\delta_\mu^\pi = \mathbb{E}_{s_0\sim\mu}[\delta_{s_0}^\pi].$$

我们指出，这个暂态访问测度并不是概率测度。此外，重要的是，最大值$_{s,\ s_0\in\mathcal{S}}\ \delta_{s_0}^\pi(s) < \infty$ 根据事实 2。

借助这些常返与暂态访问测度，我们可以在平均奖励多链马尔可夫决策过程设定下得到如下性能差异引理，该引理将在策略梯度算法的分析中发挥关键作用。

引理 3（性能差异引理）考虑一个多链马尔可夫决策过程。对于 π， $\pi' \in \Pi_+$ 和$\mu \in \mathcal{M}_+($ $\mathcal{S})$，

$$J_\mu^\pi - J_\mu^{\pi'} = \sum_{\substack{s\in\mathcal{R}\\ a\in\mathcal{A}}} d_\mu^\pi(s)\big(\pi(a \mid s) - \pi'(a \mid s)\big)Q^{\pi'}(s, a) + \sum_{\substack{s\in\mathcal{T}\\ a\in\mathcal{A}}} \delta_\mu^\pi(s)\big(\pi(a \mid s) - \pi'(a \mid s)\big)K^{\pi'}(s, a).$$

注意，引理 3可视为对以下单链设定下性能差异引理的推广。

推论4 *(Even-Dar 等人，2009，引理 4.1)* 考虑一个单链马尔可夫决策过程。对于 π， $\pi' \in \Pi_+$ 和$\mu \in \mathcal{M}_+(\mathcal{S})$，

$$J_\mu^\pi - J_\mu^{\pi'} = \sum_{s\in\mathcal{R}}\sum_{a\in\mathcal{A}} d_\mu^\pi(s)\big(\pi(a \mid s) - \pi'(a \mid s)\big)Q^{\pi'}(s, a).$$

证明 对于 $\pi' \in \Pi_+$，单链假设意味着 $K^{\pi'} = PJ^{\pi'} = c^{\pi'}\mathbf{1}$ 对某个 $c^{\pi'} \in \mathbb{R}$(Puterman，2014，定理 8.3.2）。因此，引理 3 中的第二项消失。 ■

2.4. 策略梯度定理

现在，我们可以给出平均奖励设定下的策略梯度定理。考虑以下优化问题

$$\max_{\theta\in\Theta} J_\mu^{\pi_\theta},$$

其中，$\{\pi_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$ 是关于 θ的一组可微参数化策略。基于前述工具，我们建立如下策略梯度定理。

定理 5（再生与暂态策略梯度）考虑一个多链马尔可夫决策过程。对于 $\pi_\theta \in \Pi_+$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$,

$$\nabla_\theta J_\mu^{\pi_\theta} = \sum_{s\in\mathcal{R}}\sum_{a\in\mathcal{A}} d_\mu^{\pi_\theta}(s)\nabla_\theta\pi_\theta(a\mid s)Q^{\pi_\theta}(s,a) + \sum_{s\in\mathcal{T}}\sum_{a\in\mathcal{A}} \delta_\mu^{\pi_\theta}(s)\nabla_\theta\pi_\theta(a\mid s)K^{\pi_\theta}(s,a).$$

在定理 5中，第一项涉及对再生状态求和，而第二项则对暂态状态求和。从这个角度看，定理5 自然地推广了单链马尔可夫决策过程的经典策略梯度定理。

推论6（再生策略梯度，Sutton 等人（1999年））考虑一个单链马尔可夫决策过程。对于 $\pi_\theta \in \Pi_+$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$,

$$\nabla_\theta J_\mu^{\pi_\theta} = \sum_{s\in\mathcal{R}}\sum_{a\in\mathcal{A}} d_\mu^{\pi_\theta}(s)\nabla_\theta\pi_\theta(a\mid s)Q^{\pi_\theta}(s,a).$$

证明 对于任意 $\pi \in \Pi_+$，单链假设意味着，对于某个 $c^{\pi'} \in \mathbb{R}$，有 $K^\pi = PJ^\pi = c^\pi\mathbf{1}$（Puterman, 2014，定理 8.3.2）。根据乘积法则，$\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a\mid s)K^{\pi_\theta}(s,a) = \nabla_\theta(\sum_{a\in\mathcal{A}}\pi_\theta(a\mid s)K^{\pi_\theta}(s,a)) - \sum_{a\in\mathcal{A}}\pi_\theta(a\mid s)\nabla_\theta K^{\pi_\theta}(s,a) = \nabla_\theta(c^{\pi_\theta}) - \nabla_\theta(c^{\pi_\theta}) = 0$。因此，定理 5 中的第二项消失。 ■

We a还要注意，对于单链马尔可夫决策过程，$d_\mu^\pi = g^\pi$，其中 g^π 是 P^π 的平稳分布 .

3. 策略镜像上升的收敛

在本章节中，我们研究策略镜像上升在 Π_+的一个紧子集上的收敛。

3.1. α截断策略镜像上升

我们考虑直接参数化

$$\pi_\theta(a\mid s) = \theta_{s,a},$$

其中 $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 满足$\sum_{a\in\mathcal{A}}\theta_{s,a} = 1$ 且$\theta_{s,a} \geq 0$ 对所有 $s \in \mathcal{S}$ 和 $a \in \mathcal{A}$成立。在此直接参数化下，我们不区分策略 π_θ 与参数 θ，并用 π_k表示算法的迭代。

令 $h:\ \mathcal{M}(\mathcal{A}) \to \mathbb{R}$ 为严格凸函数，并且在 $\mathcal{M}(\mathcal{A})$ 的相对内部（记作 rint $\mathcal{M}(\mathcal{A})$）上连续可微。定义由 h 生成的 Bregman散度为

$$D(p,p') = h(p) - h(p') - \langle\nabla h(p'), p - p'\rangle, \qquad \forall\, p \in \mathcal{M}(\mathcal{A}),\ p' \in \text{rint } \mathcal{M}(\mathcal{A}).$$

特别地，如果 $h(p) = \frac{1}{2}\|p\|_2^2$，那么 $D(p, p') = \frac{1}{2}\|p - p'\|_2^2$。如果 $h(p) = \sum_{a\in\mathcal{A}} p(a) \log p(a)$，那么$D(p, p') = \sum_{a\in\mathcal{A}} p(a) \log \frac{p(a)}{p'(a)}$。对于 $\rho \in \mathbb{R}_+^{|\mathcal{S}|}$，定义加权散度

$$D_\rho(\pi, \pi') = \sum_{s\in\mathcal{S}} \rho(s)\, D(\pi(\cdot \,|\, s), \pi'(\cdot \,|\, s)).$$

对于策略 π，定义

$$\rho_\mu^\pi(s) = d_\mu^\pi(s)\mathbf{1}_\mathcal{R}(s) + \delta_\mu^\pi(s)\mathbf{1}_\mathcal{T}(s),$$

其中 $\mathbf{1}_\mathcal{R}$ 和 $\mathbf{1}_\mathcal{T}$是示性函数。按照Shani等人(2020年) 和 Xiao(2022年的推导，我们考虑采用加权Bregman散度的策略镜像上升方法:

$$\pi_{k+1} = \operatorname*{argmax}_{\pi\in\mathcal{C}} \left\{ \eta_k \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \nabla J_\mu^{\pi_k}(s,a)\pi(a \,|\, s) - D_{\rho_\mu^{\pi_k}}(\pi, \pi_k) \right\},$$

其中 η_k 是步长， $\mu \in \mathcal{M}_+(\mathcal{S})$ 是具有全支撑的初始状态分布，且 $\mathcal{C}$ 是 Π_+的紧子集。对于评估 $\nabla J_\mu^{\pi_k}$，定理 5 隐含地表明

$$\nabla J_\mu^\pi(s,a) = d_\mu^\pi(s)Q^\pi(s,a)\mathbf{1}_\mathcal{R}(s) + \delta_\mu^\pi(s)K^\pi(s,a)\mathbf{1}_\mathcal{T}(s).$$

对所有 $s \in \mathcal{S}$，令 $G^\pi(s, \cdot) = Q^\pi(s, \cdot)\mathbf{1}_\mathcal{R}(s) + K^\pi(s, \cdot)\mathbf{1}_\mathcal{T}(s)$。那么，对于在 $\mathcal{S}$ 上具有全支撑的 μ，由于 $d_\mu^\pi(s) > 0$， $\delta_\mu^\pi(s') > 0$ 对任意 $s \in \mathcal{R}$ 和 $s' \in \mathcal{T}$ 均成立， π_{k+1} 的更新可按状态分解:

$$\pi_{k+1}(\cdot \,|\, s) = \operatorname*{argmax}_{p\in\mathcal{C}'} \Big\{ \sum_{a\in\mathcal{A}} \eta_k G^{\pi_k}(s,a)p(a) - D\big(p(\cdot), \pi_k(\cdot \,|\, s)\big) \Big\}, \qquad \forall s \in \mathcal{S}.$$

$\mathcal{C} = \Pi$ 的选择看似自然，但在我们的设定中，我们进一步限制策略集，以确保收敛分析中出现的某些系数有限。定义

$$\Pi_\alpha = \{\pi \,|\, \pi(a \,|\, s) \geq \alpha \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}\}, \qquad \mathcal{M}_\alpha(\mathcal{A}) = \{p \,|\, p(a) \geq \alpha \text{ for all } a \in \mathcal{A}\}$$

其中 $\alpha \in (0, 1/|\mathcal{A}|)$。事实上，在 Π_α 上，我们可以确保以下系数均为有限值。

引理 7 若 $\mu \in \mathcal{M}_+(\mathcal{S})$ 且 $\alpha \in (0, 1/|\mathcal{A}|)$，则

$$\max_{\pi\in\Pi_\alpha} \|\rho_\mu^\pi\|_1 := B_\alpha < \infty, \qquad \max_{s\in\mathcal{S},\pi,\pi'\in\Pi_\alpha} \left\| \frac{\rho_\mu^\pi(s)}{\rho_\mu^{\pi'}(s)} \right\|_\infty := C_\alpha < \infty.$$

现在，我们已准备好介绍算法 α——截断策略镜像上升。

算法 1 α-截断策略镜像上升

Input: $\alpha \in (0, 1/|\mathcal{A}|), K, \pi_0 \in \Pi_\alpha, \{\eta_k\}_{k=0}^{K-1} \subset (0, \infty), D(\cdot, \cdot)$
for $k = 0, 1, \ldots, K-1$ **do**
  for $s \in \mathcal{S}$ **do**
    $\pi_{k+1}(\cdot \,|\, s) = \operatorname{argmax}_{p\in\mathcal{M}_\alpha(\mathcal{A})} \{\eta_k \sum_{a\in\mathcal{A}} G^{\pi_k}(s,a)p(a) - D\big(p(\cdot), \pi_k(\cdot \,|\, s)\big)\}$
  end for
end for
Output: π_K

若 h 是一个勒让德函数（定义请参见附录 A ），例如欧几里得范数或负熵，则 π_{k+1} 总是存在（Xiao，2022年，引理 6）。具体而言，若 Bregman散度为欧几里得距离，则 α-截断策略镜像上升可简化为

$$\pi_{k+1}(\cdot \mid s) = \underset{p \in \mathcal{M}_\alpha}{\operatorname{argmin}} \|p(\cdot) - (\pi_k(\cdot \mid s) + \eta_k G^{\pi_k}(\cdot \mid s))\|_2^2, \qquad \forall\, s \in \mathcal{S}$$

而当 Bregman散度为 KL 散度时，它可化简为

$$\pi_{k+1}(\cdot \mid s) = \underset{p \in \mathcal{M}_\alpha}{\operatorname{argmin}} \mathrm{KL}\big(p(\cdot) \mid \pi_k(\cdot \mid s) \exp(\eta_k G^{\pi_k}(s, \cdot))/Z_s\big), \qquad \forall\, s \in \mathcal{S},$$

其中 $Z_s = \sum_{a \in \mathcal{A}} \pi_k(a \mid s) \exp(\eta_k(G^{\pi_k}(s,a))$。如果 $\alpha = 0$，这些方法分别称为投影 Q-下降法和乘法权重更新（Xiao，2022年；Freund 和 Schapire，1997年）。我们注意到，这两个优化子问题均可通过欧几里得投影算法和 KL 投影算法求解，时间复杂度为 $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$（Wang 和 Carreira-Perpin´an，2013；Herbster 和 Warmuth，2001年）。我们在附录 F 中介绍这两种投影算法。

在接下来的小节中，我们将证明 α-截断策略镜像上升的收敛性。对于给定的 μ，定义

$$\pi_\alpha \in \underset{\pi \in \Pi_\alpha}{\operatorname{argmax}} J_\mu^\pi.$$

由引理 a2， π_α存在，我们将证明 J^{π_k}相对于 J^{π_α}在 $\pi_k \in \Pi_\alpha$.

3.2. 恒定步长下的次线性收敛

我们首先证明恒定步长策略梯度算法的次线性收敛。作为分析的第一步，我们给出以下引理，该引理保证由 α-裁剪策略镜像上升生成的策略单调改进。

引理 8 考虑一个多链马尔可夫决策过程。对于 $\pi_0 \in \Pi_+$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$， α-裁剪策略镜像上升会生成一个满足以下条件的策略序列 $\{\pi_k\}_{k=1}^\infty$

$$J_\mu^{\pi_k} \le J_\mu^{\pi_{k+1}}.$$

通过引理 8，我们得到以下关于 α-裁剪策略镜像上升的收敛结果。附录 C 中的证明 Xiao（2022年）紧随其后；该文考虑了折扣奖励设定，并使用了折扣状态访问分布。

定理 9 考虑一个多链马尔可夫决策过程。对于 $\pi_0 \in \Pi_\alpha$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，采用恒定步长 $\eta > 0$ 的 α-截断策略镜像上升会生成一个满足以下条件的策略序列 $\{\pi_k\}_{k=1}^\infty$

$$J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \le \frac{1}{k+1}\left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta} + C_\alpha(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0})\right).$$

定理 9表明 $J_\mu^{\pi_k} \to J_\mu^{\pi_\alpha}$以次线性速率收敛，并且由于$J_\mu^{\pi_\alpha} \to J_{+,\,\mu}^\star$当 $\alpha \to 0$时，如果我们选择足够小的 α使得 $J_{+,\,\mu}^\star - J_\mu^{\pi_\alpha} < \frac{\epsilon}{2}$，则可以通过定理 π得到满足$J_{+\text{的策略},\,\mu}^\star - J_\mu^\pi < \epsilon$定理 9。

3.3. 自适应步长下的线性收敛

接下来，我们介绍采用自适应步长的 α-裁剪策略镜像上升的线性 收敛速率。

定理 10 考虑一个多链马尔可夫决策过程。对于 $\pi_0 \in \Pi_\alpha$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，步长为 $(C_\alpha - 1)\eta_{k+1} \geq C_\alpha \eta_k > 0$ 的 α-裁剪策略镜像上升会生成满足下述条件的策略序列 $\{\pi_k\}_{k=1}^{\infty}$

$$J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \left(1 - \frac{1}{C_\alpha}\right)^k \left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta_0(C_\alpha - 1)} + J_\mu^{\pi_\alpha} - J_\mu^{\pi_0}\right).$$

尽管自适应步长能得到线性收敛速率，但它需要已知 C_α才能设定步长。相比之下，恒定步长始终保证次线性速率。我们指出，据我们所知，定理9 和10是平均奖励多链马尔可夫决策过程上策略梯度方法的首个收敛结果。

算法 2评价器

Input: $\pi \in \Pi_+, N, H, N', H' \in \mathbb{Z}_+$
for $j = 1, \ldots, N$ **do**
  for $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ **do**
    Generate $\{(s_0, a_0), (s_1, a_1), \ldots, (s_H, a_H) \,|\, s_i \sim P(\cdot \,|\, s_{i-1}, a_{i-1}), a_i \sim \pi(\cdot \,|\, s_i)\}$
    $K^j(s_0, a_0) = \frac{1}{H+1} \sum_{i=0}^{H} r_j(s_i, a_i)$.
  end for
end for
$\hat{K}^\pi = \frac{1}{N} \sum_{j=1}^{N} K^j$.
for $j = 1, \ldots, N'$ **do**
  for $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ **do**
    Generate $\{(s_0, a_0), (s_1, a_1), \ldots, (s_{H'}, a_{H'}) \,|\, s_i \sim P(\cdot \,|\, s_{i-1}, a_{i-1}), a_i \sim \pi(\cdot \,|\, s_i)\}$
    $\hat{Q}^j(s_0, a_0) = \frac{1}{H'+1} \sum_{h=0}^{H'} \sum_{i=0}^{h} \left(r_j(s_i, a_i) - \hat{K}^\pi(s_i, a_i)\right)$.
  end for
end for
$\hat{Q}^\pi = \frac{1}{N'} \sum_{j=1}^{N'} \hat{Q}^j$
for $s \in \mathcal{S}$ **do**
  $\hat{G}^\pi(s, \cdot) = \hat{Q}^\pi(s, \cdot)\mathbf{1}_{\mathcal{R}}(s) + \hat{K}^\pi(s, \cdot)\mathbf{1}_{\mathcal{T}}(s)$
end for
Output: $\hat{G}^\pi$

4. 策略镜像上升的样本复杂度

在本章节中，我们将 α-截断策略镜像上升的分析扩展到转移概率未知的采样设定。具体而言，我们假设可以使用生成模型（Kearns 和 Singh，1998年），该模型能够针对任意给定的状态和动作，独立采样得到下一状态。

4.1. 使用生成模型近似 G^π

借助生成模型，对于给定的策略 π 以及任意状态-动作对（s，a）$\in \mathcal{S}\times\mathcal{A}$，我们可以生成时域为 H 的独立轨迹：

$$\left\{\left(s_0=s,a_0=a\right),\left(s_1,a_1\right),\ldots,\left(s_{H-1},a_{H-1}\right)\right\}.$$

利用这些样本，我们通过算法 2 所述的评价器方法来近似 $G^\pi(s,\cdot)=Q^\pi(s,\cdot)\mathbf{1}_{\mathcal{R}}(s)+K^\pi(s,\cdot)\mathbf{1}_{\mathcal{T}}(s)$，$\forall s\in\mathcal{S}$。直观地说，$K^\pi,Q^\pi$ 是对 K^π,Q^π进行 H时域截断。

为了给出生成模型下评价器方法的样本复杂度结果，我们定义期望目标时间如下。对于 $\pi\in\Pi$以及每个常返类 $\mathcal{R}_i$，

$$t_{\text{tar}}^{\pi,i}:=\sum_{s'\in\mathcal{R}_i}\rho_i^\pi(s')\,\mathbb{E}_\pi\left[t_{s'}^i\,|\,s_0=s\right]\ \text{for}\ i=1,\ldots,m,\qquad t_{\text{tar}}^\pi=\max_{1\le i\le m}t_{\text{tar}}^{\pi,i}$$

其中 $t_{s'}:=$ 下确界$\{t\ge 0:\ s_t=s'\}$ 对于 s，$s'\in\mathcal{R}_i$，以及 ρ_i^π是的平稳分布 R_i^π。众所周知，当状态空间有限时，期望目标时间总是有限的，且与起始状态无关（Levin和Peres，2017）。

定理 11 考虑一个多链马尔可夫决策过程。令 $\epsilon>0$，$\delta>0$。对于 π_0、$\pi\in\Pi_+$和 $\mu\in\mathcal{M}_+$（$\mathcal{S}$），ˆ以 $1-\delta$ 概率，评论家方法的输出满足 $\|G^\pi-G^\pi\|_\infty\le\epsilon$，其样本复杂度为

$$\mathcal{O}\left(\frac{R^3(t_{\text{tar}}^\pi)^3\,\|Q^\pi\|_\infty^4\,|\mathcal{S}||\mathcal{A}|}{\epsilon^6}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)\right).$$

4.2. 通过采样观察状态分类

ˆˆˆ要使用评论家方法计算 G^π 和 Q^π，K^π，需要对状态进行分类。

对于给定的 π 和每个常返类 $\mathcal{R}_i^\pi$，将暂态半衰期和覆盖时间定义为

$$t_{\frac{1}{2},\pi}=\min\left\{t\ge 1:\left\|(\bar{T}^\pi)^t\right\|_\infty\le\frac{1}{2}\right\},\quad t_{\text{cov}}^{\pi,i}=\max_{s\in\mathcal{R}_i}\mathbb{E}_\pi[t_{\text{cov}}^i\,|\,s_0=s]\quad\text{and}\quad t_{\text{cov}}^\pi=\max_{1\le i\le m}t_{\text{cov}}^{\pi,i}$$

其中 $\bar{T}^\pi$ 是暂态矩阵，且 $t_{\text{cov}}^i:=$ 无穷$\{t\ge 0:\ \{s_j\}_{j=0}^{t-1}=\mathcal{R}_i\}$。已知当状态空间有限时，这些量均为有限值（Lee 和 Ryu，2025年b；Levin 和 Peres，2017）。

引理12 考虑一个多链马尔可夫决策过程。令 $\delta>0$。给定 $\pi\in\Pi_+$，设 $M_1=\lceil t_{\frac{1}{2},\pi}-\log(\frac{2}{\delta})\rceil$和 $M_2=\lceil et_{\text{cov}}^\pi\log\left(\frac{2}{\delta}\right)\rceil$。以 $1-\delta$的概率，对于生成的长度为 M_1+M_2 的单条轨迹（s_0、a_0、s_1、a_1、…、$s_{M_1+M_2-1}$、$a_{M_1+M_2-1}$），对于某个 i，$\{s_j\}_{j=M_1}^{M_1+M_2-1}=\mathcal{R}_i^\pi$。

该引理表明，从任意状态s_0出发，对长度为 M_1+M_2 的轨迹进行采样，即可恢复一个常返类中的全部状态。此外，通过检查s_0 是否属于该集合，我们可以确定 s_0的分类。在生成模型下，对其余尚未分类的状态重复这一过程，即可识别所有状态的分类，样本复杂度为

$$\mathcal{O}\left((|\mathcal{T}|+m)\left(t_{\frac{1}{2},\pi}+t_{\text{cov}}^\pi\right)\log\left(\frac{|\mathcal{T}|+m}{\delta}\right)\right),$$

其中 m 表示常返类的数量。

4.3. 随机 α-截断策略镜像上升的样本复杂度

基于　继前几章的结果之后，我们现在介绍随机 α-截断策略镜像上升算法　　nt:

算法 3 随机 α-截断策略镜像上升 输入： $\alpha \in (0, 1/|\mathcal{A}|)$ $K, \pi_0 \in \Pi_\alpha, \{\eta_k\}_{k=0}^{K-1} \subset (0, \infty)$ $D(\cdot, \cdot)$ $\{N_k, H_k, N'_k, H'_k\}_{k=0}^{K-1}$ **对于** $k = 0, 1, \ldots, K-1$，**执行** $\hat{}G^{\pi_k} =$ **评价器** $(\pi_k, N_k, H_k, N'_k, H'_k)$ **对于** $s \in \mathcal{S}$，**执行** $\pi_{k+1}(s, \cdot) = \mathbf{argmax}_{p \in \mathcal{M}_\alpha(\mathcal{A})} \left\{ \eta_k \sum_{a \in \mathcal{A}} \hat{} G^{\pi_k}(s, a) p(a) - D\big(p(\cdot) \,\|\, \pi_k(\cdot \mid s)\big) \right\}$ **结束循环 结束循环 输出：** π_K

定理 13 考虑一个多链马尔可夫决策过程。设 $\epsilon > 0, \delta > 0$。对于任意 $\pi, \pi_0 \in \Pi_\alpha$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，以概率 $1-\delta$，恒定步长为 $\eta > 0$、且满足 $K = 2\left({}^{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}\!/_{\eta} + C_\alpha (J_\mu^{\pi_\alpha} - J_\mu^{\pi_0}) \right)/\epsilon$ 的随机 α-截断策略镜像上升迭代序列满足 $J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \le \epsilon$，其样本复杂度为

$$\widetilde{\mathcal{O}}\left(\frac{t_{\text{tar}}^3 \, \|Q^\pi\|_\infty^4 \, R^9 C_\alpha^6 \left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta} \right)^6 B_\alpha^6 |\mathcal{S}||\mathcal{A}|}{\epsilon^{12}} \right)$$

并且对于自适应步长 $\eta_{k+1}(C_\alpha - 1) \ge \eta_k C_\alpha > 0$ 和 $K = \text{对数}\left({}^{2(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0} + D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)\eta_0(C_\alpha - 1))}\!/_{\epsilon} \right) \log(C_\alpha/(C_\alpha - 1))$ 满足 $J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \le \epsilon$，其样本复杂度为

$$\tilde{O}\left(\frac{t_{\text{tar}}^3 \, \|Q^\pi\|_\infty^4 \, R^3 C_\alpha^6 B_\alpha^6 |\mathcal{S}||\mathcal{A}|}{\epsilon^6} \right)$$

~其中 $\mathcal{O}$ 忽略所有对数因子，$\|Q^\pi\|_\infty = \text{最大值}_{0 \le k \le K-1} \|Q^{\pi_k}\|_\infty$，以及 $t_{\text{tar}} = \text{最大值}_{0 \le k \le K-1} t^{\pi_k}$目标。

据我们所知，定理 13首次给出了平均奖励多链马尔可夫决策过程的策略梯度方法的样本复杂度结果。最后，我们简要提及，附录 E提供了针对弱可通信马尔可夫决策过程的精细分析。

5. 结论

在本文中，我们基于状态分类的不变性，利用常返与暂态访问测度的概念，首次分析了适用于平均奖励多链马尔可夫决策过程的策略梯度方法。我们的结果和证明技巧为未来研究平均奖励多链马尔可夫决策过程中的策略梯度算法铺平了道路。

一个未来方向是通过引入方差缩减技术（Li等人，2024年；Lee等人，2025）来进一步改进样本复杂度，并通过量化 α 与 ϵ 在我们定理中的依赖关系来完整刻画样本复杂度。利用我们的方法体系为平均奖励多链马尔可夫决策过程开发信赖域方法（Schulman等人，2015；Zhang和Ross，2021）也将是一个有趣的方向。

参考文献

Alekh Agarwal、Sham M Kakade、Jason D Lee和Gaurav Mahajan。策略梯度方法理论：最优性、近似与分布偏移。机器学习研究杂志，22(98):1–76，2021。Qinbo Bai、Washim Uddin Mondal和Vaneet Aggarwal。无限时域平均奖励马尔可夫决策过程的策略梯度算法遗憾分析。国际人工智能会议，2024年。Jonathan Baxter和Peter L Bartlett。直接基于梯度的强化学习。*IEEE*电路与系统国际研讨会，3:271–274，2000。Abraham Berman和Robert J Plemmons。数学科学中的非负矩阵。工业与应用数学学会，1994年。Jalaj Bhandari和Daniel Russo。策略梯度方法的全局最优性保证。运筹学，72(5):1906–1927，2024年。David Blackwell。离散动态规划。数学统计年刊，33:719–726，1962年。Victor Boone和Zihan Zhang。在平均奖励马尔可夫决策过程中实现可处理的极小极大最优遗憾。神经信息处理系统大会，2024年。Mario Bravo和Roberto Cominetti。非扩张映射的随机不动点迭代：收敛与误差界。*SIAM*控制与优化杂志，62(1):191–219，2024年。Pierre Br´emaud。马尔可夫链：吉布斯场、蒙特卡罗模拟与排队系统。施普林格科学与商业媒体，第2版，2013。Apostolos N Burnetas和Michael N Katehakis。马尔可夫决策过程的最优自适应策略。《运筹学数学》，22(1):222–255，1997年。Shicong Cen、Chen Cheng、Yuxin Chen、Yuting Wei和Yuejie Chi。带熵正则化的自然策略梯度方法的快速全局收敛。运筹学，70(4):2563–2578，2022年。Zaiwei Chen。具有自适应步长的平均奖励Q学习的非渐近保证。神经信息处理系统大会，2025。Vektor Dewanto、George Dunn、Ali Eshragh、Marcus Gallagher和Fred Roosta。平均奖励无模型强化学习：系统综述与文献图谱。*arXiv:2010.08920*，2020年。Eyal Even-Dar、Sham M Kakade和Yishay Mansour。在线马尔可夫决策过程。《运筹学数学》，34(3):726–736，2009年。Yoav Freund和Robert E Schapire。在线学习的决策理论推广及其在提升中的应用。计算机与系统科学杂志，55(1):119–139，1997年。

Swetha Ganesh 和 Vaneet Aggarwal。通过演员-评论家方法分析平均奖励单链马尔可夫决策过程的遗憾。*arXiv*预印本 *arXiv:2505.19986*，2025。Swetha Ganesh、Washim Uddin Mondal 和 Vaneet Aggarwal。无限时域平均奖励马尔可夫决策过程中采用新型策略梯度方法实现阶数最优遗憾。国际人工智能与统计学会议，2025。Mark Herbster 和 Manfred K Warmuth。跟踪最佳线性预测器。机器学习研究杂志，1：281–309，2001。Ronald A Howard。《动态规划与马尔可夫过程》。John Wiley and Sons，1960。Thomas Jaksch、Ronald Ortner 和 Peter Auer。强化学习的近最优遗憾界。机器学习研究杂志，11(51)：1563–1600，2010。Ying Jin、Ramki Gummadi、Zhengyuan Z hou 和 Jose Blanchet。用于平均奖励强化学习的可行Q学习。国际人工智能与统计学会议，2024。Yujia Jin 和 Aaron Sidford。迈向平均奖励马尔可夫决策过程样本复杂度的紧确界。*International Conference on MachineLearning*，2021。Sham M Kakade。自然策略梯度。神经信息处理系统大会，2001。Michael Kearns 和 Satinder Singh。Q学习与间接算法的有限样本收敛速率。神经信息处理系统大会，1998。Vijay Konda 和 John Tsitsiklis。演员-评论家算法。神经信息处理系统大会，1999。Navdeep Kumar、Yashaswini Murthy、Itai Shufaro、Kfir Y Levy、R Srikant 和 Shie Mannor。平均奖励马尔可夫决策过程中策略梯度的全局收敛性。国际学习表征会议，2024。John M Lee。光滑流形。载于《光滑流形导论》，第1–29页。施普林格，2003。Jongmin Lee 和 Ernest K Ryu。平均奖励马尔可夫决策过程值迭代的最优非渐近速率。国际学习表征会议，2025a。Jongmin Lee 和 Ernest K Ryu。无折扣总奖励马尔可夫决策过程的策略梯度算法为何有效。*arXiv*预印本 *arXiv:2510.18340*，2025年b。Jongmin Lee、Mario Bravo 和 Roberto Cominetti。通过锚定实现马尔可夫决策过程的近最优样本复杂度。*International Conferenceon MachineLearning*，2025。David A Levin 和 Yuval Peres。马尔可夫链与混合时间。美国数学学会，2017。Tianjiao Li、Feiyang Wu 和 Guanghui Lan。平均奖励马尔可夫决策过程的随机一阶方法。《运筹学数学》，50(4)：3125–3160，2024。

Sridhar Mahadevan。平均奖励强化学习：基础、算法与实证结果。*Machine Learning*，22(1)：159–195，1996。Peter Marbach 和 John N Tsitsiklis。基于仿真的马尔可夫奖励过程优化。*IEEE Transactions on Automatic Control*，46(2)：191–209，2001。Ji ncheng Mei、Chenjun Xiao、Csaba Szepesvari 和 Dale Schuurmans。softmax策略梯度方法的全局收敛速率。*InternationalConferenceonMachineLearning*，2020。Yashaswini Murthy 和 R Srikant。自然策略梯度与镜像下降类策略方法在平均奖励马尔可夫决策过程中的收敛性。*IEEE Conference on Decision and Control*，2023。Martin L. Puterman。马尔可夫决策过程：离散随机动态规划。John Wiley and Sons，第2版，2014。Gareth O Roberts 和 Jeffrey S Rosenthal。移位耦合与遍历平均的收敛速率。*Stochastic Models*，13(1)：147–165，1997。John Schulman、Sergey Levine、Pieter Abbeel、Michael Jordan 和 Philipp Moritz。信赖域策略优化。*International Conferenceon MachineLearning*，2015。John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。*arXiv*预印本 *arXiv:1707.06347*，2017。Paul J Schweitzer。扰动理论与有限马尔可夫链。*Journal of Applied Probability*，5 (2)：401–413，1968。Lior Shani、Yonathan Efroni 和 Shie Mannor。自适应信赖域策略优化：正则化马尔可夫决策过程的全局收敛与更快速率。*AAAIConference on Artificial Intelligence*，2020。Richard S Sutton、David McAllester、Satinder Singh 和 Yishay Mansour。结合函数逼近的强化学习策略梯度方法。神经信息处理系统大会，1999。Adrienne Tuynman、R´emy Degenne 和 Emilie Kaufmann。在缺乏先验知识的情况下寻找平均奖励马尔可夫决策过程中的优良策略。神经信息处理系统大会，2024。Yi Wan、Huizhen Yu 和 Richard S Sutton。弱可通信马尔可夫决策过程中的平均奖励Q学习收敛性。*arXiv*预印本 *arXiv:2408.16262*，2024。Mengdi Wang。原始-对偶 π 学习：遍历马尔可夫决策问题的样本复杂度与次线性运行时间。*arXiv:1710.06100*，2017。Weiran Wang 和 Miguel A Carreira-Perpin´an。投影到概率单纯形：一种高效算法及其简明证明，并附应用。*arXiv*预印本 *arXiv:1309.1541*，2013。

Chen-Yu Wei、Mehdi Jafarnia Jahromi、Haipeng Luo、Hiteshi Sharma 和 Rahul Jain。无限时域平均奖励马尔可夫决策过程中的无模型强化学习。*International Conference on Machine Learning*，2020年。Ronald J Williams。用于联结主义强化学习的简单统计梯度跟随算法。*Machine Learning*，8(3):229–256，1992年。Lin Xiao。策略梯度方法的收敛速率。机器学习研究杂志，23(282):1–36，2022年。Yiming Zhang 和 Keith W Ross。针对平均奖励准则的同策略深度强化学习。*International Conference on MachineLearning*，2021。Zihan Zhang 和 Xiangyang Ji。通过评估最优偏差函数实现强化学习的遗憾最小化。神经信息处理系统大会，2019年。Zihan Zhang 和 Qiaomin Xie。针对平均奖励马尔可夫决策过程的改进型无模型强化学习。学习理论会议，2023年。Matthew Zurek 和 Yudong Chen。针对弱连通和一般平均奖励马尔可夫决策过程的基于跨度的最优样本复杂度。神经信息处理系统大会，2024年。

附录 A. 预备知识

若函数 h 满足本质光滑性，则称其为本质光滑函数；若 h 可微，且对于每个收敛到 $\mathrm{dom}\, h$边界点的序列 $\{x_k\}$，都有 $\|\nabla h(x_k)\| \to \infty$。若函数 h 是勒让德函数，则称其为勒让德函数；这意味着它在 $\mathrm{dom}\, h$的（相对）内部本质光滑且严格凸，

事实 3（三点下降引理）（*Xiao，2022年*，引理 6）假设 $\mathcal{C} \subset \mathbb{R}^n$ 是一个闭凸集，$\phi:\ \mathcal{C} \to \mathbb{R}$ 是一个适当闭凸函数，$D(\cdot,\ \cdot)$ 是由一个 h 勒让德型函数生成的 ***Bregman*** 散度，且 $\mathrm{rint}\, \mathrm{dom}\, h \cap \mathcal{C} \neq \varnothing$。对于任意 $x \in \mathrm{rint}\, \mathrm{dom}\, h$，令

$$x^+ = \underset{u \in \mathcal{C}}{\mathrm{argmin}} \left\{ \phi(u) + D(u, x) \right\}.$$

于是，$x^+ \in \mathrm{rint}\, \mathrm{dom}\, h \cap \mathcal{C}$，且对于任意 $u \in \mathcal{C}$，

$$\phi(x^+) + D(x^+, x) \ \leq\ \phi(u) + D(u, x) - D(u, x^+).$$

在我们的设定中，$\mathcal{C} = \mathcal{M}_\alpha(\mathcal{A})$、$\phi(p) = -\eta_k \langle G^{\pi_k}(\cdot,\ s)\ p(\cdot)\ \rangle$，且 h 是负熵函数，它也属于勒让德型，并满足 $\mathrm{rint}\, \mathrm{dom}\, h \cap \mathcal{C} = \mathcal{M}_\alpha(\mathcal{A})$。因此，如果我们从 $\mathcal{M}_\alpha(\mathcal{A})$中的一个初始点开始，那么每次迭代都会保持在 $\mathcal{M}_\alpha(\mathcal{A})$中。

事实 4（霍夫丁不等式）设 $X_1, \ldots, X_n$ 是相互独立的随机变量，满足对所有 i，$a_i \leq X_i \leq b_i$。那么

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i \right] \right| \geq \epsilon \right) \leq 2exp\left(- \frac{2n^2\epsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

事实 5（目标时间引理）（*Roberts 和 Rosenthal，1997年*，推论 3）对于 $\pi \in \Pi_+$ 和 $\forall s \in \mathcal{R}_i$，

$$\left\| \frac{1}{k} \sum_{j=1}^{k} \left(R_i^\pi\right)^j(\cdot, s) \ -\ g_i^\pi(\cdot) \right\|_1 \ \leq\ \frac{2t_{\mathrm{tar}}^{\pi,i}}{k},$$

其中 g_i^π 是 R_i^π 的平稳分布。

附录 B. 第 2 节中省略的证明

B.1. 引理 3 的证明

我们先证明以下引理。

引理 14 对于任意 $\pi \in \Pi_+$，$a, a' \in \mathcal{A}$ 与 $s, s' \in \mathcal{R}_i$，$K^\pi(s, a) = K^\pi(s', a')$ 以及 $J^\pi(s) = J^\pi(s')$。

证明 对于 $s \in \mathcal{R}_i$ 与 $s' \notin \mathcal{R}_i$，$P(s' \mid s, a) = 0$。对于任意 $s, s' \in \mathcal{R}_i$，$J^\pi(s) = J^\pi(s')$，因为$R_{i,\star}^\pi = \mathbf{1} g_i^\top$，其中 g_i是 R_i^π的平稳分布。因此，$K^\pi(s, a) = PJ^\pi(s, a) = \sum_{s'' \in \mathcal{R}_i} P(s'' \mid s, a) J^\pi(s'') = J^\pi(s) = J^\pi(s')$对于任意 $a \in \mathcal{A}$。 ■

现在我们证明引理 3。

证明首先，我们有 $J^\pi - J^{\pi'} = P_\star^\pi (J^\pi - J^{\pi'}) + (P_\star^\pi - I) J^{\pi'}$，因为 $P_\star^\pi P_\star^\pi = P_\star^\pi$。

对于第一项,

$$\begin{aligned}
P_\star^\pi(J^\pi - J^{\pi'}) &= P_\star^\pi(r^\pi - J^{\pi'}) \\
&= P_\star^\pi(r^\pi + V^{\pi'} - P^{\pi'}V^{\pi'} - r^{\pi'}) \\
&= P_\star^\pi(r^\pi + P^\pi V^{\pi'} - P^{\pi'}V^{\pi'} - r^{\pi'}) \\
&= P_\star^\pi(\Theta_\pi - \Theta_{\pi'})(r + PV^{\pi'}) \\
&= P_\star^\pi(\Theta_\pi - \Theta_{\pi'})(K^{\pi'} + Q^{\pi'})
\end{aligned}$$

其中，第一个和第三个等式来自极限矩阵的性质， $P_\star^\pi = P_\star^\pi P_\star^\pi = P_\star^\pi P^\pi$；第二个和最后一个等式来自贝尔曼方程；第四个等式则是因为 $\Theta_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ 是满足 $\Theta_\pi P = P^\pi$ 和 $\Theta_\pi r = r^\pi$的策略 π 的矩阵形式。于是，对于给定的具有全支撑的 μ 和 $s \in \mathcal{R}$，我们有

$$\begin{aligned}
\mu^\top P_\star^\pi(J^\pi - J^{\pi'}) &= \mu^\top P_\star^\pi(\Theta_\pi - \Theta_{\pi'})(K^{\pi'} + Q^{\pi'}) \\
&= \sum_{s\in\mathcal{S}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} (\pi(a\,|\,s) - \pi'(a\,|\,s))(K^{\pi'}(s,a) + Q^{\pi'}(s,a)) \\
&= \sum_{s\in\mathcal{R}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} (\pi(a\,|\,s) - \pi'(a\,|\,s))(K^{\pi'}(s,a) + Q^{\pi'}(s,a)) \\
&= \sum_{s\in\mathcal{R}} d_\mu^\pi(s) \sum_{a\in\mathcal{A}} (\pi(a\,|\,s) - \pi'(a\,|\,s))Q^{\pi'}(s,a)
\end{aligned}$$

其中第三个等式来自 $d_\mu^\pi(s) = 0$ 对所有 $s \in \mathcal{T}$成立，而最后一个等式来自引理14。

对于第二项，若 $s \in \mathcal{R}_i$, $P_\star^\pi J^{\pi'}(s) = \sum_{s'\in\mathcal{R}_i} g_i^\pi(s') J^{\pi'}(s') = J^{\pi'}(s)$，则成立，因为 R_i^π, $_\star = 1(g_i^\pi)^\top$，其中 g_i^π是平稳分布，且对于 $s', s'' \in \mathcal{R}_i$， $J^{\pi'}(s') = J^{\pi'}(s'')$；由引理 14 和事实 1 可知， π, π'属于同一个常返类。否则，若 $s \in \mathcal{T}$，定义 $J_i^\pi \in \mathbb{R}^{|\mathcal{R}_i|}$和 $J_{m+1}^\pi \in \mathbb{R}^{|\mathcal{T}|}$，使得对于 $1 \le i \le m$， $J^\pi(s) = J_i^\pi(s)$，且对于 $s \in \mathcal{R}_i$成立；对于 $s \in \mathcal{T}$， $J^\pi(s) = J_{m+1}^\pi(s)$。由于 $S_i^\pi{}_{,\star} = (I - T^\pi)^{-1} S_i^\pi R_i^\pi{}_{,\star}$，我们有

$$\begin{aligned}
P_\star^\pi J^{\pi'}(s) &= \sum_{i=1}^{m} (I - T^\pi)^{-1} S_i^\pi J_i^{\pi'}(s) \\
&= (I - \bar{T}^\pi)^{-1}(P^\pi - \bar{T}^\pi) J^{\pi'}(s)
\end{aligned}$$

其中第一个等 式来自引理 14、事实 1，以及 $P_\star^\pi$和 P^π的规范形。这意味着

$$\begin{aligned}
(P_\star^\pi - I) J^{\pi'}(s) &= (I - \bar{T}^\pi)^{-1}(P^\pi - I) J^{\pi'}(s) \\
&= (I - \bar{T}^\pi)^{-1}(P^\pi - P^{\pi'}) J^{\pi'}(s) \\
&= (I - \bar{T}^\pi)^{-1}(\Theta_\pi - \Theta_{\pi'}) K^{\pi'}(s)
\end{aligned}$$

对于 $s \in \mathcal{T}$，第二个等号来自贝尔曼方程。将 μ 相乘后，即可得证。

■

B.2. 定理5的证明

证明 首先，由于暂态类和常返类的不变性，只需考虑 $\{R^{\pi}_{\star,i}\}_{i=1}^{m}, \{S^{\pi}_{\star,i}\}_{i=1}^{m}$ 的可微性，因为 $P^{\pi}_{\star}$ 的其他元素始终为零。根据（Marbach 和 Tsitsiklis，2001年，引理 1）的证明，$R^{\pi}_{\star,i}$ 的可微性得以保证。（作为一项技术细节，Marbach 和 Tsitsiklis （2001年）中 $R^{\pi}_{\star,i}$ 的可微性是在非周期性条件下证明的，但由于平稳分布的唯一性，该证明也适用于单链马尔可夫决策过程（Levin 和 Peres，2017，命题 1.29）。）而根据（Lee 和 Ryu，2025年b，定理 4.7）的证明，$(I - T^{\pi})^{-1}$ 的可微性也得以保证。这些结果还表明，$S^{\pi}_{i,\star} = (I - T^{\pi})^{-1} S^{\pi}_{i} R^{\pi}_{i,\star}$ 以及 $P^{\pi}_{\star}$ 也都是可微的。

对于流形 S，我们将 S 的切空间记为 $\mathrm{T}S$ (Lee, 2003)。考虑 Θ 一个可微流形，以及微分 $\nabla_{\theta}\pi_{\theta} : \mathrm{T}\Theta \to \mathrm{T}\Pi_{+}$，其中 $\pi_{\theta} : \Theta \to \Pi_{+}$，且 $\mathrm{T}\Pi_{+} = \{u \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mid \sum_{a\in\mathcal{A}} u(a \mid s) = 0 \, \forall s \in \mathcal{S}\}$。（注意，这里 Π_{+} 是嵌入欧几里得空间的单纯形。）根据前面的论证，由于 $J^{\pi}_{\mu} : \Pi_{+} \to \mathbb{R}$ 是光滑函数，我们也可以考虑微分$\nabla_{\pi} J^{\pi}_{\mu} : \mathrm{T}\Pi_{+} \to \mathrm{T}\mathbb{R}$，其中 $\mathrm{T}\mathbb{R} = \mathbb{R}$。

根据引理 3，对于任意 $u \in \mathbf{T}\Pi_{+}$ 和 $\epsilon > 0$ 满足 $\pi + \epsilon u \in \Pi_{+}$，有

$$J^{\pi+\epsilon u}_{\mu} - J^{\pi}_{\mu} = \epsilon \sum_{\substack{s\in\mathcal{R}\\ a\in\mathcal{A}}} d^{\pi+\epsilon u}_{\mu}(s) u(a \mid s) Q^{\pi}(s,a) + \epsilon \sum_{\substack{s\in\mathcal{T}\\ a\in\mathcal{A}}} \delta^{\pi+\epsilon u}_{\mu}(s) u(a \mid s) K^{\pi}(s,a)$$

由 d^{π}_{μ} 和 δ^{π}_{μ} 在 Π_{+} 上的连续性，两边除以 ϵ 并令 $\epsilon \to 0$，可将微分形式化为

$$\nabla_{\pi} J^{\pi}_{\mu}(s,a) = d^{\pi}_{\mu}(s) Q^{\pi}(s,a) + \delta^{\pi}_{\mu}(s) K^{\pi}(s,a).$$

然后，根据链式法则，

$$\nabla_{\theta} J^{\pi}_{\mu} = \sum_{\substack{s\in\mathcal{R}\\ a\in\mathcal{A}}} d^{\pi}_{\mu}(s) \nabla_{\theta}\pi(a \mid s) Q^{\pi}(s,a) + \sum_{\substack{s\in\mathcal{T}\\ a\in\mathcal{A}}} \delta^{\pi}_{\mu}(s) \nabla_{\theta}\pi(a \mid s) K^{\pi}(s,a).$$

■

我们简要指出，引理 2可直接由定理5得出。

附录 C. 章节3中省略的证明

C.1. 引理7的证明

证明 正如我们在定理 5 的证明中所示，d^{π} 和 δ^{π} 在 Π_{+} 上连续。对于每个 i，$\mathcal{R}_i$ 在 $\pi \in \Pi_{\alpha}$ 上是单一常返类，因此 R^{π}_{i}的平稳分布 g^{π}_{i} 始终为正（Levin 和 Peres，2017，命题 1.19）。由连续性可知 $\min_{s\in\mathcal{R},\ \pi\in\Pi_{\alpha}} d^{\pi}_{\mu}(s) > 0$，且 $\|d^{\pi}_{\mu}\|_1 = 1$。另一方面，对于 $s \in \mathcal{T}$，根据定义，$\min_{\pi\in\Pi_{\alpha}} \delta^{\pi}_{\mu}(s) \geq \mu(s)$，并由连续性可知 $\max_{\pi\in\Pi_{\alpha}} \delta^{\pi}_{\mu}(s) < \infty$。 ■

C.2. 引理8的证明

我们证明了引理的一个更详细版本8。

引理15 对于 $\mu \in \mathcal{M}_+(\mathcal{S})$， α-策略镜像上升生成一列策略 $\{\pi_k\}_{k=1}^{\infty}$，满足

$$\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s)) \le 0, \qquad \forall\, s\in\mathcal{S},$$

and

$$J_\mu^{\pi_k} \le J_\mu^{\pi_{k+1}}.$$

证明 应用事实 3 并结合 $\mathcal{C}=\mathcal{M}_\alpha(\mathcal{A})$， $\phi(p)=-\eta_k\sum_{a\in\mathcal{A}} G^{\pi_k}(a,\ s)p(a)$，可得

$$\begin{aligned}
&\eta_k\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)p(a) + D\big(\pi_{k+1}(\cdot\,|\,s),\ \pi_k(\cdot\,|\,s)\big)\\
&\le \eta_k\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)\pi_{k+1}(a\,|\,s) + D\big(p,\pi_k(\cdot\,|\,s)\big) - D\big(p,\pi_{k+1}(\cdot\,|\,s)\big)
\end{aligned}$$

对于任意 $p\in\mathcal{M}_\alpha(\mathcal{A})$。整理各项并将两边同时除以 η_k，得到

$$\begin{aligned}
&\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(p(a)-\pi_{k+1}(a\,|\,s)) + \frac{1}{\eta_k}D\big(\pi_{k+1}(\cdot\,|\,s),\pi_k(\cdot\,|\,s)\big)\\
&\le\ \frac{1}{\eta_k}D\big(p,\pi_k(\cdot\,|\,s)\big) - \frac{1}{\eta_k}D\big(p,\pi_{k+1}(\cdot\,|\,s)\big). \qquad (*)
\end{aligned}$$

将 $p=\pi_k(\cdot\,|\,s)$ 代入前一个不等式，得到

$$\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s)) \le -\frac{1}{\eta_k}D\big(\pi_{k+1}(\cdot\,|\,s),\pi_k(\cdot\,|\,s)\big)-\frac{1}{\eta_k}D\big(\pi_k(\cdot\,|\,s),\pi_{k+1}(\cdot\,|\,s)\big).$$

然后，第一个结论由 Bregman散度的非负性得出；第二个结论则将 $\rho_\mu^{\pi_{k+1}}(\cdot)$ 乘以等式两边，利用性能差引理得出。 ■

C.3. 定理9 的证明

证明 考虑前面的不等式（$*$）。令 $p=\pi_\alpha(\cdot\,|\,s)\in\mathcal{M}_\alpha$，并在内积中加上再减去 $\pi_k(\cdot\,|\,s)$。于是我们有

$$\begin{aligned}
&\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s)) + \sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s)-\pi_k(a\,|\,s))\\
&\le\ \frac{1}{\eta_k}D\big(\pi_\alpha(\cdot\,|\,s),\pi_k(\cdot\,|\,s)\big) - \frac{1}{\eta_k}D\big(\pi_\alpha(\cdot\,|\,s),\pi_{k+1}(\cdot\,|\,s)\big).
\end{aligned}$$

这意味着

$$\begin{aligned}
&\sum_{s\in\mathcal{S}}\rho_\mu^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s)) + \sum_{s\in\mathcal{S}}\rho_\mu^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s)-\pi_k(a\,|\,s))\\
&\le\ \frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k) - \frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1}).
\end{aligned}$$

对于第一项，

$$\begin{aligned}&\sum_{s\in\mathcal{S}}\rho_\mu^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}}G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s))\\&=\sum_{s\in\mathcal{S}}\rho_\mu^{\pi_{k+1}}(s)\frac{\rho_\mu^{\pi_\alpha}(s)}{\rho_\mu^{\pi_{k+1}}(s)}\sum_{a\in\mathcal{A}}G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s))\\&\geq C_\alpha(J_\mu^{\pi_k}-J_\mu^{\pi_{k+1}}),\end{aligned}$$

其中，最后一个不等式由引理15 和 3得出。

对于第二项，由引理3可得，

$$\sum_{s\in\mathcal{S}}\rho_\mu^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}}G^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s)-\pi_k(a\,|\,s))=J_\mu^{\pi_\alpha}-J_\mu^{\pi_k}.$$

因此，我们有

$$J_\mu^{\pi_\alpha}-J_\mu^{\pi_k}\;\leq\;\frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k)-\frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1})+C_\alpha(J_\mu^{\pi_{k+1}}-J_\mu^{\pi_k}).$$

对所有 $k\geq 0$设定 $\eta_k=\eta$ 并对 k 求和，得到

$$\sum_{i=0}^{k}\left(J_\mu^{\pi_\alpha}-J_\mu^{\pi_i}\right)\;\leq\;\frac{1}{\eta}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_0)-\frac{1}{\eta}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1})+C_\alpha(J_\mu^{\pi_{k+1}}-J_\mu^{\pi_0}).$$

由于 $J_\mu^{\pi_k}$根据引理 8 是非递减的，且 Bregman散度非负，因此得出

$$J_\mu^{\pi_\alpha}-J_\mu^{\pi_k}\;\leq\;\frac{1}{k+1}\left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_0)}{\eta}+C_\alpha(J_\mu^{\pi_\alpha}-J_\mu^{\pi_0})\right).$$

■

C.4. 定理的证明10

定义 $U_k^{\pi_\alpha}=J_\mu^{\pi_\alpha}-J_\mu^{\pi_k}$，我们首先证明以下关键引理。

引理 16 对于 $\pi_0\in\Pi_\alpha$ 和给定的 $\mu\in\mathcal{M}_+(\mathcal{S})$，采用步长为 $\eta_k>0$的 α-裁剪策略镜像上升会生成一系列策略 $\{\pi_k\}_{k=1}^{\infty}$，满足：

$$C_\alpha\left(U_{k+1}^{\pi_\alpha}-U_k^{\pi_\alpha}\right)+U_k^{\pi_\alpha}\;\leq\;\frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k)-\frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1})$$

证明 在定理的前一个证明中 9，我们证明了

$$\begin{aligned}&\sum_{s\in\mathcal{S}}\rho_\mu^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}}G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s))+\sum_{s\in\mathcal{S}}\rho_\mu^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}}G^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s)-\pi_k(a\,|\,s))\\&\leq\;\frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k)-\frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1}).\end{aligned}$$

对于第一项

$$\sum_{s\in\mathcal{S}}\rho_{\mu}^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}}G^{\pi_k}(a,s)(\pi_k(a\,|\,s)-\pi_{k+1}(a\,|\,s))$$
$$\geq C_\alpha(J_\mu^{\pi_k}-J_\mu^{\pi_{k+1}})$$

其中该不等式来自引理15 和3。

对于第二项，由引理3,

$$\sum_{s\in\mathcal{S}}\rho_{\mu}^{\pi_\alpha}(s)\sum_{a\in\mathcal{A}}G^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s)-\pi_k(a\,|\,s))=J_\mu^{\pi_\alpha}-J_\mu^{\pi_k}.$$

代入后，我们得到所需结果。

■

我们现在可以证明定理10

证明由引理16,

$$C_\alpha\left(U_{k+1}^{\pi_\alpha}-U_k^{\pi_\alpha}\right)+U_k^{\pi_\alpha}\ \leq\ \frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k)-\frac{1}{\eta_k}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1}).$$

将两边同除以 C_α 并整理各项，我们得到

$$U_{k+1}^{\pi_\alpha}+\frac{1}{\eta_kC_\alpha}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1})\ \leq\ \left(1-\frac{1}{C_\alpha}\right)\left(U_k^{\pi_\alpha}+\frac{1}{\eta_k(C_\alpha-1)}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k)\right).$$

由于步长满足条件 $\eta_{k+1}(C_\alpha-1)\ \geq\eta_kC_\alpha>0$，我们有

$$U_{k+1}^{\pi_\alpha}+\frac{1}{\eta_{k+1}(C_\alpha-1)}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_{k+1})\ \leq\ \left(1-\frac{1}{C_\alpha}\right)\left(U_k^{\pi_\alpha}+\frac{1}{\eta_k(C_\alpha-1)}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k)\right).$$

因此，通过递推,

$$U_k^{\pi_\alpha}+\frac{1}{\eta_k(C_\alpha-1)}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_k)\ \leq\ \left(1-\frac{1}{C_\alpha}\right)^k\left(U_0^{\pi_\alpha}+\frac{1}{\eta_0(C_\alpha-1)}D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_0)\right).$$

■

附录 D. 第 4 章节中省略的证明4

D.1. 定理11

定义

$$V^{k+1}=P^\pi V^k+r^\pi\qquad k=0,1,\ldots$$

其中 $V^0=0$。

事实 6（经典结果，（Puterman，2014，定理9.4.1））考虑一个一般的（多链）马尔可夫决策过程。那么，对于 $k\geq1$，序列$\{V^k\}_{k=0}^\infty$ 呈现出如下速率

$$\left\|\frac{V^k}{k}-J^\pi\right\|_\infty\leq\frac{2\left\|V^\pi\right\|_\infty}{k}.$$

利用事实 6，我们首先研究对 K^π的近似所需的样本复杂度。

引理 17 设 $\epsilon > 0$ 和 $\delta > 0$。对于给定的 $\pi \in \Pi$，以 $1-\delta$ 的概率，评价器中的 K^π 满足 $\left\| \hat{K}^\pi - K^\pi \right\|_\infty \leq \epsilon$，其样本复杂度为

$$NH|\mathcal{S}||\mathcal{A}| = \mathcal{O}\left(\frac{(\|V^\pi\|_\infty + R)R^2|\mathcal{S}||\mathcal{A}|}{\epsilon^3} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta} \right) \right).$$

证明 由于

$$\hat{K}(s_0, a_0) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{H+1} \sum_{i=0}^{H} r_j(s_i, a_i),$$

and

$$\mathbb{E}_\pi \left[\frac{1}{N}\sum_{j=1}^{N} \frac{1}{H+1} \sum_{i=0}^{H} r_j(s_i, a_i) \right] = \mathbb{E}_\pi \left[\frac{1}{H+1} \sum_{i=0}^{H} r_j(s_i, a_i) \right]$$

根据生成模型样本的独立性。此外，我们有

$$\begin{aligned}
\left| \mathbb{E}_\pi \left[\frac{1}{H+1} \sum_{i=0}^{H} r_j(s_i, a_i) \right] - K^\pi(s_0, a_0) \right| &= \left| \frac{1}{H+1}(PV^H + r)(s_0, a_0) - PJ^\pi(s_0, a_0) \right| \\
&= \left| P\left(\frac{V^H}{H+1} - J^\pi \right)(s_0, a_0) + \frac{r(s_0, a_0)}{H+1} \right| \\
&= \left| \frac{H}{H+1} P\left(\frac{V^H}{H} - J^\pi \right)(s_0, a_0) + \frac{(r-K)(s_0, a_0)}{H+1} \right| \\
&\leq \frac{2(\|V^\pi\|_\infty + R)}{H+1}
\end{aligned}$$

其中，第一个等式来自贝尔曼方程，最后一个不等式来自事实 6和 $\|K\|_\infty \leq R$。此外，根据霍夫丁不等式，当 $N = \frac{2R^2}{(\epsilon')^2} \log(\frac{2}{\delta})$ 时，我们得到——

$$\text{Prob}\left(\left| \hat{K}^\pi(s_0, a_0) - \mathbb{E}_\pi \left[\frac{1}{N}\sum_{j=1}^{N} \frac{1}{H+1} \sum_{i=0}^{H} r_j(s_i, a_i) \right] \right| \leq \epsilon' \right) \geq 1 - \delta.$$

并且对 $(s_0, a_0) \in \mathcal{S} \times \mathcal{A}$ 使用并集界，可得 $\left\| \hat{K}^\pi - \mathbb{E}\left[\hat{K}^\pi \right] \right\|_\infty < \epsilon'$。设 $H = \frac{4}{(\|V^\pi\|_\infty + R)\epsilon}$和 $\epsilon' = \epsilon/2$。于是，根据三角不等式 $\left\| K^\pi - \hat{K}^\pi \right\|_\infty \leq \left\| K^\pi - \mathbb{E}\left[\hat{K}^\pi \right] \right\|_\infty + \left\| \mathbb{E}\left[\hat{K}^\pi \right] - \hat{K}^\pi \right\|_\infty \leq \epsilon$，其样本复杂度为

$$NH|\mathcal{S}||\mathcal{A}| = \frac{32|\mathcal{S}||\mathcal{A}|(\|V^\pi\|_\infty + R)R^2}{\epsilon^3} \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta} \right)$$

■

现在，我们建立以下引理，用于使用评论家方法评估 Q^π。

引理18 考虑一个一般的（多链）马尔可夫决策过程。那么，对于 $k \geq 1$的任意常返状态 $s \in \mathcal{R}$，$\{V^k\}_{k=0}^{\infty}$ 呈现出如下速率

$$\left| \frac{1}{k} \sum_{i=1}^{k} (V^i - iJ^\pi)(s) - V^\pi(s) \right| \leq \frac{2t_{\text{tar}} \left\| V^\pi \right\|_\infty}{k}.$$

证明首先,

$$\begin{aligned} V^k &= \sum_{i=0}^{k-1} (P^\pi)^i r^\pi \\ &= \sum_{i=0}^{k-1} (P^\pi)^i (J^\pi + (I - P^\pi)V^\pi) \\ &= kJ^\pi + V^\pi - (P^\pi)^k V^\pi. \end{aligned}$$

Thus

$$\begin{aligned} \frac{1}{k} \sum_{i=1}^{k} (V^i - iJ^\pi) &= V^\pi - \frac{1}{k} \sum_{i=0}^{k-1} (P^\pi)^i V^\pi \\ &= V^\pi + \left(P_\star^\pi - \frac{1}{k} \sum_{i=0}^{k-1} (P^\pi)^i \right) V^\pi \end{aligned}$$

其中，第一个等号来自 $V^0 = 0$，最后一个等号来自 $P_\star^\pi V^\pi = 0$ （Puterman，2014，A.5节）。然后，根据随机矩阵的规范形式和事实 5，对于 $s \in \mathcal{R}$,

$$\left\| P_\star^\pi(s \,|\, \cdot) - \frac{1}{k} \sum_{i=0}^{k-1} (P^\pi)^i (s \,|\, \cdot) \right\|_1 \leq \frac{2t_{\text{tar}}^\pi}{k} \qquad \forall s \in \mathcal{R}$$

其中 $t_{\text{tar}}^\pi = \max_{1 \leq j \leq m} t^{\pi,\ j}$ 目标。 ■

利用前述引理，我们得出以下样本复杂度结果。

引理19 令 $\epsilon > 0$ 和 $\delta > 0$。对于给定的 $\pi \in \Pi_+$，以 $1 - \delta$ 概率

$$\left\| \hat{Q}^\pi(\cdot, \cdot\cdot) \mathbf{1}_{\mathcal{R}}(\cdot) - Q^\pi(\cdot, \cdot\cdot) \mathbf{1}_{\mathcal{R}}(\cdot) \right\|_\infty \leq \frac{(2t_{\text{tar}} + 1) \left\| V^\pi \right\|_\infty}{H' + 1} + \frac{(H' + 2) \left\| K^\pi - \hat{K}^\pi \right\|_\infty}{2} + \epsilon.$$

具有样本复杂度

$$N' H' |\mathcal{S}||\mathcal{A}| = \mathcal{O}\left(\frac{2R^2 H'(H' + 2)^2 |\mathcal{S}||\mathcal{A}|}{\epsilon^2} \log(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}) \right).$$

证明 $\hat{Q}^j(s_0,a_0) = \frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - \hat{K}^\pi(s_i,a_i)\right)$。首先，我们可以按如下方式分解

$$
\begin{aligned}
&\left|\frac{1}{N'}\sum_{j=1}^{N'}\hat{Q}^j(s_0,a_0) - Q^\pi(s_0,a_0)\right| \\
&\le \frac{1}{N'}\sum_{j=1}^{N'}\left|\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - \hat{K}^\pi(s_i,a_i)\right) - \frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right| \\
&+ \left|\frac{1}{N'}\sum_{j=1}^{N'}\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right) - \mathbb{E}_\pi\left[\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right]\right| \\
&+ \frac{1}{N'}\sum_{j=1}^{N'}\left|\mathbb{E}_\pi\left[\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right] - Q^\pi(s_0,a_0)\right|
\end{aligned}
$$

对于第一项，

$$
\begin{aligned}
&\left|\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - \hat{K}^\pi(s_i,a_i)\right) - \frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right| \\
&= \left|-\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\hat{K}^\pi(s_i,a_i) + \frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}K^\pi(s_i,a_i)\right| \\
&\le \frac{(H'+2)\left\|K^\pi - \hat{K}^\pi\right\|_\infty}{2}
\end{aligned}
$$

对于最后一项，

$$
\begin{aligned}
&\left|Q^\pi(s_0,a_0) - \mathbb{E}_\pi\left[\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right]\right| \\
&= \left|(PV^\pi + r - K^\pi)(s_0,a_0) - \mathbb{E}_\pi\left[\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right]\right| \\
&= \left|PV^\pi(s_0,a_0) - \mathbb{E}_\pi\left[\frac{1}{H'+1}\sum_{h=1}^{H'}\sum_{i=1}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right]\right| \\
&= \left|PV^\pi(s_0,a_0) - P\left(\frac{1}{H'+1}\sum_{i=1}^{H'}(V^i - iJ^\pi)\right)(s_0,a_0)\right| \\
&= \left|\frac{1}{H'+1}PV^\pi(s_0,a_0) - \frac{H'}{H'+1}P\left(\frac{1}{H'}\sum_{i=0}^{H'}(V^i - iJ^\pi) - V^\pi\right)(s_0,a_0)\right| \\
&\le \frac{(2t^\pi_{\mathrm{tar}}+1)\left\|V^\pi\right\|_\infty}{H'+1}
\end{aligned}
$$

其中，最后一个不等式由引理 18得出。

对于第二项，令 $Q'_j = \frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i, a_i) - K^\pi(s_i, a_i)\right)$，根据霍夫丁不等式，有 $N' = \frac{2R^2}{\epsilon^2}(H'+2)^2)\log(\frac{2}{\delta})$，

$$\text{Prob}\left(\left|\frac{1}{N'}\sum_{j=1}^{N'} Q'_j(s_0,a_0) - \mathbb{E}\left[\frac{1}{N'}\sum_{j=1}^{N'}\frac{1}{H'+1}\sum_{h=0}^{H'}\sum_{i=0}^{h}\left(r_j(s_i,a_i) - K^\pi(s_i,a_i)\right)\right]\right| \le \epsilon\right) \ge 1-\delta.$$

根据对 $(s, a) \in \mathcal{S}\times\mathcal{A}$ 使用并集界，我们得出结论。 ■

现在我们可以证明定理。**证明** 在引理 19 中，令 $H' = \frac{3}{(2t^\pi_{\text{tar}}+1)\epsilon'}\|V^\pi\|_\infty$，$\epsilon = \frac{\epsilon'}{3}$。那么样本复杂度为

$$\begin{aligned} N'H'|\mathcal{S}||\mathcal{A}| &= \frac{2R^2H'(H'+2)^2|\mathcal{S}||\mathcal{A}|}{\left(\frac{\epsilon'}{3}\right)^2}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right) \\ &= \mathcal{O}\left(\frac{R^2(t^\pi_{\text{tar}})^3\|Q^\pi\|^3_\infty|\mathcal{S}||\mathcal{A}|}{(\epsilon')^5}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)\right)\end{aligned}$$

其中我们使用了 $\|V^\pi\|_\infty \le \|Q^\pi\|_\infty$. 此外，根据引理 17，样本复杂度为

$$\begin{aligned} NH|\mathcal{S}||\mathcal{A}| &= \frac{32|\mathcal{S}||\mathcal{A}|(\|V^\pi\|_\infty + R)R^2}{\left(\frac{2}{3(H'+2)}\right)^3}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right) \\ &= \mathcal{O}\left(\frac{(t^\pi_{\text{tar}})^3\|Q^\pi\|^4_\infty R^3|\mathcal{S}||\mathcal{A}|}{(\epsilon')^3}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)\right)\end{aligned}$$

我们可以得到 $\left\|K^\pi - \hat{K}^\pi\right\|_\infty \le \frac{2\epsilon'}{3}(H'+2)$。那么这意味着 $\left\|\hat{Q}^\pi(\cdot,\cdot)\mathbf{1}_{\mathcal{R}}(\cdot) - Q^\pi(\cdot,\cdot)\mathbf{1}_{\mathcal{R}}(\cdot)\right\|_\infty \le \frac{\epsilon'}{3} + \frac{\epsilon'}{3} + \frac{\epsilon'}{3} \le \epsilon'$

因此，结合前面的结果，以 $1-\delta$ 概率，我们有 $\left\|G^\pi - \hat{G}^\pi\right\|_\infty \le \epsilon'$，总样本复杂度为

$$\mathcal{O}\left(\frac{R^3(t^\pi_{\text{tar}})^3\|Q^\pi\|^4_\infty|\mathcal{S}||\mathcal{A}|}{(\epsilon')^6}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)\right)$$

■

D.2. 引理12的证明

将 $\mathcal{F}_t = \{\sigma(\{s_i : 0 \le i \le t\})\}$记为马尔可夫链中采样过程生成的自然滤过。

首先，我们证明以下引理。

引理20 令 $s_0 \in \mathcal{R}_i$ 和 $\pi \in \Pi_+$。对于给定的确定性 t、$n > 0$,

$$\textit{Prob}((t^i_{\text{cov}} - t)_+ \ge n \,|\, \mathcal{F}_t) \le \frac{\mathbb{E}[(t^i_{\text{cov}} - t)_+ \,|\, \mathcal{F}_t]}{n} \le \frac{t^{\pi,i}_{\text{cov}}}{n}$$

证明 第一个不等式由马尔可夫不等式得到。对于第二个不等式，定义从 t 开始的新覆盖时间 t'_{cov}，即 $t'_{\text{cov}} = \inf\{u \geq 0 : \{s_j\}_{j=t}^{u+t-1} = \mathcal{R}_i\}$ 根据覆盖时间的定义，随后 $t^i_{\text{cov}} - t \leq t'_{\text{cov}}$ 几乎必然成立，这意味着 $\mathbb{E}[(t^i_{\text{cov}} - t)_+ \,|\, \mathcal{F}_t] \leq \mathbb{E}[t'_{\text{cov}} \,|\, \mathcal{F}_t] = \mathbb{E}[t'_{\text{cov}} \,|\, s_t] \leq t^{\pi,\ i}_{\text{cov}}$。 ■

根据引理20 以及 t^π_{cov}的定义，我们可以直接得到以下推论 .

推论21 令 $s_0 \in \mathcal{R}$ 和 $\pi \in \Pi_+$。对于给定的确定性 $t, n > 0$,

$$Prob((t_{\text{cov}} - t)_+ \geq n \,|\, \mathcal{F}_t) \leq \frac{\mathbb{E}[(t_{\text{cov}} - t)_+ \,|\, \mathcal{F}_t]}{n} \leq \frac{t^\pi_{\text{cov}}}{n}$$

其中 $t_{\text{cov}} =$ 下确界$\{u \geq 0 : \{s_j\}_{j=0}^{u-1} = \mathcal{R}_i$ 对于某个 $i\}$

现在，我们来证明引理12。

证明 假设 s_0是暂态。如果 $k = \lceil t_{\frac{1}{2},\ \pi} \log(\frac{1}{\delta}) \rceil$ 使得 $\left\| (\bar{T}^\pi)^k \right\|_\infty < \delta$根据暂态半衰期的定义，这意味着以 $1 - \delta$ 概率， s_k 是常返的。

现在，假设 s_0 是常返的。那么，根据马尔可夫不等式和 $\mathbb{E}[t_{\text{cov}} \,|\, s_0 \in \mathcal{R}] \leq t^\pi_{\text{cov}}$，Prob($t_{\text{cov}} \geq 3t^\pi_{\text{cov}} \,|\, s_0 \in \mathcal{R}$) $\leq$Prob($t_{\text{cov}} \geq et^\pi_{\text{cov}} \,|\, s_0 \in \mathcal{R}$) $\leq e^{-1}$。因此，我们有 Prob($t_{\text{cov}} \geq 3kt^\pi_{\text{cov}} | \, s_0 \in \mathcal{R}$) $\leq e^{-k}$，因为根据归纳法，

$$\begin{aligned}
&\text{Prob}(t_{\text{cov}} \geq 3kt^\pi_{\text{cov}} \,|\, s_0 \in \mathcal{R}) \\
&= \mathbb{E}[\mathbf{1}_{t_{\text{cov}} \geq 3(k-1)t^\pi_{\text{cov}}} \text{Prob}(t_{\text{cov}} \geq 3kt^\pi_{\text{cov}} \,|\, \mathcal{F}_{3(k-1)t^\pi_{\text{cov}}}) \,|\, s_0 \in \mathcal{R}] \\
&= \mathbb{E}[\mathbf{1}_{t_{\text{cov}} \geq 3(k-1)t^\pi_{\text{cov}}} \text{Prob}((t_{\text{cov}} - 3(k-1)t^\pi_{\text{cov}})_+ \geq 3t^\pi_{\text{cov}} \,|\, \mathcal{F}_{3(k-1)t^\pi_{\text{cov}}}) \,|\, s_0 \in \mathcal{R}] \\
&\leq e^{-1}\mathbb{E}[\mathbf{1}_{t_{\text{cov}} \geq 3(k-1)t^\pi_{\text{cov}}} \,|\, s_0 \in \mathcal{R}] \\
&\leq e^{-k}.
\end{aligned}$$

其中倒数第二个不等式来自推论21。因此，Prob($t_{\text{cov}} \leq 3kt^\pi_{\text{cov}} \,|\, s_0 \in \mathcal{R}$) $\geq 1 - \delta$，样本复杂度为 $\lceil 3t^\pi_{\text{cov}} \log \frac{1}{\delta} \rceil$。

定义 $A = \{s_{M_1} \in \mathcal{R}\}, B = \{\cup_{i=M_1}^{M_1+M_2-1}\{s_i\} = \mathcal{R}_j$，其中 $j\}$。那么，Prob($A \cap B$) = Prob(A)Prob($B \,|\, A$) =Prob(A)Prob($B \,|\, s_{M_1} \in \mathcal{R}$) $\geq (1 - \frac{\delta}{2})(1 - \frac{\delta}{2}) \geq 1 - \delta$，样本复杂度为 $M_1 + M_2$，其中倒数第二个不等式来自前面的论证。 ■

D.3. 不精确 α 截断策略镜像上升

首先，我们考虑不精确策略镜像上升

$$\pi_{k+1}(\cdot \,|\, s) = \underset{p \in \mathcal{M}_\alpha(\mathcal{A})}{\text{argmax}} \left\{ \eta_k \sum_{a \in \mathcal{A}} \hat{G}^{\pi_k}(s,a)p(a) - D(p(\cdot), \pi_k(\cdot \,|\, s)) \right\}, \qquad \forall s \in \mathcal{S},$$

ˆ其中 G^{π_k} 是 G^{π_k} 的不精确评估。我们首先研究以下假设下不精确策略镜像上升的收敛性质。

ˆ**假设1** 不精确评估 G^{π_k} 满足

$$\left\| \hat{G}^{\pi_k} - G^{\pi_k} \right\|_\infty \leq \epsilon, \qquad \forall k \geq 0.$$

我们首先证明关键引理，即引理8的对应结果。

引理22 给定 $\pi_0 \in \Pi_\alpha$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，带步长 $\eta_k > 0$ 的非精确 α-截断策略镜像上升生成策略序列 $\{\pi_k\}_{k=1}^{\infty}$，满足

$$\sum_{a\in\mathcal{A}} \hat{G}^{\pi_k}(s,a)(\pi_k(\cdot\,|\,s) - \pi_{k+1}(\cdot\,|\,s)) \le 0, \qquad \forall s \in \mathcal{S},$$

and

$$J_\mu^{\pi_k} - J_\mu^{\pi_{k+1}} \le 2B_\alpha \epsilon.$$

证明 第一个不等式可直接由引理15中的相同论证得到。根据引理 3,

$$\begin{aligned} J_\mu^{\pi_k} - J_\mu^{\pi_{k+1}} &= \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_{k+1}}(s) \sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) \\ &= \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_{k+1}}(s) \sum_{a\in\mathcal{A}} \hat{G}^{\pi_k}(a,s)(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) \\ &\quad + \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_{k+1}}(s) \sum_{a\in\mathcal{A}} (G^{\pi_k}(a,s) - \hat{G}^{\pi_k}(a,s))(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)). \end{aligned}$$

对于第二项,

$$\begin{aligned} \sum_{a\in\mathcal{A}} (G^{\pi_k}(a,s) - \hat{G}^{\pi_k}(a,s))(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) &\le \left\| G^{\pi_k} - \hat{G}^{\pi_k} \right\|_\infty \|\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)\|_1 \\ &\le 2 \left\| \hat{G}^{\pi_k} - G^{\pi_k} \right\|_\infty \\ &\le 2\epsilon, \end{aligned}$$

其中，第一个不等式由霍尔德不等式得出，最后一个不等式由假设 1得出。由于 最大值 $_{\pi\in\Pi_\alpha} \|\rho^\pi\|_1 \le B_\alpha$，即可得到所需结果。 ■

以下事实将在定理 23的证明中使用。

事实 7 *[Xiao (2022年)]* 假设 $0 < \alpha < 1, b > 0$，且非负序列 $\{a_k\}$ 满足

$$a_{k+1} \le \alpha a_k + b, \qquad \forall k \ge 0.$$

那么对于所有 $k \ge 0$,

$$a_k \le \alpha^k a_0 + \frac{b}{1-\alpha}.$$

现在，我们给出非精确策略镜像上升的收敛结果。

定理23 考虑一个多链马尔可夫决策过程。在假设 *1*下，对于 $\pi_0 \in \Pi_\alpha$ 并给定 $\mu \in \mathcal{M}_+(\mathcal{S})$，非精确的 α-截断策略镜像上升采用恒定步长 $\eta > 0$ 生成策略序列 $\{\pi_k\}_{k=1}^{\infty}$，满足

$$J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \frac{1}{k+1}\left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta} + C_\alpha(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0})\right) + (2C_\alpha + k + 2)B_\alpha\epsilon.$$

并使用自适应步长 $\eta_{k+1}(C_\alpha - 1) \geq \eta_k C_\alpha > 0$ 生成满足以下条件的策略序列 $\{\pi_k\}_{k=1}^{\infty}$

$$J^{\pi_\alpha} - J_\mu^{\pi_k} \leq \left(1 - \frac{1}{C_\alpha}\right)^k \left(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0} + \frac{1}{\eta_0(C_\alpha - 1)} D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)\right) + 4C_\alpha B_\alpha \epsilon. \tag{1}$$

证明 沿用定理 9 中证明的相同论证,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} \rho_\mu^{\pi_\alpha}(s) \sum_{a\in\mathcal{A}} \hat{G}^{\pi_k}(a,s)(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) + \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_\alpha}(s) \sum_{a\in\mathcal{A}} \hat{G}^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s) - \pi_k(a\,|\,s)) \\
&\leq \frac{1}{\eta_k} D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_k) - \frac{1}{\eta_k} D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_{k+1}).
\end{aligned}$$

对于第一项,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} \rho_\mu^{\pi_\alpha}(s) \sum_{a\in\mathcal{A}} \hat{G}^{\pi_k}(a,s)(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) \\
&\geq C_\alpha \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_{k+1}}(s) \sum_{a\in\mathcal{A}} \hat{G}^{\pi_k}(a,s)(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) \\
&= C_\alpha \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_{k+1}}(s) \sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) \\
&+ C_\alpha \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_{k+1}}(s) \sum_{a\in\mathcal{A}} (\hat{G}^{\pi_k}(a,s) - G^{\pi_k}(a,s))(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) \\
&\geq C_\alpha \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_{k+1}}(s) \sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_k(a\,|\,s) - \pi_{k+1}(a\,|\,s)) - 2C_\alpha B_\alpha\epsilon \\
&\geq C_\alpha(J_\mu^{\pi_k} - J_\mu^{\pi_{k+1}}) - 2C_\alpha B_\alpha\epsilon
\end{aligned}$$

倒数第二个不等式由引理22 和假设 1 得出，最后一个不等式则由引理3 得出。

对于第二项,

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} \rho_\mu^{\pi_\alpha}(s) \sum_{a\in\mathcal{A}} \hat{G}^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s) - \pi_k(a\,|\,s)) \\
&= \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_\alpha}(s) \sum_{a\in\mathcal{A}} G^{\pi_k}(a,s)(\pi_\alpha(a\,|\,s) - \pi_k(a\,|\,s)) \\
&+ \sum_{s\in\mathcal{S}} \rho_\mu^{\pi_\alpha}(s) \sum_{a\in\mathcal{A}} (\hat{G}^{\pi_k}(a,s) - G^{\pi_k}(a,s))(\pi_\alpha(a\,|\,s) - \pi_k(a\,|\,s)) \\
&\geq J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} - 2B_\alpha\epsilon.
\end{aligned}$$

其中，最后一个不等式由引理15 和 3 以及假设 1。首先，考虑恒定
步长。我们有

$$J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \frac{1}{\eta_k} D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_k) - \frac{1}{\eta_k} D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_{k+1}) + C_\alpha (J_\mu^{\pi_{k+1}} - J_\mu^{\pi_k}) + 2(C_\alpha + 1) B_\alpha \epsilon$$

并对 k 求和，得到

$$\sum_{i=0}^{k} \left(J_\mu^{\pi_\alpha} - J_\mu^{\pi_i}\right) \leq \frac{1}{\eta} D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0) - \frac{1}{\eta} D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_{k+1}) + C_\alpha (J_\mu^{\pi_{k+1}} - J_\mu^{\pi_0}) + 2(k+1)(C_\alpha + 1) B_\alpha \epsilon$$

根据引理 22，$-J_\mu^{\pi_k} - 2B_\alpha \epsilon \leq -J_\mu^{\pi_{k-1}}$这意味着

$$J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \frac{1}{k+1} \left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta} + C_\alpha (J_\mu^{\pi_\alpha} - J_\mu^{\pi_0}) \right) + k B_\alpha \epsilon + 2(C_\alpha + 1) B_\alpha \epsilon.$$

其次，考虑自适应步长。

$$C_\alpha \left(U_{k+1} - U_k - 2B_\alpha \epsilon\right) + U_k \leq \frac{1}{\eta_k} D_k - \frac{1}{\eta_k} D_{k+1} + 2B_\alpha \epsilon,$$

其中 $U_k = J_\mu^{\pi_\alpha} - J_\mu^{\pi_k}$ 和 $D_k = D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_k)$。这意味着

$$C_\alpha (U_{k+1} - U_k) + U_k \leq \frac{1}{\eta_k} D_k - \frac{1}{\eta_k} D_{k+1} + 2B_\alpha (C_\alpha + 1) \epsilon.$$

两边除以 C_α，并对各项进行整理，结合自适应步长可得

$$U_{k+1} + \frac{1}{\eta_{k+1}(C_\alpha - 1)} D_{k+1} \leq \left(1 - \frac{1}{C_\alpha}\right) \left(U_k + \frac{1}{\eta_k (C_\alpha - 1)} D_k \right) + 4B_\alpha \epsilon.$$

最后，由事实 7结合 $a_k = U_k + \eta_k \frac{1}{(C_\alpha - 1)} D_k$，$\alpha = 1 - \frac{1}{C_\alpha}$，$b_k = 4B_\alpha \epsilon$，可得

$$U_k \leq \left(1 - \frac{1}{C_\alpha}\right)^k \left(U_0 + \frac{1}{\eta_0 (C_\alpha - 1)} D_0 \right) + 4 C_\alpha B_\alpha \epsilon.$$

■

D.4. 定理13

证明 首先，考虑恒定步长 $K = 2 \left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta} + C_\alpha (J_\mu^{\pi_\alpha} - J_\mu^{\pi_0}) \right) / \epsilon'$。根据定理 11 并对 $0 \leq k \leq K - 1$ 使用并集界，可得 $\left\| G^{\pi_k} - \hat{G}^{\pi_k} \right\|_\infty \leq \frac{\epsilon'}{(2C_\alpha + K + 2) B_\alpha}$，其样本复杂度为

$$K(HN + H'N')|\mathcal{S}||\mathcal{A}| = \mathcal{O}\left(\left(\frac{t_{\text{tar}}^3 \|Q^\pi\|_\infty^4 R^9 C_\alpha^6 \left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta} \right)^6 B_\alpha^6 |\mathcal{S}||\mathcal{A}|}{(\epsilon')^{12}} \right) \log \left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta} \right) \right)$$

对所有 $0 \leq k \leq K-1$。然后，根据定理 23，可得 $J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \frac{\epsilon'}{2} + \frac{\epsilon'}{2} \leq \epsilon'$.

对于自适应步长 $\eta_{k+1}(C_\alpha - 1) \geq \eta_k C_\alpha$ 其中 $K = \text{对数}\left(2(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0} + \frac{\rho_\mu^{\pi_\alpha}}{D(\pi_\alpha, \pi_0)} \eta_0 (C_\alpha - 1))/\epsilon\right)^{\text{对数}(C_\alpha/(C_\alpha-1))}$ 根据定理 11 以及对 $0 \leq k \leq K-1$, $\left\|G^{\pi_k} - \hat{G}^{\pi_k}\right\|_\infty \leq \frac{\epsilon'}{8 C_\alpha B_\alpha}$ 的并集界，其样本复杂度为

$$K(NH + N'H')|\mathcal{S}||\mathcal{A}| = \mathcal{O}\left(\frac{R^3 C_\alpha^6 B_\alpha^6 t_{\text{tar}}^3 \|Q^\pi\|_\infty^4 |\mathcal{S}||\mathcal{A}|}{(\epsilon')^6} K \log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right)\right)$$

对于所有 $0 \leq k \leq K-1$。那么，根据定理 23，我们有 $J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \frac{\epsilon'}{2} \pm \frac{\epsilon'}{2} \leq \epsilon'$. ■

附录 E. 弱可通信马尔可夫决策过程

在本章节中，我们假设马尔可夫决策过程是弱通信的。我们将证明，通过将 α 选取为 ϵ的函数， α截断策略镜像上升算法可以获得满足 $J_\mu^\star - J_\mu^\pi \leq \epsilon$的 ϵ最优策略（即不限于正策略）。

E.1. J^π 在弱通信下的连续性

我们首先论证 J^π 在弱连通MDP中最优策略处的连续性。

引理 24 在弱连通*MDP*中，映射 $\pi \mapsto J^\pi$和 $\pi \mapsto J_\mu^\pi$在最优策略处连续，并且对任意固定的 $\mu \in \mathcal{M}(\mathcal{S})$，有$J_{+,\mu}^\star = J_\mu^\star$。

证明 对于弱连通MDP，最优平均奖励是均匀向量（Puterman，2014，定理 8.32）。然后，在引理 3的证明中，我们有 $(P_\star^{\pi_\star} - I)J^\star = 0$，并且我们有

$$J_\mu^\pi - J_\mu^\star = \sum_{s \in \mathcal{R}} \sum_{a \in \mathcal{A}} d_\mu^\pi(s)\big(\pi(a \mid s) - \pi_\star(a \mid s)\big) Q^{\pi_\star}(s,a).$$

根据霍尔德不等式，$J_\mu^\star - J_\mu^\pi \leq \sum_{s \in \mathcal{R}} d_\mu^\pi(s) \big\|\pi(\cdot \mid s) - \pi_\star(\cdot \mid s)\big\|_1 \|Q^{\pi_\star}\|_\infty$。因此，对于 $\pi \in \Pi_+$，$\lim_{\pi \to \pi_\star} J_\mu^\pi = J_\mu^\star$。

■

给定 $\mu \in \mathcal{M}(\mathcal{S})$，我们称 π是一个 ϵ-最优策略，如果 $J_\mu^\star - J_\mu^\pi \leq \epsilon$。然后，引理24保证我们可以找到 ϵ-最优策略，对于任意 $\epsilon > 0$，只需考虑 Π_+中的策略。

E.2. 弱连通MDP中的策略梯度定理

我们首先陈述以下关于弱可通信马尔可夫决策过程的关键引理。

引理 25 考虑一个弱连通马尔可夫决策过程。对于每个 $\pi \in \Pi_+$, $J^\pi = c^\pi \mathbf{1}$ 存在某个$c^\pi \in \mathbb{R}$。

证明 根据（Puterman，2014，命题 8.3.1），存在策略 π，使得 P^{π} 具有单个常返类和暂态状态。接着根据可达性的定义，对于 $\pi' \in \Pi_+$，$P^{\pi'}$ 也具有相同的单个常返类和暂态状态。最后，由于 $d^{\pi'} = \mathbf{1}(g^{\pi'})^{\top}$ 其中 $g^{\pi'}$ 是 $P^{\pi'}$的唯一平稳分布，我们得出 $J^{\pi'} = (g^{\pi'})^{\top} r^{\pi'} \mathbf{1}$。 ■

将引理 25 应用于引理3的证明中，可直接得到弱可通信马尔可夫决策过程的以下性能差异引理。

C推论 26 考虑一个弱可通信马尔可夫决策过程。对于 π、$\pi' \in \Pi_+$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，

$$J_{\mu}^{\pi} - J_{\mu}^{\pi'} = \sum_{s \in \mathcal{R}} \sum_{a \in \mathcal{A}} d_{\mu}^{\pi}(s)\big(\pi(a \mid s) - \pi'(a \mid s)\big) Q^{\pi'}(s,a).$$

同样，将引理 25 应用于定理5的证明中，可直接得到弱可通信马尔可夫决策过程的以下策略梯度定理。

推论 27 考虑一个弱可通信马尔可夫决策过程。对于 $\pi_\theta \in \Pi_+$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，

$$\nabla_\theta J_{\mu}^{\pi_\theta} = \sum_{s \in \mathcal{R}} \sum_{a \in \mathcal{A}} d_{\mu}^{\pi_\theta}(s) \nabla_\theta \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a).$$

E.3. 弱可通信马尔可夫决策过程中策略镜像上升的收敛性

若马尔可夫决策过程是弱可通信的，则我们可以明确给出找到一个 ϵ-最优策略所需的迭代次数，其中 α 的取值选为 ϵ 的函数。

我们首先给出具有恒定步长的 α-裁剪策略镜像上升的次线性收敛速率。

推论 28 考虑弱连通*MDP*。对于任意给定的 $\epsilon \in (0,\ 1)$，令 $\alpha = \frac{\epsilon}{(|\mathcal{A}|+1)\|Q^{\pi^\star}\|_\infty}$。那么，对于 $\pi_0 \in \Pi_\alpha$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，$\alpha$-裁剪策略镜像上升在恒定步长 η 下生成 ϵ-最优策略，适用于 $k \geq \frac{2}{\epsilon}\left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\ \pi_0)}{\eta} + C_\alpha(J^{\pi_\alpha} - J^{\pi_0})\right)$.

证明 按照引理24的证明，如果 $\alpha = \frac{\epsilon}{(|\mathcal{A}|+1)\|Q^{\pi^\star}\|_\infty}$，则有

$$\begin{aligned} J_{\mu}^{\star} - J_{\mu}^{\pi} &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_{\mu}^{\pi}(s)(\pi^\star(a \mid s) - \pi(a \mid s)) Q^{\pi_\star}(s,a) \\ &\leq (|\mathcal{A}|+1)\alpha \left\|Q^{\pi_\star}\right\|_\infty \\ &\leq \frac{\epsilon}{2}. \end{aligned}$$

根据定理9，当 $k \geq \frac{2}{\epsilon}\left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\ \pi_0)}{\eta} + C_\alpha(J^{\pi_\alpha} - J^{\pi_0})\right)$，$J_{\mu}^{\pi_\alpha} - J_{\mu}^{\pi_k} \leq \frac{\epsilon}{2}$，从而证明完毕。 ■

接下来，我们给出采用自适应步长的 α 截断策略镜像上升的线性 收敛速率。

推论29 考虑一个弱通信的马尔可夫决策过程。对于任意给定的 $\epsilon \in (0,\ 1)$，令 $\alpha = \frac{\epsilon}{2(|\mathcal{A}|+1)\|Q^{\pi^\star}\|_\infty}$。那么，对于 $\pi_0 \in \Pi_\alpha$ 和 $\mu \in \mathcal{M}_+(\mathcal{S})$，采用步长 $(C_\alpha - 1)\eta_{k+1} \geq C_\alpha \eta_k > 0$ 的 α 截断策略镜像上升可生成 ϵ 最优策略

$$k \geq \frac{\log\left(2(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0} + \frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta_0(C_\alpha - 1)})/\epsilon\right)}{\log(C_\alpha/(C_\alpha - 1))}$$

.

证明 沿用推论 28 的证明，令 $\alpha = \frac{\epsilon}{(|\mathcal{A}|+1)\|Q^{\pi^\star}\|_\infty}$ 则存在 $\pi \in \Pi_\alpha$ 使得 $\|\pi^\star - \pi\|_\infty \leq \alpha$。此外，我们有 $J_\mu^\star - J_\mu^\pi \leq \frac{\epsilon}{2}$。根据定理 10，当 $K \geq \frac{\log\left(4(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0} + D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)\eta_0(C_\alpha - 1))/\epsilon\right)\log(C_\alpha/(C_\alpha - 1))}{}$，$J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \frac{\epsilon}{2}$，由此完成证明。 ■

据我们所知，推论 28和29 是针对平均奖励弱可通信马尔可夫决策过程的策略梯度方法的首批收敛结果。

E.4. 弱通信马尔可夫决策过程中策略镜像上升的样本复杂度

如果马尔可夫决策过程是弱通信的，我们可以明确给出寻找一个 ϵ-最优策略的样本复杂度。注意，对于 $\pi \in \Pi_+$，在弱通信马尔可夫决策过程中有 $\|d_\mu^\pi\|_1 = B_\alpha = 1$。

推论 30 考虑一个弱通信马尔可夫决策过程。令 $\epsilon > 0$、$\delta > 0$，并设 $\alpha = \frac{\epsilon}{2(|\mathcal{A}|+1)\|Q^{\pi^\star}\|_\infty}$。对于任意 π、$\pi_0 \in \Pi_\alpha$，并给定 $\mu \in \mathcal{M}_+(\mathcal{S})$，以概率 $1-\delta$，采用恒定步长 $\eta > 0$ 的随机 α-裁剪策略上升迭代，以及 $K = 4\left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta} + C_\alpha(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0})\right)/\epsilon$，会生成 ϵ-最优策略，其样本复杂度为

$$\widetilde{\mathcal{O}}\left(\frac{t_{\text{tar}}^3 \|Q^\pi\|_\infty^4 R^9 C_\alpha^6 \left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)}{\eta}\right)^6 |\mathcal{S}||\mathcal{A}|}{\epsilon^{12}}\right)$$

且采用自适应步长 $\eta_{k+1}(C_\alpha - 1) \geq \eta_k C_\alpha > 0$ 和 $K = \frac{\log\left(4(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0} + D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha, \pi_0)\eta_0(C_\alpha - 1))/\epsilon\right)}{\log(C_\alpha/(C_\alpha - 1))}$ 可生成 ϵ-最优策略，其样本复杂度为

$$\tilde{O}\left(\frac{t_{\text{tar}}^3 \|Q^\pi\|_\infty^4 R^3 C_\alpha^6 |\mathcal{S}||\mathcal{A}|}{\epsilon^6}\right)$$

其中 $\mathcal{O}$ 忽略所有对数因子，$\|Q^\pi\|_\infty = \max_{0 \leq k \leq K-1} \|Q^{\pi_k}\|_\infty$，且 $t_{\text{tar}} = \max_{0 \leq k \leq K-1} t^{\pi_k}{}_{\text{tar}}$。

证明 沿用推论 28 的证明，令 $\alpha = \frac{\epsilon}{(|\mathcal{A}|+1)\|Q^{\pi^\star}\|_\infty}$ 那么存在 $\pi \in \Pi_\alpha$ 使得 $\|\pi^\star - \pi\|_\infty \leq \alpha$。此外，我们有 $J_\mu^\star - J_\mu^\pi \leq \frac{\epsilon}{2}$。

首先，对于恒定步长，根据定理 13，$K = 4\left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_0)}{\eta} + C_\alpha(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0})\right)/\epsilon$，$J_\mu^{\pi_\alpha} - J_\mu^{\pi_k} \leq \frac{\epsilon}{2}$，其样本复杂度为

$$K(HN + H'N')|\mathcal{S}||\mathcal{A}| = \mathcal{O}\left(\left(\frac{t_{\text{tar}}^3 \|Q^\pi\|_\infty^4 R^9 C_\alpha^6 \left(\frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_0)}{\eta}\right)^6 |\mathcal{S}||\mathcal{A}|}{\epsilon^{12}}\right)\log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right)\right).$$

那么，我们有 $J_\mu^\star - J_\mu^{\pi_k} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \leq \epsilon$.

对于自适应步长 $\eta_{k+1}(C_\alpha - 1) \geq \eta_k C_\alpha$ with $K = \frac{\log\left(4(J_\mu^{\pi_\alpha} - J_\mu^{\pi_0} + \frac{D_{\rho_\mu^{\pi_\alpha}}(\pi_\alpha,\pi_0)}{\eta_0(C_\alpha-1)})/\epsilon\right)}{\log(C_\alpha/(C_\alpha-1))}$，其样本复杂度为

$$K(HN + H'N')|\mathcal{S}||\mathcal{A}| = \left(\frac{R^3 C_\alpha^6 t_{\text{tar}}^3 \|Q^\pi\|_\infty^4 |\mathcal{S}||\mathcal{A}|}{\epsilon^6} K \log\left(\frac{2K|\mathcal{S}||\mathcal{A}|}{\delta}\right)\right).$$

那么，我们有 $J_\mu^\star - J_\mu^{\pi_k} \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \leq \epsilon$. ■

据我们所知，推论30 是平均奖励弱连通马尔可夫决策过程策略梯度方法中首个样本复杂度结果。

附录 F. 投影算法

算法 4到 $\mathcal{M}_\alpha(\mathcal{A})$ 的欧几里得投影

输入： $q \in \mathbb{R}^d, \alpha \in [0, 1/d]$

将 q 排序为满足 $q'_1 \geq q'_2 \geq \cdots \geq q'_d$的 q'

求 $\rho =$ 的最大值 $\left\{1 \leq j \leq d : \quad q'_j + \frac{1}{j}\left(1 - d\alpha - \sum_{i=1}^j q'_i\right) > 0\right\}$

$\lambda = \frac{1}{\rho}\left(1 - d\alpha - \sum_{i=1}^\rho q'_i\right)$

对于 $i = 1, \cdots\cdots, d$ **do**

$p_i =$ 最大值$\{q_i + \lambda + \alpha, \alpha\}$

结束循环

输出： p

算法 5对 $\mathcal{M}_\alpha(\mathcal{A})$进行 KL 投影

输入： $w \in \mathcal{M}_+(\mathcal{S})$ 其中 $|\mathcal{S}| = d, \alpha \in [0, 1/d]$

设置 $W = \{1, \ldots, d\}, C_\# = 0, C_\% = 0$

当 $W \neq \emptyset$**时执行**

在 $\{w_i$ 中查找中位数 $\omega : \ i \in W\}$

$L = \{i \in W : w_i < \omega\}, \quad L_\# = |L|, \quad L_\% \leftarrow \sum_{i \in L} w_i$

$M = \{i \in W : \ w_i = \omega\}, \quad M_\# = |M|, \quad M_\% = \sum_{i \in M} w_i$

$H = \{i \in W : \ w_i > \omega\}$

$m_0 = \dfrac{1 - (C_\# + L_\#)\alpha}{1 - (C_\% + L_\%)}$

如果 $\omega\, m_0 < \alpha$ **则**

$C_\# = C_\# + L_\# + M_\#$

$C_\% = C_\% + L_\% + M_\%$

$W = H$

else

$W = L$

结束如果

结束当

$m_0 = \dfrac{1 - C_\# \alpha}{1 - C_\%}$

对于 $i = 1, \ldots, d$ **do**

如果 $w_i < \omega$ **则**

$p_i = \alpha$

else

$p_i = m_0\, w_i$

结束如果

结束循环

输出： p