一种近似　上升方法以证明PP的收敛　O

Leif Döring[1], Daniel Schmidt[1], Moritz Melcher[1], Sebastian Kassing[2], Benedikt Wille[1], TilmanAach[1], 和 Simon Weissmann[1]

[1]数学研究所，曼海姆大学，68138曼海姆，德国，{leif.doering, daniel.schmidt, moritz.melcher, benedikt.wille, tilman.aach, simon.weissmann}@uni-mannheim.de[2]数学与信息学系，伍珀塔尔大学，42119伍珀塔尔，德国，{kassing}@uni-wuppertal.de

arXiv:2602.03386v1 [cs.LG] 3 Feb 2026

摘要

近端策略优化（PPO）是使用最为广泛的深度强化学习算法之一，但其理论基础仍然不完整。最重要的是，收敛以及对近端策略优化基本优势的理解仍悬而未决。在标准理论假设下，我们表明，近端策略优化的策略更新方案（在多次使用的轨迹上使用代理梯度执行多个训练轮次的小批量更新）可以被解释为近似策略梯度上升。我们展示了如何控制由替代梯度累积的偏差，并运用随机重排技术证明了近端策略优化的收敛定理，从而阐明近端策略优化成功的原因。此外，我们发现了近端策略优化中常用的截断广义优势估计里一个此前被忽视的问题。该几何加权方案会在回合边界处使无限质量坍缩到最长的 k步优势估计器上。实证评估表明，简单的权重修正能够在具有强烈终止信号的环境中带来显著改进，例如Lunar Lander。

1 引言

强化学习（RL）已成为一种强大的范式，通过与环境交互来训练自主智能体进行序贯决策 [33]。近年来，策略梯度方法已成为许多成功应用的基础，涵盖游戏对弈 [31, 20] 、机器人学 [14, 21] 以及大语言模型对齐 [22, 5]。从本质上说，策略梯度方法旨在通过沿参数化期望总回报的梯度 $J(\theta)$ 更新策略 π_θ ，从而对其进行优化。策略梯度定理 [34, 38] 为该梯度提供了一个无偏估计量，可利用当前策略的样本进行计算。演员-评论家方法利用这一思想，在收集轨迹采样、估计价值函数（评论家）和更新策略之间交替进行。一个典型例子是优势演员-评论家 [19], ，它会在收集新的同策略数据之前，对每个批次执行一次更新。在这些方法中，近端策略优化（PPO） [30] 因其简单、稳定且实证表现出色，已成为应用最广泛的算法之一。PPO 最初被提出，作为信赖域策略优化（TRPO） [28]的一种实用的一阶近似。TRPO 提出通过最大化代理目标函数来改进参考策略 $\pi_{\theta_{\text{old}}}$ ，同时满足限制策略变化幅度的信赖域约束：

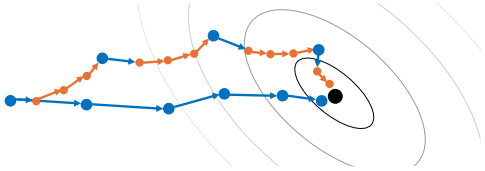


图 1：PPO 与优势演员-评论家策略参数更新的示意图。蓝色箭头表示策略梯度$\nabla_\theta J(\theta)$步骤，橙色箭头表示一系列周期，每个周期都采用偏差逐渐增大的替代梯度 $g^{\text{clip}}{}_{\text{PPO}}(\theta,\ \theta_{\text{old}})$步骤。对于随机近似，蓝点表示重新采样的时刻。

最大化代理目标函数，同时满足限制策略变化幅度的信赖域约束：

$$\begin{aligned} &\text{maximize}_\theta\ \hat{\mathbb{E}}_t\left[r_t(\theta)\hat{\mathbb{A}}_t^{\pi_{\text{old}}}\right], \\ \text{subject to}\quad &\hat{\mathbb{E}}_t\left[\text{KL}(\pi_{\theta_{\text{old}}}(\cdot\,;\,s_t)\,\|\,\pi_\theta(\cdot\,;\,s_t))\right]\ \leq\ \delta, \end{aligned}$$

其中 $r_t(\theta)\ =\ \frac{\pi_\theta(a_t;s_t)}{\pi_{\theta_{\text{old}}}(a_t;s_t)}$ ，$\hat{}\ \mathbb{A}_t^{\pi_{\text{old}}}$是一个优势估计。然而，求解约束问题需要二阶信息。PPO 被提出，作为信赖域原理的一种可实现的放宽形式。在 [30],的裁剪变体中，目标函数为

$$L(\theta) = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{\mathbb{A}}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{\mathbb{A}}_t\right)\right],$$

$\hat{}$其中 $\mathbb{A}_t$ 通过在旧策略$\pi_{\theta_{\text{old}}}$下计算的（截断）广义优势估计（GAE） [29], 获得。尽管 PPO 在实践中的成功有目共睹，但对 PPO 的理论理解仍基本处于空白状态，而且其在实践中的决定性优势也难以确定 [6]。例如，似乎还没有收敛结果考虑到样本重用：转移缓冲区会在每个训练轮次中随机打乱。理论缺乏发展的一个主要原因可能是，PPO 与 TRPO 之间的联系相当启发式，因此难以将其作为定理的基础。我们探讨以下基本问题：

PPO 有什么良好的理论基础？理论能为实际应用带来哪些启示？

我们的论文换了一个视角。我们不再关注与信赖域策略优化的联系，而是只将近端策略优化重新视为一种经过良好组织、能够重用样本的策略梯度方法。近端策略优化具有周期结构：先进行一次优势演员-评论家更新，然后进行若干步替代梯度，见图 1。优势演员-评论家与近端策略优化首个周期中各步骤之间的关系，早先已在 [8] 中有所观察。可视化中的蓝色箭头表示优势演员-评论家的梯度步，橙色箭头表示近端策略优化额外的替代梯度步；随着周期推进，这些步的可信度会逐渐降低。

对 *PPO* 理论与实践的主要贡献：

- 本文通过梯度代理函数给出了一种形式化方法 $g_{\text{clip old}}$ of $\nabla_\theta J$ （第 4 节），使其随机近似与实际的 PPO 实现接近，并采用了 PPO 更新机制的大部分特性（我们略去 KL 正则化和非对称裁剪）。

- 形式为 $|\nabla_\theta J(\theta) - g_{\text{clipppo}}(\theta, \theta_{\text{旧}})| \leq C|\theta - \theta_{\text{旧}}|$ 的精确梯度偏差估计在定理 4.2 中导出。这表明图 1 中的橙色箭头可以有多可信。

- 收敛性证明在定理 5.1 和 6.2 中给出，展示了额外的有偏替代梯度对随机策略梯度和确定性策略梯度的影响。我们将近端策略优化与随机重排（RR）理论联系起来。分析表明，近端策略优化基于周期的更新结构通过聚合裁剪梯度估计，隐式地控制了有效步长。

- 实际应用：我们对近端策略优化所做的相应有限时间建模，凸显了近端策略优化在有限时域截断广义优势估计的一个副作用。我们将该效应称为尾部质量塌缩，并提出一种简单的修正。实验表明，在 Lunar Lander 上有显著改进。

2 策略梯度基础

虽然许多策略梯度结果是在无限时域折扣设定中陈述的，但我们直接采用有限时域截断，以贴近实际的近端策略优化实现。我们假设有限状态空间与动作空间 $\mathcal{S}$ 与 $\mathcal{A}$ 以及固定的初始分布 μ。价值函数记为 $V_t^\pi(s) = \mathbb{E}^\pi[\sum_{i=t}^{T-1}\gamma^{i-t}R_i \,|\, S_t = s]$与 $Q_t^\pi(s,a) = \mathbb{E}^\pi[\sum_{i=t}^{T-1}\gamma^{i-t}R_i \,|\, S_t = s, A_t = a]$。此外，我们使用可微参数化策略类 $\{\pi_\theta\}_{\theta\in\mathbb{R}^d}$，其所谓的得分函数为 $\nabla_\theta \log \pi_\theta(a;\ s)$。优化目标是最大化参数化价值函数$J(\theta) := V_0^{\pi_\theta}(\mu) = \sum_{s\in\mathcal{S}} V_0^{\pi_\theta}(s)\mu(s)$。若假定奖励有界，则 $J_*:=\sup_\theta J(\theta)$存在。根据似然比恒等式，策略梯度具有随机梯度表示$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_\theta}[\nabla_\theta \log \pi_\theta(A_t;\ S_t)\ R_t^T]$，其中待实现奖励定义为 $R_t^T:= \sum_{i=t}^{T-1}\gamma^{i-t}R_i$。由此得到的简单策略梯度估计器因方差过高而众所周知过于嘈杂，不适合实际优化。为降低方差，常用的策略梯度表示使用平均待实现奖励并减去一个基线。在折扣有限时间设定中，这是$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_\theta}[\nabla_\theta \log \pi_\theta(A_t;\ S_t)\ \mathbb{A}_t^{\pi_\theta}(S_t,A_t)],$，其中 $\mathbb{A}_t^{\pi_\theta} = Q_t^{\pi_\theta} - V_t^{\pi_\theta}$称为优势函数。算法上，优势函数策略梯度造成了一种结构性困难。为了用梯度上升改进演员，需要评估当前策略，即需要计算/估计 $\mathbb{A}_t^{\pi_\theta}$。算法上的解决方案即所谓的演员-评论家。优势演员-评论家算法在用于改进策略的梯度步与用于产生估计$\hat{}\ \mathbb{A}_t^{\pi_\theta}$的估计步之间交替。优势演员-评论家 [19] 使用神经网络进行优势函数建模，并使用广义优势估计进行优势估计。

注 2.1。 众所周知，演员-评论家算法的实现通常会忽略折扣因子 γ^t，理论方面的讨论可参见 [35]，实验方面的讨论可参见 [41] 和 [4]。由于对于近似的无限时域问题而言，该因子在数学上是必要的，因此我们保留 γ^t，并承认省略 γ 并无害处。这与 [3], 中讨论的形式主义与实现不匹配相一致；该文建议关注强化学习算法的结构性理解，而不是那些能够改进基准测试的构造。

3 相关工作

本文延续了一系列证明策略梯度算法收敛性的理论论文，但据我们所知，关于 PPO 风格策略更新机制的研究寥寥无几。由于非凸优化景观，策略梯度的分析通常颇具挑战；人们一般依赖目标函数的 L-平滑性，而这需要对策略类作出相当强的假设。一些研究关注收敛到驻点[40, 23]。在表格型 Softmax 参数化等强策略假设下，可以证明

包括梯度支配在内的额外结构条件，并据此推导出收敛至全局最优解及其收敛速率（例如 [1, 17, 25]）。关于演员-评论家算法的理论结果，可参见 [12] 及其中的参考文献。大多数理论论文分析的是折扣无限时域 MDP 设置。由于 PPO 的实现会进行截断，我们决定研究有限时域情形。在有限时域中，最优策略未必是平稳的； [11]提出了一种用于寻找非平稳策略的策略梯度算法。秉承 PPO 的理念，本文分析了对最优平稳策略的搜索。

与关于原始策略梯度收敛理论的海量文献相比，关于近端策略优化的研究则较为有限。其原因在于裁剪机制，尤其是代理偏差和数据重用。[15] 利用无限维镜像下降，证明了一种神经网络 PPO 变体的收敛性。关于近期的两项收敛结果，我们请参阅 [9]和 [16]，但需注意，二者都不允许重用重洗牌数据。 [9] 基本证明了代理梯度步不会损害原始策略梯度方案，而 [16] 则针对一种特定的策略设置开展研究（特别是概率有正的下界，从而可以证明梯度支配性质）。据我们所知，目前尚无纳入样本重用的研究。为纳入多样本使用，我们的工作建立在有限和设置下随机重排的既有研究成果之上，这类研究与监督学习相关，例如 [18]。此外，我们还请读者参阅 [27]及其中的参考文献。

最后，由于我们也对广义优势估计器的实际应用有所贡献，这里简要介绍一些相关工作。广义优势估计最初是为无限时域 MDP 提出的， [30] 建议通过在 T处直接截断，将其用于有限时域。近期，广义优势估计的截断被用于大语言模型领域的轨迹子集 [2] ，也被用于经典环境 [32]。据我们所知，我们对尾部质量塌缩的观察以及对塌缩质量的重新加权，都是新的。

相关理论文章中的典型假设是奖励有界，且得分函数有界或满足利普希茨条件。我们将在这些假设下开展工作，因为它们允许我们使用上升不等式。此外，我们假设可以使用一个表现良好的评论家。

假设 3.1. • 有界奖励： $|R_t| \leq R_*$。

- 有界得分函数： $\forall \theta \in \mathbb{R}^d, s \in \mathcal{S}, a \in \mathcal{A}$:

$$\left|\nabla_\theta \log \pi_\theta(a;s)\right| \leq \Pi_*$$

- 利普希茨得分函数： $\forall \theta \in \mathbb{R}^d, s \in \mathcal{S}, a \in \mathcal{A}$:

$$\left|\nabla_\theta \log \pi_\theta(a;s) - \nabla_{\theta'} \log \pi_{\theta'}(a;s)\right| \leq L_s \left|\theta - \theta'\right|$$

- 可使用 $\hat{}$ 优势估计 $\mathbb{A}_t$ ，其有界于 A_* 且具有均匀的估计偏差 δ .

4 重新思考 PPO

与策略梯度方案的起源不同，近端策略优化并非作为基于轨迹展开的随机梯度近似被提出，而是作为一种直接算法，并（非常成功地）聚焦于实现细节。在我们的分析中，我们以不同方式引入近端策略优化：通过推导精确策略梯度 $\nabla_\theta J$ 的代理函数 g^{clip} 近端策略优化，而近端策略优化正是其自然的随机近似。我们首先说明为何在策略梯度中加入 g^{clip} 近端策略优化 代理梯度步是合理的，然后证明对 $\nabla_\theta J$ 和 g^{clip} 近端策略优化 进行循环使用的有偏梯度上升（见图 1）确实具有理论优势（第5节）。最后，我们给出近端策略优化的收敛结果（第6节）。

下面是一个出发点。众所周知，数据多次使用是监督学习中一种成功的收敛加速方法，尤其是与随机重排结合时。在随机重排（随机重排）中，样本的小批次会被重复用于多个随机梯度下降步，并在对数据的完整遍历（称为训练轮次）之间进行重排。在在线强化学习中，这会带来问题，因为梯度步依赖于当前的采样策略。因此需要一种循环变体，其内层循环对循环开始时收集的数据执行随机重排策略更新。将采样与更新解耦的一种有原则的方法是重要性采样（IS）：

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}^{\pi_{\theta_{\text{old}}}} \Big[\prod_{i=0}^{t} \frac{\pi_\theta(A_i\,;\,S_i))}{\pi_{\theta_{\text{old}}}(A_i\,;\,S_i))} \nabla_\theta \log \pi_\theta(A_t\,;\,S_t)\, \mathbb{A}_t^{\pi_\theta}(S_t, A_t) \Big],$$

其中，$\pi_{\theta_{\text{old}}}$ 是用于采样轨迹展开的策略。原则上，可以尝试按以下方式实现数据重用：在某个参数实例 θ_{old}下采样轨迹展开，使用小批量估计梯度，并针对轨迹展开数据执行若干轮经重要性采样校正的梯度步。然后将当前参数记为 θ_{old} ，并开始下一个周期。策略梯度（或优势演员-评论家）使用新采样的轨迹展开数据，在 θ_{old}执行一次采样梯度步；而采用样本重用的策略梯度则在 θ_{old} 执行一次采样策略梯度步，此外还会使用固定的轨迹展开数据执行一个周期的 IS 梯度步。

问题：（i）当 π_θ偏离采样分布$\pi_{\theta_{\text{old}}}$时，重要性采样权重可能会在 t 上累积，导致方差极大。（ii）策略梯度涉及 $\mathbb{A}_t^{\pi_\theta}$，尽管数据来自$\pi_{\theta_{\text{old}}}$；无法使用来自$\pi_{\theta_{\text{old}}}$ 的轨迹展开通过广义优势估计来估算这一项。（iii）重要性比率要求使用基于轨迹展开的估计器，而基于状态转移的估计器方差更小（时间相关性更低）。

我们现在说明，PPO 可以被理解为对 (i)-(iii) 的一种可实现的回应。为了解决 (i)-(iii)，需要

- 舍弃除一个以外的所有重要性采样比率，并对剩余比率进行裁剪，
- 将 $\mathbb{A}_t^\pi$ 替换为 $\mathbb{A}^{\pi_\theta}$旧的t ，以便从π_θ旧的轨迹采样中进行广义优势估计，
- 将 $\frac{1}{T}\sum_{t=0}^{T}$视为期望，并通过从转移缓冲区中均匀采样来估计它；该缓冲区由展平轨迹缓冲区得到（需要第一步）。

前两项导出如下近似

$$\begin{aligned}
&\sum_{t=0}^{T-1} \gamma^t \mathbb{E}^{\pi_{\theta_{\text{old}}}} \Big[\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\text{old}}}(A_t\,;\,S_t)} \mathbb{1}_{|\frac{\pi_\theta(A_t;S_t)}{\pi_{\text{old}}(A_t;S_t)}-1|\le\epsilon} \times \nabla_\theta \log \pi_\theta(A_t\,;\,S_t)\, \mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t) \Big] \\
&= \sum_{t=0}^{T-1} \gamma^t \mathbb{E}^{\pi_{\theta_{\text{old}}}} \Big[\frac{\nabla_\theta \pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\text{old}}}(A_t\,;\,S_t)} \times \mathbb{1}_{|\frac{\pi_\theta(A_t;S_t)}{\pi_{\text{old}}(A_t;S_t)}-1|\le\epsilon} \mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t) \Big] \\
&=: g_{\text{PPO}}^{\text{clip}}(\theta, \theta_{\text{old}})
\end{aligned}$$

$\nabla_\theta J(\theta)$，其中我们利用了 $\pi_\theta(a;\, s)\nabla_\theta \log\, \pi_\theta(a;\, s) \;=\; \nabla_\theta \pi_\theta(a;\, s)$。这正是 PPO 梯度替代目标期望的形式化表达。均匀采样的视角将引出第6节中的真正 PPO 采样器。与 PPO 相比，这里有两处细微改动。

注释 4.1. 在 PPO 中会忽略折扣因子 γ^t ，但本文不会。我们的理论也可以用 $\gamma = 1$ 同样加以推导。接下来，我们独立于优势函数裁剪重要性采样比率，而 PPO 使用非对称

裁剪。我们不讨论非对称裁剪，因为具体选择很大程度上取决于问题（例如参见大语言模型背景下的 [26] ）。
需要强调的是，延迟优势以及舍弃/截断重要性采样比率会引入偏差。为避免阻碍收敛，这种偏差在更新周期中不应增长过快。尽管这一事实已为人所知，本文的一项主要结果给出了替代梯度偏差的界：

定理 4.2（替代梯度偏差控制）。 在假设 *3.1* 下，有

$$\left|\nabla_\theta J(\theta) - g_{PPO}^{clip}(\theta, \theta_{old})\right| \leq R\,|\theta - \theta_{old}|,$$

其中常数 R 的具体形式见定理 *C.7*。

我们列出了常数的具体形式，以便感兴趣的读者可以直接将该界用于 $\gamma = 1$ 或 $T = \infty$。

证明思路。 我们利用性能差异引理，写出

$$\nabla_\theta J(\theta) = \nabla_\theta\big(V(\pi_\theta) - V(\pi_{\theta_{\text{old}}})\big) = \nabla_\theta \sum_{t=0}^{T-1} \gamma^t \mathbb{E}^{\pi_\theta}\left[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t)\right],$$

由此可得

$$\nabla_\theta J(\theta) - g_{\text{PPO}}(\theta, \theta_{\text{old}}) = \sum_{t=0}^{T-1} \gamma^t \nabla_\theta \left(\mathbb{E}^{\pi_\theta}[g_t^\theta(S_t)] - \mathbb{E}^{\pi_{\theta_{\text{old}}}}[g_t^\theta(S_t)]\right),$$

其中 $g_t^\theta(s) := \mathbb{E}_{A\sim\pi_\theta(\cdot\,;s)}[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s, A)]$。这里 g_{PPO} 表示未裁剪的替代目标。右侧可用$\pi_{\theta_{\text{old}}}$ 与 π_θ 之间的全变差距离来估计；对于有界得分函数，该距离关于 $|\theta-\theta_{\text{old}}|$线性有界。最后，通过若干重要性比率计算将裁剪纳入该估计。完整细节请参见附录 C。 □

现在换一个角度查看图 1，可以更好地说明该图的意图。每个周期内，精确梯度近端策略优化执行一次优势演员-评论家步，并附带额外的（橙色）替代梯度；随着该方案偏离重采样点，这些梯度的偏差会增大（与最优方向的对齐程度降低）。
该定理表明，只要参数保持在采样参数附近（信赖域）内，偏差就很小，额外的代理梯度步不会损害策略梯度。因此，向优势演员-评论家加入采样的近端策略优化型代理梯度步无需额外样本，并不必然有害，而且具有多项优势。这些优势包括：用状态转移的小批次代替完整轨迹采样来实现方差缩减（减少时间相关性），以及进行更多价值网络更新（例如[37]）。

5 确定性收敛

在讨论循环随机重排随机梯度下降背景下的近端策略优化之前，我们先来探讨更简单的精确梯度情形。假设我们可以显式获取梯度 $\nabla_\theta J$，此外还能获取精确的替代目标梯度 g^{clip} PPO。我们想知道，在尝试优化 J 时，使用替代梯度是否能带来优势。为模拟稍后讨论近端策略优化时的情形，我们假设可以免费使用替代梯度，即只统计使用真实策略梯度 $\nabla_\theta J$ 的梯度步数 C。我们假设 J 是 L- 光滑的（见命题 B.5 ），并将该方法与标准梯度

上升法进行比较，其最优步长已知为 $\eta = \frac{1}{L}$。结果表明，这个问题在很大程度上取决于问题参数，例如梯度的利普希茨常数和初始化时的误差。举例来说，取 $f(x) = -x^2$，那么梯度上升一步即可收敛。额外的有偏梯度步会使收敛变差。

在实践中，光滑常数 L 是未知的，而对于 $\eta \ll \frac{1}{L}$，情况要清晰得多。在这种情形下，额外的有偏梯度步实际上可能有益。假设存在若干周期 $c = 0,...,C-1$，其周期长度为 K。更新规则为

$$\theta_{c,e+1} = \theta_{c,e} + \eta\, g^{\text{clip}}_{\text{PPO}}(\theta_{c,e}, \theta_{c,0}), \quad e = 0, \ldots, K-1$$
$$\theta_{c+1,0} = \theta_{c,K}$$

由于 $g^{\text{clip}}{}_{\text{近端策略优化}}(\theta_{c,0}, \theta_{c,0}) = \nabla_\theta J(\theta_{c,0})$，循环代理梯度上升方法会先执行正确的梯度步，随后执行偏差逐渐增大的代理梯度步（将图 1 与$K = 4$对比）。借助上升引理和定理 4.2，我们可以推导出沿参数序列的 J收敛性如下。

定理 5.1 (确定性 PPO 收敛)。假设 C 为周期数，K 为周期长度，R来自定理 *4.2*，G来自命题 *C.10*，$\Delta_0 := J_* - J(\theta_{0,0})$为初始最优性差距，且 L 是 $\nabla_\theta J$ 的利普希茨常数（J 是 L-光滑的，见命题 *B.5*）。如果学习率 η 小于 $\frac{1}{L}$，那么

$$\min_{c=0,...,C-1,\ e=0,...,K-1} |\nabla_\theta J(\theta_{c,e})|^2 \le \frac{2\Delta_0}{\eta CK} + \frac{1}{6}\eta^2 (K-1)(2K-1)R^2G^2.$$

证明见附录 D.2。选择 $K = 1$ 可恢复针对 L-光滑函数的梯度上升的标准收敛速率，其中最优步长为 $\eta = \frac{1}{L}$。对于 $K > 1$，我们考虑较宽松的上界 $\frac{2\Delta_0}{\eta CK} + \frac{1}{3}\eta^2 K^2 R^2 G^2$。固定周期长度 K进行优化，可得最优步长 $\eta_* = \min(\frac{c}{K}, \frac{1}{L})$；固定 $\eta \le \frac{1}{L}$进行优化，可得最优周期长度[1] $K_* = \frac{c}{\eta}$，且 $c = (\frac{3\Delta_0}{CR^2G^2})^{\frac{1}{3}}$。如果 $\eta^* < \frac{1}{L}$，两种情况都会导致

$$\min_{0\le c\le C-1,\ 0\le e\le K-1} |\nabla_\theta J(\theta_{c,e})|^2 \le \left(\frac{3\Delta_0 RG}{C}\right)^{\frac{2}{3}}.$$

总之，在精确梯度步预算 C 固定的情况下，额外的有偏梯度步可以改善收敛，条件是 $\eta \ll \frac{1}{L}$；或者，在步长最优（但通常未知）的情况下，若Δ_0 和 L 相较于 R 和 G足够大，也可以改善收敛。正如我们将在下一节讨论的，这正是近端策略优化的优势。估计代理梯度可以帮助弥补过小的学习率，且不会增加采样成本；它们只使用周期中第一个梯度步所生成的轨迹展开（图 1 中的蓝点）。

6 重新洗牌分析：近端策略优化的收敛

现在我们转向近端策略优化（PPO），将确定性分析中的精确代理梯度替换为使用转移缓冲区样本的采样代理梯度。我们强调，我们的贡献

[1]为简便起见，我们允许 K_* 在这里为非整数。

算法1 循环重洗牌 PPO 替代目标上升

要求： 初始 θ、步长 η、周期 C、训练轮次 K、批次大小 $B, m := N/B$。

对 $c = 0, \dots, C-1$ **执行** ▷ 周期

  $\theta_{\text{old}} \leftarrow \theta$ ▷ 固定采样参数

  在 π_θ 下采样 n 轨迹展开 $_{\text{old}}$

  估计优势 $\hat{\mathbb{A}}$ ▷ 评论家步骤，例如使用广义优势估计

  填充缓冲区 $\{(s^i,\ a^i,\ r^i,\ \hat{\mathbb{A}}^i,\ t^i)\}_{i=0}^{N-1}$

  对于 $e = 0, \dots, K-1$ **执行** ▷ 训练轮次

    生成随机排列 $\sigma = (\sigma_0,\ \dots\dots,\ \sigma_{N-1})$

    对于 $k = 0, \dots\dots, m-1$ **执行** ▷ 小批量更新

      $\mathcal{B}_k \leftarrow \{\sigma_{kB}, \dots, \sigma_{(k+1)B-1}\}$

      按照式 (1) 计算替代目标 $\hat{g}^{\text{clip}}(\theta,\ \theta_{\text{old}};\ \mathcal{B}_k)$

      $\theta \leftarrow \theta + \eta\, \hat{g}^{\text{clip}}(\theta,\ \theta_{\text{old}};\ \mathcal{B}_k)$ ▷ 这是 $\theta_{c,e,k+1}$

    结束循环

  结束循环

结束循环

主要在于概念层面。除了一些细微修改（在梯度中进行折扣，以及围绕 1 对称裁剪）之外，这只是对标准 PPO 策略更新机制的形式化。基于这一形式化，我们的主要贡献是下面的定理 6.2。

注释 6.1. 在分析中，我们假设可以使用性能良好的优势估计器。例如，在精确评论家的玩具假设下，这一抽象条件便得以满足。为便于数学处理，我们采用了尽可能弱的假设，并且该假设能一致地控制估计误差。尽管这一假设并不理想，但鉴于当前深度学习理论（在价值预测方面）的发展状况，它不可避免。

我们先改写 g^{clip} 近端策略优化$(\theta,\ \theta_{\text{old}})$。均匀地、独立于过程选择时间步 $U \sim \mathcal{U}\{0, \dots, T-1\}$。于是，替代时间总和可以写成均匀分布下的期望：

$$g^{\text{clip}}_{\text{PPO}}(\theta, \theta_{\text{old}}) = T\, \mathbb{E}_{U\sim\mathcal{U}}\Big[\mathbb{E}^{\pi_{\theta_{\text{old}}}}\Big[\gamma^U \frac{\nabla_\theta \pi_\theta(A_U; S_U)}{\pi_{\theta_{\text{old}}}(A_U; S_U)} \mathbb{1}_{\left|\frac{\pi_\theta(A_U;S_U)}{\pi_{\theta_{\text{old}}}(A_U;S_U)}-1\right| \le \epsilon}\, \mathbb{A}_U^{\pi_{\theta_{\text{old}}}}(S_U, A_U)\Big]\Big],$$

也就是说，它是对均匀采样的时间索引和在 $\pi_{\theta_{\text{old}}}$ 下的马尔可夫决策过程取双重期望。实践中，这个联合期望按周期利用采样转移进行近似。在一个周期内，固定采样参数 θ_{old}，在 $\pi_{\theta_{\text{old}}}$ 下收集长度为 T 的 n 条轨迹展开，并使用截断（或有限时间）广义优势估计计算所有优势（见第 7 节）。接下来，将得到的数据展平为一个大小为 N 的转移缓冲区 $\{(s^i,\ a^i,\ r^i,\ \hat{\mathbb{A}}^i,\ t^i)\}_{i=0}^{N-1}$: $= nT$，其中 $(s^i,\ a^i,\ r^i,\ t^i)$ 遍历轨迹展开中遇到的所有 $\hat{}$ 状态-动作-奖励-时间元组，而 $\mathbb{A}_i$ 表示根据轨迹为 $\mathbb{A}_{t_i}^{\pi_{\theta_{\text{old}}}}(s^i,\ a^i)$ 计算出的优势估计（例如，实践中使用广义优势估计）。注意，我们在标准转移缓冲区中附加了状态转移的时间索引，以便对梯度进行折扣。这不会造成任何实际困难。

在一个周期内，PPO 实现会使用重新洗牌后的小批量，对转移缓冲区进行多次遍历。我们将这一机制形式化如下。设 $\sigma = (\sigma_0, \dots, \sigma_{N-1})$ 是 $\{0, \dots, N-1\}$ 的一个随机排列（重新洗牌），并将排列后的索引划分为大小为 B 的连续小批量：对于 $k = 0, \dots, m-1$，其中 $m := \frac{N}{B}$，定义 $\mathcal{B}_k := \{\sigma_{kB}, \dots, \sigma_{(k+1)B-1}\} \subseteq$

$\{1,\dots,N\}$。单个 PPO 更新步骤使用从小批量采样得到的替代目标梯度

$$\hat{g}^{\mathrm{clip}}(\theta,\theta_{\mathrm{old}};\mathcal{B}_k)=\frac{1}{B}\sum_{i\in\mathcal{B}_k} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta,\theta_{\mathrm{old}}), \tag{1}$$

其中，每个状态转移的贡献为

$$g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta,\theta_{\mathrm{old}}):=T\gamma^{t^i}\frac{\nabla_\theta\pi_\theta(a^i;s^i)}{\pi_{\theta_{\mathrm{old}}}(a^i;s^i)}\,\mathbb{1}_{\left|\frac{\pi_\theta(a^i;s^i)}{\pi_{\theta_{\mathrm{old}}}(a^i;s^i)}-1\right|\le\epsilon}\hat{\mathbb{A}}^i.$$

一个训练轮次对应于对缓冲区的一次遍历，即使用 m 步，依次遍历由 σ生成的索引批次 $\mathcal{B}_0,\dots,\mathcal{B}_{m-1}$。在同一个周期内，近端策略优化会重复此过程 K 个训练轮次：每个训练轮次开始时都会重新抽取一个排列，然后遍历数据。为方便读者，我们在算法1中给出了对近端策略优化策略更新的解读伪代码。

上述过程生成了一系列参数 $\theta_{c,e,k}$ ，其索引由周期 c、训练轮次 e 以及该训练轮次内第k次小批量更新确定。下文中，我们用 $\theta_{c,e,k}$ 表示周期 c的训练轮次 e 中第（$k+1$）次小批量更新之前的参数。特别地，$\theta_{c,0,0}$ 是周期 c 的初始化，并在该周期中起到 θ_{old} 的作用；$\theta_{c,e,0}$ 是训练轮次的起始点，而 $\theta_{c,e,m}$ 是训练轮次的终点。因此，近端策略优化可以视为一种周期性随机重排方法。回顾一下，随机重排是一种用于有限和目标函数的随机梯度下降式方法：在每个训练轮次开始时，先将数据点打乱，然后使用每个数据点各进行一次小批量梯度步。与每一步都独立地有放回抽取索引的随机梯度下降不同，随机重排在一个训练轮次内采用无放回抽样，这通常可以减少冗余并改善收敛。关于随机重排在监督学习中的收敛结果（这些结果启发了我们的收敛证明），请参见 [18] 及其中的参考文献。下面给出近端策略优化的主要收敛结果：

定理 6.2。 假设满足假设 *3.1*，且学习率 η 小于 $\frac{1}{Lm}$。那么，对于任意 $p,q\in(0,\,1)$，都有

$$\begin{aligned}\min_{c=0,\dots,C-1,\ e=0,\dots,K-1}\mathbb{E}\big[|\nabla J(\theta_{c,e,0})|^2\big]\le{}&\frac{2\Delta_0}{\eta CKm}+6\eta^2(7B_1^2+L^2)K^2m^2G^2\\&+42\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p}\\&+\frac{42\eta^qB_3^2|\mathcal{A}|^q\Pi_*^qK^qm^qG^q}{\epsilon^q}+\frac{12\sigma^2}{N}+12T^2\Pi_*^2\delta^2\,,\end{aligned} \tag{2}$$

其中的常数取自定理 *5.1*，其他常数见附录*D* *.4.*

与第 5 节不同，随机情形带来了需要更加谨慎处理的挑战。最值得注意的是，迭代点是随机的，并且随机依赖于每个周期开始时收集的样本。因此，定理 4.2 中所建立的确定性分析所使用的偏差项不再能够直接处理，因为现在这两个量都是随机的，并且通过采样过程相互耦合。为克服这一问题，我们的分析转而直接处理每个周期内生成的采样梯度。我们不再将随机迭代点处计算的精确替代梯度与真实梯度进行比较，而是比较中间步骤处的采样梯度与周期开始时的采样梯度。这种路径级偏差分解（见引理 D.3）使我们能够控制新鲜采样所引入的依赖，同时保留对梯度一致性的有意义刻画。

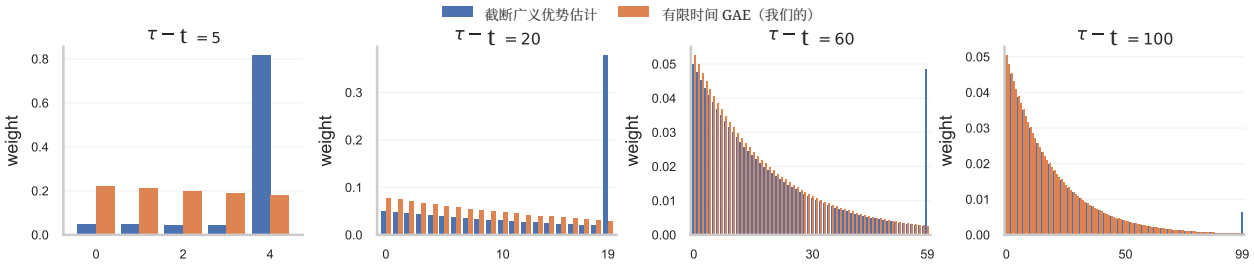


图 2：**GAE 尾部质量塌缩**。比较截断广义优势估计与我们的有限时间（重新归一化）估计器分配给 k 步优势估计量 $\hat{\mathbb{A}}_t^{(k)}$ 的权重。每个面板对应不同的剩余时域 $\tau - t$，并展示了对于截断广义优势估计，几何尾部权重如何塌缩到最后一个可用估计器 $\hat{\mathbb{A}}_t^{(\tau-t-1)}$ 上，而有限时间变体则通过有限时域重新归一化将该质量重新分配到可观测的 k 步估计器上。

设定 $p = q = 1$ 并假设 $\eta \leq \frac{1}{Lm}$，我们可以将式 (2) 中的上界改写为

$$\frac{2\Delta_0}{\eta CKm} + c_1\eta^2K^2m^2 + \frac{c_2\eta Km}{\epsilon} + \frac{12\sigma^2}{N} + 12T^2\Pi_*^2\delta^2$$

对于适当的常数 c_1，c_2。为了更好地量化该界在小步长下的表现，我们平衡 $\frac{1}{\eta}$ 与 η 的项（忽略 η^2），从而得到合适的周期大小

$$K = \frac{1}{\eta m}\sqrt{\frac{2\Delta_0\epsilon}{Cc_2}}$$

以及相应的上界

$$\min_{c=0,\ldots,C-1,\ e=0,\ldots,K-1} \mathbb{E}\left[|\nabla J(\theta_{c,e,0})|^2\right] \leq 2\sqrt{\frac{2\Delta_0 c_2}{C\epsilon}} + \frac{2c_1\Delta_0\epsilon}{Cc_2} + \frac{12\sigma^2}{N} + 12T^2\Pi_*^2\delta^2 .$$

与确定性情形一样，结果表明，基于周期的更新方案降低了对步长选择的敏感性。较小的学习率可以通过复用相同轨迹采样进行额外更新来弥补，而不会削弱收敛保证。这种行为可以解释为一种隐式信赖域机制，其中许多小幅裁剪更新自适应地控制有效步长。

7 有限时间 GAE

在收敛理论中，我们需要基于抽象的评论家假设进行分析。本节将揭示 PPO 中的理论与实现差距（参见 [30] 中的 (11)、(12)），该差距是在截断原始广义优势估计器时产生的。详情和证明见附录 E。

回顾一下，无限时域马尔可夫决策过程中的原始 GAE 如下所示。受 k 步贝尔曼期望算子的启发，k 步前向估计量 $\mathbb{A}_t^{(k)} := \sum_{t=0}^{k} \gamma^t R_t + \gamma^{k+1} V^\pi(S_{t+k+1}) - V^\pi(S_t)$ 是 $\mathbb{A}^\pi(S_t, A_t)$ 的条件无偏估计量。用价值网络近似 V 替代真实价值函数后，该估计量记为 $\hat{\mathbb{A}}_t^{(k)}$。当 k 较大时（接近蒙特卡罗优势估计量），估计方差占主导；当 k 较小时，价值函数

近似偏差占主导。对 k 步估计量进行几何混合，得到 GAE $\hat{\mathbb{A}}_t^{\infty} := (1-\lambda)\sum_{k=0}^{\infty}\lambda^k \hat{\mathbb{A}}_t^{(k)}$。有一个简单却重要的技巧，使 GAE 尤具吸引力。利用伸缩和消去可得 $\hat{\mathbb{A}}_t^{\infty} = \sum_{\ell=0}^{\infty}(\gamma\lambda)^{\ell}\,\delta_{t+\ell}$，其中时序差分误差为 $\delta_t = R_t + \gamma V(S_{t+1}) - V(S_t)$。对于有限时域马尔可夫决策过程（甚至终止马尔可夫决策过程），无限时域设定并不适用。在近端策略优化中（见 [30] 中的 (11)），GAE 通常通过在轨迹展开结束后丢弃时序差分误差来截断 τ:

$$\hat{\mathbb{A}}_t := \sum_{\ell=0}^{\tau-t-1}(\gamma\lambda)^{\ell}\,\delta_{t+\ell}.$$

虽然在 PPO 中 $\tau = T$ 被视为固定值，但截断同样可以应用于终止时刻。截断表示尤其有用，因为它允许使用终端条件 $\mathbb{A}_\tau := 0$回溯 $\mathbb{A}_t = \delta_t + \gamma\lambda\mathbb{A}_{t+1}$。GAE 的思想是以权重 $(1-\lambda)\lambda^k$对 k-步优势估计器进行几何混合，但在进行截断时这一点会失效。所有超过 τ 的 k-步估计器的质量都会塌缩到最长的非平凡估计器上。

命题 7.1（截断广义优势估计的尾部质量塌缩）。 对于 $t \le \tau - 1$，实际使用的广义优势估计器满足

$$\hat{\mathbb{A}}_t = \sum_{k=0}^{\tau-t-2}\underbrace{(1-\lambda)\lambda^k}_{GAE\ weights}\hat{\mathbb{A}}_t^{(k)} + \underbrace{\lambda^{\tau-t-1}}_{collapsed\ tail\text{-}mass}\hat{\mathbb{A}}_t^{(\tau-t-1)}.$$

我们将这一现象称为尾部质量塌缩，参见图 2 中的蓝色柱状条。接下来，我们提出一种新的估计器，使用经过归一化的几何权重，使其总和仅填充 $\{0, ..., \tau - t - 1\}$。

De定义 7.2（有限时间广义优势估计）。 我们将有限时间广义优势估计器定义为

$$\hat{\mathbb{A}}_t^{\tau} := \frac{1-\lambda}{1-\lambda^{\tau-t}}\sum_{k=0}^{\tau-t-1}\lambda^k\,\hat{\mathbb{A}}_t^{(k)}.$$

如果 $\tau = T$，则称该估计器为固定时刻方法；否则称为终止时间 GAE。

图 2 中的橙色条形展示了有限时间 GAE 的几何权重。通过对可用后缀 $k \in \{0, \ldots, \tau - t - 1\}$ 所支持的不同 k 步估计器上的几何权重重新归一化，我们的估计器避免了截断广义优势估计（蓝色）在轨迹展开末尾附近出现的尾部质量向 $\hat{\mathbb{A}}_t^{(\tau-t-1)}$ 强烈塌缩。

直观而言，最长前瞻项 $\hat{\mathbb{A}}_t^{(\tau-t-1)}$ 受自举的影响最小，因此其值函数近似（自举）偏差往往较小；但它通常方差较高，因为它累积了最长的折扣时序差分误差之和。因此，对于固定的 γ 和 λ，我们的有限时间重新归一化在方差、偏差和自举之间进行权衡。同时，它恢复了 GAE 几何混合解释预期的有限时域类比，而不是隐式地将未观测到的尾部质量塌缩到单个估计器上。

至于 截断广义优势估计，我们的有限时间 GAE 同样满足简单的反向递归 :

$\hat{}$**命题 7.3**。 使用终端条件 $\mathbb{A}_\tau^{\tau} := 0$，有限时间估计器满足以下递归

$$\hat{\mathbb{A}}_t^{\tau} = \delta_t + \gamma\lambda\,\frac{1-\lambda^{\tau-t-1}}{1-\lambda^{\tau-t}}\,\hat{\mathbb{A}}_{t+1}^{\tau}, \quad t = \tau-1, \ldots, 0.$$

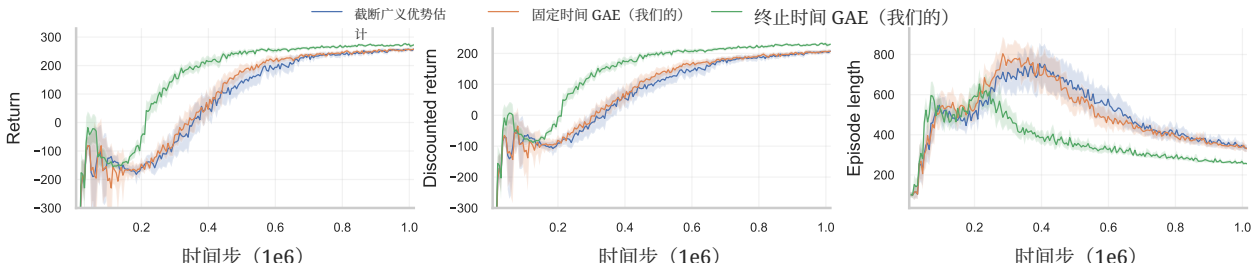


图 3：采用 Stable-Baselines3 Zoo 默认 PPO超参数的 LunarLander-v3 学习曲线。左（中）：平均评估（折扣）回报，即每个回合的（折扣）奖励之和。右：平均评估回合长度。曲线对 20 个随机种子取平均，阴影区域为随机种子的标准误。两种方法使用完全相同的默认超参数；仅优势估计器不同（截断广义优势估计 vs. 我们的有限时间广义优势估计）。其他指标与消融实验见附录 E

为展示如何简单地适配截断广义优势估计，我们在算法 2 中提供了伪代码。更多实现细节请参见附录 E。

在附录 E.9 中，我们进行了一个简化的玩具示例计算，以理解尾部质量坍缩重加权的方差效应。结果表明，在回合结束附近，广义优势估计内部的协方差会降低，见图 4，因此我们的有限时间 GAE 估计器在严重依赖回合结束的环境中应当是有益的，例如 Lunar Lander，其中着陆前不久的动作至关重要。

实验：我们在 LunarLander-v3 上评估这一效应，使用 Stable-Baselines3 的 PPO 实现 [24]，并仅修改优势估计（如算法 2 所示）。在图 3 中，我们报告了标准超参数设置下的开箱即用结果，比较了截断广义优势估计（蓝色）、带 $\tau = T = 1000$ 的固定时刻方法变体（绿色），以及带$\tau = \min\{T, \text{termination}\}$ 的终止时间变体（橙色）。学习曲线表明，终止时间估计器的学习速度明显更快。它更早达到高回报，并实现更短的回合长度（更快着陆）。一个合理的解释是，终止感知变体恰好在 $\tau - t$ 较小的高影响区间降低了估计的方差，从而产生更稳定的策略更新和更快的学习。相比之下，固定时域GAE 的表现与截断广义优势估计相似，这与理论一致。附录 E 提供了鲁棒性检验，包括针对每个估计器分别优化超参数的实验。作为健全性检查，我们在附录 E.8 中对 Ant 进行了一项小规模实验；有限时间 GAE 的表现也非常好。

算法 2 有限时间 GAE（我们的方法）

要求： 策略 π_θ、价值估计 V、折扣 γ、GAE参数 λ

使用 $v_t \leftarrow V(s_t)$ 为 $t = 0, \ldots, \tau$ 生成轨迹展开 $\{(s_t, a_t, r_t, v_t)\}_{t=0}^{\tau-1}$

$\hat{\mathbb{A}}_\tau^\tau \leftarrow 0$ ▷ 边界条件

对于 $t = \tau - 1, \ldots, 0$ **do**

  $\delta_t \leftarrow r_t + \gamma v_{t+1} - v_t$

  $\hat{\mathbb{A}}_t^\tau \leftarrow \delta_t + \gamma\lambda \frac{1-\lambda^{\tau-t-1}}{1-\lambda^{\tau-t}} \hat{\mathbb{A}}_{t+1}^\tau$

循环结束

返回 $\{\hat{\mathbb{A}}_t^\tau\}_{t=0}^{\tau-1}$

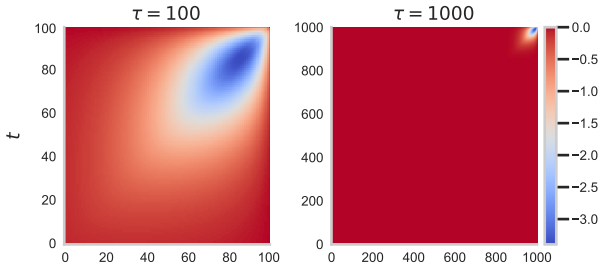


图 4：在简化假设下（见附录 E），固定 $\gamma = 0.999$、$\lambda = 0.95$ 和 $\tau = 200$（左）以及 $\tau = 1000$（右）时，截断广义优势估计（PPO）与有限时间 GAE（本文方法）之间的协方差热图差异。色标表示 $\mathrm{Cov}[\mathbb{A}_t^T, \mathbb{A}_s^T] - \mathrm{Cov}[\mathbb{A}_t, \mathbb{A}_s]$。

8 结论与未来工作

本文在通常的策略梯度假设下，弥补了 PPO 理论方面的空白。所有出现的常数都非常大，因此应将其视为帮助理解结构的依据，而非直接的实践见解（策略梯度理论一贯如此）。我们进行了偏差分析（定理 4.2），由此可以在精确梯度情形（定理 5.1）以及采用 RR 的原始 PPO 情形（定理 6.2）中推导出收敛结论。估计结果揭示了一个事实：额外的有偏 PPO 更新能够改善学习效果。PPO 通过增加额外的（无成本）有偏梯度步，来弥补较小（更安全）步长带来的影响。尽管这是一项理论研究，我们也发现，实践中使用的截断广义优势估计存在尾部质量塌缩。令人欣喜的是，只需对 GAE 做微小改动，就能显著改善例如 Lunar Lander 训练（以及 Ant 训练）的效果。鉴于问题的难度以及技术论证篇幅较长，我们将进一步的工作留待未来开展。

当前，人们对策略梯度算法的严格收敛性研究抱有浓厚兴趣。PPO 的有偏策略梯度解释为优化理论打开了大门，同时也清晰展示了如何应用随机论证，例如随机重排中的相关方法。我们相信，本文或许能够开启一些有趣的未来工作。(i) 正则化是一个尤其活跃的研究领域。由于我们对 PPO 的解释与策略梯度理论较为接近，因此将 KL 正则化加入分析似乎是合理的。(ii) 鉴于人们对不使用评论家网络的 PPO 变体日益感兴趣，研究我们的分析如何应用于 GRPO 的变体将会很有意义。(iii) 哪种非对称裁剪可以得到形式化分析？能否理解其中的形式差异？(iv) 我们认为，定理 6.2 中的分析还可以改进。首先，可以证明多轨迹展平的方差缩减效应；其次，在随机重排分析中减少与周期起点的比较。

对于有限时间 GAE，后续工作将开展全面的实验研究，以了解我们的有限时间 GAE 在何种情况下优于或逊于截断广义优势估计。在理论方面，若能通过玩具示例量化 GAE 中的偏差—方差权衡，将会很有意思。

参考文献

[1] Alekh Agarwal、Sham M. Kakade、Jason D. Lee 和 Gaurav Mahajan。《策略梯度方法的理论：最优性、近似与分布偏移》。《机器学习研究杂志》，22(98):1–76，2021。

[2] 字节跳动 Seed。《截断近端策略优化》，2025年。arXiv:2506.15050。

[3] Pablo Samuel Castro。《强化学习研究中的形式化与实现差距》，2025年。[4] Fengdi Che、Gautham Vasan 和 A. Rupam Mahmood。《同策略策略梯度方法中的折扣因子不匹配校正》。ICML'23，2023年。[5] Paul F Christiano、Jan Leike、Tom Brown、Miljan Martic、Shane Legg 和 Dario Amodei。《从人类偏好中进行深度强化学习》。载于神经信息处理系统进展，第4299–4307页，2017年。[6] Logan Engstrom、Andrew Ilyas、Shibani Santurkar、Dimitris Tsipras、Firdaus Janoos、Larry Rudolph 和 Aleksander Madry。《实现对深度策略梯度的影响：关于近端策略优化和信赖域策略优化的案例研究》。载于国际学习表征会议，2020年。[7] Ilyas Fatkhullin、Anas Barakat、Anastasia Kireeva 和 Niao He。《随机策略梯度方法：提升费舍尔非退化策略的样本复杂度》。载于第*40*届国际机器学习会议论文集，第202卷机器学习研究论文集，第9827–9869页。PMLR，2023年7月23日至29日。[8] Shengyi Huang、Anssi Kanervisto、Antonin Raffin、Weixun Wang、Santiago Ontañón 和 Rousslan Fernand Julien Dossa。《优势演员-评论家是近端策略优化的一种特殊情况》，2022年。[9] Ruinan Jin、Shuai Li 和 Baoxiang Wang。《关于裁剪近端策略优化的驻点收敛》。载于 B. Kim、Y. Yue、S. Chaudhuri、K. Fragkiadaki、M. Khan 和 Y. Sun 编，国际学习表征会议，第202卷，第11594–11611页，2024年。[10] Sham Kakade 和 John Langford。《近似强化学习的近似最优性》。载于第十九届国际机器学习会议论文集，ICML '02，第267–274页，美国加利福尼亚州旧金山，2002年。Morgan Kaufmann出版社。[11] Sara Klein、Simon Weissmann 和 Leif Döring。《超越平稳性：随机 softmax 策略梯度方法的收敛分析》。国际学习表征会议，2024年。[12] Harshat Kumar、Alec Koppel 和 Alejandro Ribeiro。《函数逼近下强化学习演员-评论家方法的样本复杂度》。机器学习，112(7):2433–2467，2023年2月。[13] David Asher Levin。马尔可夫链与混合时间。ProQuest Ebook Central。罗得岛州普罗维登斯，第二版，2017年。[14] Sergey Levine、Chelsea Finn、Trevor Darrell 和 Pieter Abbeel。《深度视觉运动策略的端到端训练》。机器学习研究期刊，17(1):1334–1373，2016年。[15] Boyi Liu、Qi Cai、Zhuoran Yang 和 Zhaoran Wang。《神经信赖域/近端策略优化可实现全局最优策略》。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Garnett 编，神经信息处理系统进展，第32卷。Curran Associates, Inc.，2019年。[16] Yin Liu、Qiming Dai、Junyu Zhang 和 Zaiwen Wen。《裁剪近端策略优化的非渐近全局收敛》，2025年。arXiv:2512.16565。

[17] Jincheng Mei、Chenjun Xiao、Csaba Szepesvari 和 Dale Schuurmans。关于 Softmax 策略梯度方法的全局收敛速率。发表于第*37*届国际机器学习会议，收入机器学习研究论文集第119卷，第6820–6829页。PMLR，2020年7月13日至18日。[18] Konstantin Mishchenko、Ahmed Khaled 和 Peter Richtárik。随机重排：简明分析与显著改进。发表于神经信息处理系统进展，33：17309–17320，2020年。[19] Volodymyr Mnih、Adria Puigdomenech Badia、Mehdi Mirza、Alex Graves、Timothy Lillicrap、Tim Harley、David Silver 和 Koray Kavukcuoglu。用于深度强化学习的异步方法。发表于国际机器学习会议，第1928–1937页。PMLR，2016年。[20] OpenAI，：Christopher Berner、Greg Brockman、Brooke Chan、Vicki Cheung、Przemysław Dębiak、Christy Dennison、David Farhi、Quirin Fischer、Shariq Hashme、Chris Hesse、Rafal Józe- fowicz、Scott Gray、Catherine Olsson、Jakub Pachocki、Michael Petrov、Henrique P. d. O. Pinto、Jonathan Raiman、Tim Salimans、Jeremy Schlatter、Jonas Schneider、Szymon Sidor、Ilya Sutskever、Jie Tang、Filip Wolski 和 Susan Zhang。利用大规模深度强化学习玩 Dota 2，2019年。[21] OpenAI、Ilge Akkaya、Marcin Andrychowicz、Maciek Chociej、Mateusz Litwin、Bob McGrew、Alex Petron、Alex Paino、Matthias Plappert、Glenn Powell 等。用机器人手解魔方。*arXiv* 预印本 *arXiv:1910.07113*，2019年。[22] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。训练语言模型，使其遵循人类反馈中的指令。发表于神经信息处理系统进展，35：27730–27744，2022年。[23] Matteo Papini、Matteo Pirotta 和 Marcello Restelli。平滑策略与安全策略梯度。机器学习，111(11)：4081–4137，2022年11月。[24] Antonin Raffin、Ashley Hill、Adam Gleave、Anssi Kanervisto、Maximilian Ernestus 和 Noah Dormann。Stable-Baselines3：可靠的强化学习实现。*J. Mach. Learn.Res.*，22(1)，2021年1月。[25] Samuel Robertson、Thang Chu、Bo Dai、Dale Schuurmans、Csaba Szepesvari 和 Jincheng Mei。使用任意学习率的 REINFORCE 都会收敛到最优策略。发表于神经信息处理系统进展（*NeurIPS*），2025年。[26] Nicolas Le Roux、Marc G. Bellemare、Jonathan Lebensold、Arnaud Bergeron、Joshua Greaves、Alex Fréchette、Carolyne Pelletier、Eric Thibodeau-Laufer、Sándor Toth 和 Sam Work。渐缩离策略 REINFORCE：面向大语言模型的稳定高效强化学习，2025年。[27] Itay Safran 和 Ohad Shamir。随机打乱下的随机梯度下降效果如何？发表于计算学习理论年会，2019年。[28] John Schulman、Sergey Levine、Philipp Moritz、Michael Jordan 和 Pieter Abbeel。信赖域策略优化。发表于第*32*届国际机器学习会议论文集——国际机器学习会议第*37*卷，ICML'15，第1889–1897页。JMLR.org，2015年。

[29] John Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel。使用广义优势估计进行高维连续控制，2018年。[30] 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年。[31] David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driess- che、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot 等。用深度神经网络和树搜索掌握围棋。自然，529(7587)：484–489，2016年。[32] Xiulei Song、Yizhao Jin、Greg Slabaugh 和 Simon Lucas。用于近端策略优化的部分优势估计器，2023年。[33] Richard S Sutton 和 Andrew G Barto。强化学习：导论。麻省理工学院出版社，2018年。[34] Richard S Sutton、David McAllester、Satinder Singh 和 Yishay Mansour。采用函数逼近的强化学习策略梯度方法。发表于神经信息处理系统进展，第1057–1063页，1999年。[35] Philip S. Thomas。自然演员-评论家算法中的偏差。发表于第*31*届国际机器学习会议论文集——国际机器学习会议第*32*卷，ICML'14，第I–441–I–448页。JMLR.org，2014年。[36] Emanuel Todorov、Tom Erez 和 Yuval Tassa。MuJoCo：用于基于模型控制的物理引擎。发表于*2012*年*IEEE/RSJ*智能机器人与系统国际会议，第5026–5033页，2012年。[37] Tao Wang、Ruipeng Zhang 和 Sicun Gao。改进价值估计可显著增强原始策略梯度，2025年。[38] Ronald J Williams。用于联结主义强化学习的简易统计梯度跟随算法。机器学习，8(3)：229–256，1992年。[39] Rui Yuan、Robert M. Gower 和 Alessandro Lazaric。原始策略梯度的一般样本复杂度分析。发表于Gustau Camps-Valls、Francisco J. R. Ruiz 和 Isabel Valera 编，第*25*届人工智能与统计国际会议论文集，收入机器学习研究论文集第151卷，第3332–3380页。PMLR，2022年3月28日至30日。[40] Kaiqing Zhang、Alec Koppel、Hao Zhu 和 Tamer Başar。策略梯度方法全局收敛至（几乎）局部最优策略。*SIAM*控制与优化期刊，58(6)：3586–3612，2020年。[41] Shangtong Zhang、Romain Laroche、Harm van Seijen、Shimon Whiteson 和 Remi Tachet des Combes。深入探究演员-评论家算法中的折扣不匹配，2022年。

A 符号与初步结果

我们先固定一些符号。我们用 $|\cdot|$ 表示欧几里得范数，用 $|\cdot|_\infty$ 表示$\mathbb{R}^d$上的最大范数，用 $\|\cdot\|_\infty$ 表示状态和/或动作空间上的最大范数。梯度 $\nabla_\theta \pi_\theta(s,a)$是指关于策略参数的导数。为避免混淆，我们有时会从梯度中省略标识符 θ。回想一下，我们考虑的是折扣有限时域马尔可夫决策过程，其价值函数

$$V^\pi(s) \coloneqq \mathbb{E}_s^\pi\Big[\sum_{t=0}^{T-1} \gamma^t R_t \mid S_0 = s\Big]$$

and

$$V_t^\pi(s) \coloneqq \mathbb{E}^\pi\Big[\sum_{i=t}^{T-1} \gamma^{i-t} R_i \mid S_t = s\Big] \quad \text{and} \quad Q_t^\pi(s,a) \coloneqq \mathbb{E}^\pi\Big[\sum_{i=t}^{T-1} \gamma^{i-t} R_i \mid S_t = s, A_t = a\Big].$$

为完整起见，我们还设定 $V_T^\pi \equiv 0$ 和 $Q_T^\pi \equiv 0$。我们还将优势函数定义为 $\mathbb{A}_t^\pi(s,a) \coloneqq Q_t^\pi(s,a) - V_t^\pi(s)$，并用 $d^{\pi,s'}_t$ 表示边缘状态分布，即

$$d_t^{\pi,s'}(s) = \mathbb{P}_{s'}^\pi(S_t = s).$$

在本节中，我们将优势函数视为已知，就如同我们能够使用一个完美的评论家。我们将始终使用连续可微的参数化策略族 $\{\pi_\theta\}_{\theta\in\mathbb{R}^d}$，并将 $J(\theta) = V^{\pi_\theta}(\mu) \coloneqq \sum_s V^{\pi_\theta}(s)\mu(s)$以及 $d_t^\pi \coloneqq d^{\pi,\mu}_t$简记为针对某个固定初始状态分布 μ的形式。我们将始终假设

$$\pi_\theta(a\,;\,s) > 0 \quad \text{for all } \theta \in \mathbb{R}^d, (s,a) \in \mathcal{S}\times\mathcal{A}, \tag{3}$$

以确保似然比定义良好。通常，最后应用一次 Softmax 归一化即可满足这一条件。

为了证明近端策略优化的收敛性，我们需要对底层马尔可夫决策模型（有界奖励）和策略（得分函数有界且利普希茨连续）作出假设，这些假设将推出参数化价值函数的 L-光滑性，见命题 B.5。这些假设在策略梯度方法的收敛性分析中是标准的。

假设 A.1（有界奖励）。 奖励的绝对值一致有界，且上界为 R_*。

请注意，在假设 A.1 下，价值函数、 Q-函数和优势函数也都是有界的。对我们而言最相关的是，对所有 $t \le T-1$，都有 $||\mathbb{A}_t^\pi||_\infty \le \frac{1-\gamma^T}{1-\gamma} 2R_*$。我们在确定性设置中使用这一界。在随机分析中，我们假设可以获得有偏且有界的优势估计。

假设 A.2（有偏且有界的优势估计）。 存在常数 $A_* < \infty$ 和 $\delta \ge 0$，使得对于任意 θ 以及每个 $t \in \{0,\dots,T-1\}$，我们都可以获得满足以下条件的优势估计 $\hat{\mathbb{A}}_t$

$$\mathbb{E}^{\pi_\theta}[|\mathbb{E}^{\pi_\theta}\big[\hat{\mathbb{A}}_t \mid S_t, A_t\big] - \mathbb{A}_t^{\pi_\theta}(S_t,A_t)|^2] \le \delta^2 \qquad \text{and} \qquad \big|\hat{\mathbb{A}}_t\big| \le A_* \quad \text{a.s.}$$

假设可以访问理论评论家，即可轻松满足无偏且有界的优势估计器假设。

假设 A.3（有界得分函数）。 得分函数是有界的，即

$$\Pi_* := \sup_\theta ||\nabla_\theta \log(\pi_\theta)||_\infty < \infty.$$

注意，有界得分函数意味着有界梯度，因为

$$|\nabla_\theta \pi_\theta(a\,;\,s)| = \pi_\theta(a\,;\,s)|\nabla_\theta \log \pi_\theta(a\,;\,s)| \le \Pi_*,$$

并且根据中值定理，策略具有利普希茨连续性。

假设 A.4（利普希茨得分函数）。 存在 $L > 0$，使得对于所有（s, a）$\in \mathcal{S} \times \mathcal{A}$以及所有 $\theta, \theta' \in \mathbb{R}^d$,

$$|\nabla \log \pi_\theta(a\,;\,s) - \nabla \log \pi_{\theta'}(a\,;\,s)| \le L_{\mathrm{s}}|\theta - \theta'|.$$

例如，我们可参阅 [40]第 7 页，了解满足这些假设的示例策略。

下文中，我们用 TV 表示全变差距离

$$\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{a\in\mathcal{A}} \big|P(a) - Q(a)\big| = \sup_{B\subseteq\mathcal{A}} \big|P(B) - Q(B)\big| = \frac{1}{2}\sup_{||f||_\infty \le 1} \Big|\int_{\mathcal{A}} f dP - \int_{\mathcal{A}} f dQ\Big| \quad (4)$$

对于有限动作空间 $\mathcal{A}$上的概率测度 P 和 Q。

引理 A.5. 在假设 *A.3* 下，有

$$\mathrm{TV}\big(\pi_\theta(\,\cdot\,;\,s), \pi_{\theta'}(\,\cdot\,;\,s)\big) \le \tfrac{1}{2}\Pi_*\,|\theta - \theta'|, \quad \textit{for all } \theta, \theta' \in \mathbb{R}^d \textit{ and } s \in \mathcal{S}. \quad (5)$$

证明。 假设 A.3 与 $\nabla_\theta \pi_\theta(a\,;\,s) = \pi_\theta(s\,;\,a)\nabla_\theta \log\, \pi_\theta(s;\,a)$ 一起推出，对所有 $s \in \mathcal{S}$ 均有

$$\begin{aligned}
\mathrm{TV}(\pi_\theta(\,\cdot\,;\,s), \pi_{\theta'}(\,\cdot\,;\,s)) &= \frac{1}{2}\sum_a \big|\pi_\theta(a\,;\,s) - \pi_{\theta'}(a\,;\,s)\big| \\
&\le \frac{1}{2}\sum_a \int_0^1 |\nabla \pi_{\varphi(t)}(a\,;\,s)^\top(\theta - \theta')|\,dt \\
&\le \frac{1}{2}\int_0^1 \underbrace{\sum_a \pi_{\varphi(t)}(a\,;\,s)}_{=1}\,\Pi_*|\theta - \theta'|\,dt = \frac{1}{2}\Pi_*|\theta - \theta'|,
\end{aligned} \quad (6)$$

其中 $\varphi(t) = (1-t)\theta + t\theta'$。 □

B 参数化价值函数的性质

本节汇总价值函数的基本性质，尤其是其 L-光滑性。尽管文献中对参数化价值函数的光滑性已有充分研究，但现有证明通常依赖略强的假设，或针对无限时域或有限时域非折扣马尔可夫决策过程；例如，参见 [1, 39, 23, 7]。为方便读者，我们给出自包含的证明，这些证明与所引文献中的证明有所不同。下文还将运用这里提出的技术，证明近端策略优化的梯度偏差估计。

首先，我们回顾折扣有限时域马尔可夫决策过程的标准策略梯度定理:

命题 B.1. 在假设 *A.1* 和 *A.3* 下，价值函数关于策略参数的梯度为

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}^{\pi_\theta}\Big[\nabla_\theta \log\big(\pi_\theta(A_t\,;\,S_t)\big)\mathbb{A}_t^{\pi_\theta}(S_t, A_t)\Big]. \tag{7}$$

引理 B.2 （ *J*的利普希茨连续性）。 在假设 *A.1* 和 *A.3* 下，对于所有 $t \in\{0,\dots,T\}$、 $s \in \mathcal{S}$ 以及 θ、$\theta' \in \mathbb{R}^d$，有

$$\big|V_t^{\pi_\theta}(s) - V_t^{\pi_{\theta'}}(s)\big| \le \frac{1-(T-t+1)\gamma^{T-t}+(T-t)\gamma^{T-t+1}}{(1-\gamma)^2}\,\Pi_* R_*\,|\theta-\theta'|. \tag{8}$$

特别地，

$$\big|J(\theta) - J(\theta')\big| \le \frac{1-(T+1)\gamma^{T}+T\gamma^{T+1}}{(1-\gamma)^2}\,\Pi_* R_*\,|\theta-\theta'|.$$

证明。 证明通过对 t 进行反向归纳来完成。对于 $t=T$，由于 $V_T \equiv 0$，该结论成立。假设该界在时刻 $t+1$成立。为使用归纳假设，我们应用（有限时间）贝尔曼递推

$$V_t^{\pi_\theta}(s) - V_t^{\pi_{\theta'}}(s) = (T_{\pi_\theta}V_{t+1}^{\pi_\theta})(s) - (T_{\pi_{\theta'}}V_{t+1}^{\pi_{\theta'}})(s),$$

其中（$T_\pi V$）（s）$= \mathbb{E}_{a\sim\pi(\cdot\,;\,s),\ s'\sim p(\cdot\,;\ s,\ a)}[r(s,\ a) + \gamma V(s')]$。现在我们将差分解为

$$|V_t^{\pi_\theta}(s) - V_t^{\pi_{\theta'}}(s)| \le \underbrace{|T_{\pi_\theta}(V_{t+1}^{\pi_\theta} - V_{t+1}^{\pi_{\theta'}})(s)|}_{\text{(A)}} + \underbrace{|(T_{\pi_\theta} - T_{\pi_{\theta'}})V_{t+1}^{\pi_{\theta'}}(s)|}_{\text{(B)}}.$$

由于T_{π_θ} 是最大范数压缩映射，我们得到

$$\text{(A)} \le \gamma\|V_{t+1}^{\pi_\theta} - V_{t+1}^{\pi_{\theta'}}\|_\infty.$$

为处理项（B），我们显式使用贝尔曼算子。根据定义

$$(T_\pi V)(s) = \sum_{a\in\mathcal{A}} \pi(a\,;\,s)\big(\mathbb{E}[r(s,a)] + \gamma\mathbb{E}_{s'\sim p(\cdot\,;\,s,a)}[V(s')]\big).$$

因此，

$$(T_{\pi_\theta} - T_{\pi_{\theta'}})V_{t+1}^{\pi_{\theta'}}(s) \sum_a \big(\pi_\theta(a\,;\,s) - \pi_{\theta'}(a\,;\,s)\big)\big(\mathbb{E}[r(s,a)] + \gamma\mathbb{E}_{s'\sim p(\cdot\,;\,s,a)}[V_{t+1}^{\pi_{\theta'}}(s')]\big).$$

由于奖励有界，对于所有 a，有

$$\big|\mathbb{E}[r(s,a)] + \gamma\mathbb{E}_{s'\sim p(\cdot\,;\,s,a)}[V_{t+1}^{\pi_{\theta'}}(s')]\big| \le R_* \sum_{j=0}^{T-t-1}\gamma^j.$$

回顾式 (6)，$\sum_a\big|\pi_\theta(a;\ s) - \pi_{\theta'}(a\,;\ s)\big| \le \Pi_*|\theta-\theta'|$，则可得

$$\text{(B)} \le \Big(\sum_a |\pi_\theta(a\,;\,s) - \pi_{\theta'}(a\,;\,s)|\Big)\Big(R_*\sum_{j=0}^{T-t-1}\gamma^j\Big) \le \Pi_*|\theta-\theta'|\Big(R_*\sum_{j=0}^{T-t-1}\gamma^j\Big).$$

将这些结果合并代入递推关系，即可得到关于 t 的结论，其中恒等式 (8) 可通过直接运用有限等比级数公式计算得出。将 $t=0$ 代入公式，即可得到关于 $|J(\theta) - J(\theta')|$ 的结果。

□

由于马尔可夫决策过程是定义在状态-动作积空间上的随机过程，因此自然会想到将相关量分解为仅依赖于状态的部分，以及在给定状态的条件下依赖于动作的部分。对于全变差距离，可以按如下方式进行分解。

引理 B.3 (总变差的边缘分解). 令 ν, ν' 为定义在 $\mathcal{S} \times \mathcal{A}$上的概率测度，其形式为

$$\nu(s,a) = d(s)\pi(a\,;\,s), \qquad \nu'(s,a) = d'(s)\pi'(a\,;\,s).$$

Then

$$\mathrm{TV}(\nu,\nu') \;\le\; \mathrm{TV}(d,d') + \sup_{s\in\mathcal{S}} \mathrm{TV}\big(\pi(\,\cdot\,;\,s), \pi'(\,\cdot\,;\,s)\big)\,.$$

证明。 根据定义， $\mathcal{S} \times \mathcal{A}$上两个测度之间的全变差距离为:

$$\mathrm{TV}(\nu,\nu') = \sup_{B\subset\mathcal{S}\times\mathcal{A}} \Big| \sum_{(s,a)\in B} \big(d(s)\pi(a\,;\,s) - d'(s)\pi'(a\,;\,s)\big)\Big|.$$

对于给定的集合 $B \subset \mathcal{S} \times \mathcal{A}$，定义 $B_s := \{a \in \mathcal{A} : (s,a) \in B\}$。则

$$\sum_{(s,a)\in B} d(s)\pi(a\,;\,s) = \sum_{s\in\mathcal{S}} d(s) \sum_{a\in B_s} \pi(a\,;\,s),$$

并且对于 ν'，也存在类似的分解。
加上并减去混合项 $\sum_{s\in\mathcal{S}} d'(s)\sum_{a\in B_s}\pi(a\,;\,s)$，得到

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} d(s) \sum_{a\in B_s} \pi(a\,;\,s) - \sum_{s\in\mathcal{S}} d'(s) \sum_{a\in B_s} \pi'(a\,;\,s) \\
&= \sum_{s\in\mathcal{S}} \big[d(s) - d'(s)\big] \sum_{a\in B_s} \pi(a\,;\,s) + \sum_{s\in\mathcal{S}} d'(s) \sum_{a\in B_s} \big[\pi(a\,;\,s) - \pi'(a\,;\,s)\big].
\end{aligned}$$

取绝对值并利用三角不等式可得

$$\mathrm{TV}(\nu,\nu') \le \sup_B \Big| \sum_{s\in\mathcal{S}} (d(s) - d'(s)) \sum_{a\in B_s} \pi(a\,;\,s)\Big| + \sup_B \Big| \sum_{s\in\mathcal{S}} d'(s) \sum_{a\in B_s} \big(\pi(a\,;\,s) - \pi'(a\,;\,s)\big)\Big|.$$

对于第一项，注意到对于所有 s， $0 \le \sum_{a\in B_s}\pi(a\,;\,s) \le 1$。因此，一个上界为 $\sup_{C\subset\mathcal{S}} \big|\sum_{s\in C}\big[d(s) - d'(s)\big]\big| =$ 全变差$(d,\ d')$。 对于第二项，我们使用上界

$$\sum_{s\in\mathcal{S}} d'(s) \sup_{C\subset\mathcal{A}} \Big| \sum_{a\in C} \big[\pi(a\,;\,s) - \pi'(a\,;\,s)\big]\Big| = \sup_{s\in\mathcal{S}} \mathrm{TV}\big(\pi(\,\cdot\,;\,s), \pi'(\,\cdot\,;\,s)\big),$$

其中，最后一个等式由 $\mathcal{A}$ 上全变差距离的定义得出。 □

引理 B.4。 边缘状态分布之间的全变差距离可以分解为

$$\mathrm{TV}\big(d_t^{\pi'}, d_t^{\pi}\big) \le \sum_{i=0}^{t-1} \mathbb{E}^{\pi}\left[\mathrm{TV}\big(\pi'(\,\cdot\,;\,S_i), \pi(\,\cdot\,;\,S_i)\big)\right].$$

证明。 对于所有 $t \in \{1, \dots, T\}$，我们有 $d_t^{\pi}(s) = \sum_{s' \in \mathcal{S}} d_{t-1}^{\pi}(s') \sum_{a' \in \mathcal{A}} \pi(a' ;\ s') p(s ;\ s', a')$，因此

$$
\begin{aligned}
\sum_{s \in \mathcal{S}} \big| d_t^{\pi'}(s) - d_t^{\pi}(s) \big| &= \sum_{s \in \mathcal{S}} \Big| \sum_{s' \in \mathcal{S}} \Big(d_{t-1}^{\pi'}(s') \sum_{a' \in \mathcal{A}} \pi'(a' ;\ s') p(s ;\ s', a') - d_{t-1}^{\pi}(s') \sum_{a' \in \mathcal{A}} \pi(a' ;\ s') p(s ;\ s', a') \Big) \Big| \\
&= \sum_{s \in \mathcal{S}} \Big| \sum_{s' \in \mathcal{S}} \big(d_{t-1}^{\pi'}(s') - d_{t-1}^{\pi}(s') \big) \sum_{a' \in \mathcal{A}} \pi'(a' ;\ s') p(s ;\ s', a') \\
&\qquad + \sum_{s' \in \mathcal{S}} d_{t-1}^{\pi}(s') \sum_{a' \in \mathcal{A}} \big(\pi'(a' ;\ s') - \pi(a' ;\ s') \big) p(s ;\ s', a') \Big| \\
&\le \sum_{s \in \mathcal{S}} \Big(\sum_{s' \in \mathcal{S}} \big| d_{t-1}^{\pi'}(s') - d_{t-1}^{\pi}(s') \big| \sum_{a' \in \mathcal{A}} \pi'(a' ;\ s') p(s ;\ s', a') \\
&\qquad + \sum_{s' \in \mathcal{S}} d_{t-1}^{\pi}(s') \sum_{a' \in \mathcal{A}} \big| \pi'(a' ;\ s') - \pi(a' ;\ s') \big| p(s ;\ s', a') \Big) \\
&= \sum_{s' \in \mathcal{S}} \big| d_{t-1}^{\pi'}(s') - d_{t-1}^{\pi}(s') \big| \sum_{a' \in \mathcal{A}} \pi'(a' ;\ s') \sum_{s \in \mathcal{S}} p(s ;\ s', a') \\
&\qquad + \sum_{s' \in \mathcal{S}} d_{t-1}^{\pi}(s') \sum_{a' \in \mathcal{A}} \big| \pi'(a' ;\ s') - \pi(a' ;\ s') \big| \sum_{s \in \mathcal{S}} p(s ;\ s', a') \\
&= \sum_{s' \in \mathcal{S}} \big| d_{t-1}^{\pi'}(s') - d_{t-1}^{\pi}(s') \big| + \sum_{s' \in \mathcal{S}} d_{t-1}^{\pi}(s') \sum_{a' \in \mathcal{A}} \big| \pi'(a' ;\ s') - \pi(a' ;\ s') \big|,
\end{aligned}
$$

即全变差$\big(d_t^{\pi'},\ d_t^{\pi}\big) \le$全变差$\big(d_{t-1}^{\pi'},\ d_{t-1}^{\pi}\big) + \mathbb{E}^{\pi}\big[$全变差$\big(\pi'(\ \cdot\ ;\ S_{t-1})\ \pi(\ \cdot\ ;\ S_{t-1})\big)\big]$。利用 $d_0^{\pi'} = d_0^{\pi}$ 并递归地应用此不等式，即得所述结论。 □

命题 B.5 (L关于 J的光滑性。 在假设 ***A.1***、***A.3*** 和 ***A.4*** 下，对所有θ、 $\theta' \in \mathbb{R}^d$,

$$
\begin{aligned}
\big| \nabla J(\theta) - \nabla J(\theta') \big| &\le R_* \sum_{t=0}^{T-1} \gamma^t \Big(L \sum_{k=0}^{T-t-1} \gamma^k + \Pi_*^2 \gamma \sum_{k=0}^{T-t-2} \gamma^k \Big(\sum_{j=0}^{T-t-2-k} \gamma^j \Big) + (t+1) \Pi_*^2 \sum_{k=0}^{T-t-1} \gamma^k \Big) \big| \theta - \theta' \big| \\
&= \underbrace{R_* \left(\frac{L_s (1-\gamma^T)}{(1-\gamma)^2} - \frac{L_s T \gamma^T}{1-\gamma} + \Pi_*^2 \Big(\frac{1+\gamma-2\gamma^T}{(1-\gamma)^3} - \frac{(2T-1)\gamma^T}{(1-\gamma)^2} - \frac{T^2 \gamma^T}{1-\gamma} \Big) \right)}_{=:L} |\theta - \theta'|.
\end{aligned}
$$

证明。 我们将策略梯度写成得分函数形式

$$
\nabla J(\theta) = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{s \sim d_t^{\pi_\theta},\, a \sim \pi_\theta(\,\cdot\,;\, s)} \big[\nabla_\theta \log \pi_\theta(a ;\ s)\, Q_t^{\pi_\theta}(s, a) \big].
$$

对于 $t = 0$、.....、$T-1$，我们写作,

$$
\phi_t^{\theta}(s, a) := \gamma^t \nabla_\theta \log \pi_\theta(a ;\ s)\, Q_t^{\pi_\theta}(s, a),
$$

使得 $\nabla J(\theta) = \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d_t^{\pi_\theta},\ a \sim \pi_\theta(\,\cdot\,;\ s)} [\phi_t^{\theta}(s,\ a)]$。对于固定的 $t \in \{0, \dots, T-1\}$，我们将其分解

$$
\begin{aligned}
&\mathbb{E}_{s \sim d_t^{\pi_\theta},\, a \sim \pi_\theta(\,\cdot\,;\, s)} [\phi_t^{\theta}(s, a)] - \mathbb{E}_{s \sim d_t^{\pi_{\theta'}},\, a \sim \pi'_\theta(\,\cdot\,;\, s)} [\phi_t^{\theta'}(s, a)] \\
&= \underbrace{\mathbb{E}_{s \sim d_t^{\pi_\theta},\, a \sim \pi_\theta(\,\cdot\,;\, s)} [\phi_t^{\theta}(s, a) - \phi_t^{\theta'}(s, a)]}_{(A)} + \underbrace{\mathbb{E}_{s \sim d_t^{\pi_\theta},\, a \sim \pi_\theta(\,\cdot\,;\, s)} [\phi_t^{\theta'}(s, a)] - \mathbb{E}_{s \sim d_t^{\pi_{\theta'}},\, a \sim \pi_{\theta'}(\,\cdot\,;\, s)} [\phi_t^{\theta'}(s, a)]}_{(B)}.
\end{aligned}
$$

我们首先求出 (A) 的一个界。为此，改写

$$\begin{aligned}\phi_t^{\theta}(s,a)-\phi_t^{\theta'}(s,a)=&\gamma^t\big(\nabla_\theta\log\pi_\theta(a\,;\,s)-\nabla_\theta\log\pi_{\theta'}(a\,;\,s)\big)Q_t^{\pi_\theta}(s,a)\\&+\gamma^t\nabla_\theta\log\pi_{\theta'}(a\,;\,s)\big(Q_t^{\pi_\theta}(s,a)-Q_t^{\pi_{\theta'}}(s,a)\big).\end{aligned}$$

注意到

$$\sum_{k=0}^{T-t-2}\gamma^k\Big(\sum_{j=0}^{T-t-2-k}\gamma^j\Big)=\frac{1-(T-t)\,\gamma^{T-t-1}+(T-t-1)\,\gamma^{T-t}}{(1-\gamma)^2},$$

因此，根据引理 B.2，有

$$\|Q_t^{\pi_\theta}-Q_t^{\pi_{\theta'}}\|_\infty\le\gamma\|V_{t+1}^{\pi_\theta}-V_{t+1}^{\pi_{\theta'}}\|_\infty\le\big|\theta-\theta'\big|\Pi_*R_*\gamma\sum_{k=0}^{T-t-2}\gamma^k\Big(\sum_{j=0}^{T-t-2-k}\gamma^j\Big).$$

结合得分函数的利普希茨性和有界性假设，以及由于奖励有界性可知的事实 $\|Q_t^{\pi_\theta}\|_\infty\le R_*\sum_{k=0}^{T-t-1}\gamma^k$，我们得到

$$\|(A)\|_\infty\le\gamma^tR_*\Big(L_{\mathrm{s}}\sum_{k=0}^{T-t-1}\gamma^k+\Pi_*^2\gamma\sum_{k=0}^{T-t-2}\gamma^k\Big(\sum_{j=0}^{T-t-2-k}\gamma^j\Big)\Big)\big|\theta-\theta'\big|.$$

现在转向 (B)，即分布偏移。首先注意到

$$\|(B)\|_\infty\le2\|\phi_t^{\theta'}\|_\infty\,\mathrm{TV}\big(\nu(\theta)\,,\,\nu(\theta')\big)\,,\tag{9}$$

其中，$\nu(\theta)$ 是 $\mathcal{S}\times\mathcal{A}$ 上的测度，满足 $\nu(s,a)=d_t^{\pi_\theta}(s)\pi_\theta(a\mid s)$。利用总变差的对偶刻画可以得到这一点，即 $\mathrm{TV}(\mu,\nu)=\frac{1}{2}\sup_{\|f\|_\infty\le1}\big|\int f\,d\mu-\int f\,d\nu\big|$。令 $\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ 为有界可测函数，并定义 $f:=\frac{\phi}{\|\phi\|_\infty}$，使得 $\|f\|_\infty\le1$。那么

$$|\mathbb{E}_\nu[\phi]-\mathbb{E}_{\nu'}[\phi]|=\Big|\int\phi\,d\nu-\int\phi\,d\nu'\Big|=\|\phi\|_\infty\Big|\int f\,d\nu-\int f\,d\nu'\Big|\le2\|\phi\|_\infty\,\mathrm{TV}\big(\nu,\nu'\big)\,.$$

我们现在估计（9）式右端。首先，利用假设 A.1 和 A.3 可得

$$\|\phi_t^{\theta'}\|_\infty\le\gamma^t\Pi_*R_*\sum_{k=0}^{T-t-1}\gamma^k.$$

接下来，利用引理 B.3 和引理 A.5 可得

$$\mathrm{TV}\big(\nu(\theta)\,,\,\nu(\theta')\big)\le\mathrm{TV}\big(d_t^{\pi_\theta}\,,\,d_t^{\pi'_\theta}\big)+\frac{1}{2}\Pi_*|\theta-\theta'|.$$

此外，引理 B.4 与引理 A.5 一起表明，对所有 $t\in\{0,\dots,T\}$

$$\mathrm{TV}\big(d_t^{\pi'},d_t^{\pi}\big)\le\sum_{i=0}^{t-1}\mathbb{E}^{\pi}\left[\mathrm{TV}\big(\pi'(\,\cdot\,;\,S_i),\pi(\,\cdot\,;\,S_i)\big)\right]\le\frac{t}{2}\Pi_*|\theta-\theta'|.$$

结合上述估计可得

$$\|(B)\|_\infty\le\gamma^t(t+1)\Pi_*^2R_*\sum_{k=0}^{T-t-1}\gamma^k|\theta-\theta'|.$$

对 $t=0,\ \ldots\ldots,\ T-1$ 中 (A) 和 (B) 的界求和，即可得到结果；命题陈述中的最后一个等式可通过仔细计算并应用几何级数公式推导得出。

□

在本文中，我们关注折扣有限时域马尔可夫决策过程。然而，自然会问该证明在极限无限时域折扣情形（$T\to\infty,\ \gamma<1$）以及有限时域无折扣情形（$T<\infty,\ \gamma=1$）中会得到什么。

注记 B.6. 尽管我们的证明技巧不同于 [40] 在无限时域折扣情形中所使用的技巧，但它在极限 $T\to\infty$中恢复了完全相同的光滑常数。在该情形下，光滑性估计简化为

$$\big|\nabla J(\theta)-\nabla J(\theta')\big|\le R_*\Big(\frac{L_{\mathrm{s}}}{(1-\gamma)^2}+\frac{\Pi_*^2(1+\gamma)}{(1-\gamma)^3}\Big)\big|\theta-\theta'\big|$$

这与 [40]的引理 3.2 中所述的界一致。

注记 B.7. 在有限时域无折扣情形（$\gamma=1$ 和 $T<\infty$）中，同样的论证仍然成立（几何和得以简化），从而得到

$$\big|\nabla J(\theta)-\nabla J(\theta')\big|\le R_*\Big(\frac{L_{\mathrm{s}}T(T+1)}{2}+\frac{\Pi_*^2T(2T^2+3T+1)}{6}\Big)\big|\theta-\theta'\big|,$$

这略小于上界

$$R_*\big(L_{\mathrm{s}}T^2+\Pi_*^2T^3\big)\big|\theta-\theta'\big|$$

这是在 [23] 中于稍强的假设下推导得到的。

这两条注释体现了无限时域折扣马尔可夫决策过程与有效时域为 $T=(1-\gamma)^{-1}$ 的有限时域无折扣马尔可夫决策过程之间众所周知的对应关系。特别地，请回忆：无限时域折扣马尔可夫决策过程的价值函数，与时间时域为期望为 $(1-\gamma)^{-1}$ 的独立几何随机变量的无折扣马尔可夫决策过程的价值函数相同。

C 策略梯度偏差理论

现在我们来讨论本文的主要贡献之一：对近端策略优化中使用的代理梯度偏差进行界定。在接下来的两节中，我们将证明定理 4.2。

C.1 未裁剪代理梯度偏差

本节中，我们估计真实策略梯度与代理梯度之间的差异

$$g_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})=\sum_{t=0}^{T-1}\gamma^t\mathbb{E}_\mu^{\pi_{\theta_{\mathrm{old}}}}\Big[\frac{\nabla_\theta\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\mathbb{A}_t^{\pi_{\theta_{\mathrm{old}}}}(S_t,A_t)\Big].$$

在下　　接下来，我们将偏差界从近端策略优化转移到裁剪梯度 g^{clip} 近端策略优化 。

这些估计基于折扣有限时域马尔可夫决策过程的性能差异引理的一个变体（对于无限时域折扣马尔可夫决策过程，参见 [10, 引理 6.1] 和[28, 方程 (1)] ）。为方便读者，我们将补充证明。

命题 C.1（性能差异恒等式）。对于任意两个策略 π，$\tilde{\pi}$，

$$V^{\tilde{\pi}}(\mu) - V^{\pi}(\mu) = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\mu}^{\tilde{\pi}}\big[\mathbb{A}_t^{\pi}(S_t, A_t)\big].$$

特别地，对于任意两个参数化策略 π_θ,$\pi_{\theta_{\text{old}}}$，

$$\nabla_\theta J(\theta) = \nabla_\theta \left(V^{\pi_\theta}(\mu) - V^{\pi_{\theta_{\text{old}}}}(\mu)\right) = \nabla_\theta \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\mu}^{\pi_\theta}\big[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t)\big]. \tag{10}$$

证明。首先回顾一下，$Q_t^{\pi}(s,\ a) = \mathbb{E}_{S_t=s,\ A_t=a}^{\pi}\left[R_t + \gamma V_{t+1}^{\pi}(S_{t+1})\right]$ 和 $\mathbb{E}_{S\sim d_t^{\pi},\ A\sim\pi(\cdot\,;\ S)}\left[A_t^{\pi}(S,\ A)\right] = 0$。因此，

$$\begin{aligned}
V^{\tilde{\pi}}(\mu) - V^{\pi}(\mu) &= \sum_{t=0}^{T-1} \gamma^t \left(\mathbb{E}^{\tilde{\pi}}[R_t] - \mathbb{E}^{\pi}[R_t]\right) \\
&= \sum_{t=0}^{T-1} \gamma^t \Big(\mathbb{E}_{S\sim d_t^{\tilde{\pi}}, A\sim\tilde{\pi}(\cdot\,;S)}\big[\underbrace{\mathbb{E}_{S_t=S,A_t=A}^{\pi}[R_t]}_{=Q_t^{\pi}(S,A)-\gamma\mathbb{E}_{S'\sim p(\cdot\,;S,A)}[V_{t+1}^{\pi}(S')]}\big] - \mathbb{E}_{S\sim d_t^{\pi}, A\sim\pi(\cdot\,;S)}\big[\mathbb{E}_{S_t=S,A_t=A}^{\pi}[R_t]\big]\Big) \\
&= \sum_{t=0}^{T-1} \gamma^t \Big(\mathbb{E}_{S\sim d_t^{\tilde{\pi}}, A\sim\tilde{\pi}(\cdot\,;S)}\big[A_t^{\pi}(S,A) - \gamma\mathbb{E}_{S'\sim p(\cdot\,;S,A)}[V_{t+1}^{\pi}(S')] + V_t^{\pi}(S)\big] \\
&\qquad -\mathbb{E}_{S\sim d_t^{\pi}, A\sim\pi(\cdot\,;S)}\big[A_t^{\pi}(S,A) - \gamma\mathbb{E}_{S'\sim p(\cdot\,;S,A)}[V_{t+1}^{\pi}(S')] + V_t^{\pi}(S)\big]\Big) \\
&= \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{S\sim d_t^{\tilde{\pi}}, A\sim\tilde{\pi}(\cdot\,;S)}[A_t^{\pi}(S,A)] \\
&\quad + \sum_{t=0}^{T-1} \gamma^t \left(\gamma\mathbb{E}_{S\sim d_t^{\pi}, A\sim\pi(\cdot\,;S)}\big[\mathbb{E}_{S'\sim p(\cdot\,;S,A)}[V_{t+1}^{\pi}(S')]\big] - \mathbb{E}_{S\sim d_t^{\pi}, A\sim\pi(\cdot\,;S)}[V_t^{\pi}(S)]\right) \\
&\quad - \sum_{t=0}^{T-1} \gamma^t \left(\gamma\mathbb{E}_{S\sim d_t^{\tilde{\pi}}, A\sim\tilde{\pi}(\cdot\,;S)}\big[\mathbb{E}_{S'\sim p(\cdot\,;S,A)}[V_{t+1}^{\pi}(S')]\big] - \mathbb{E}_{S\sim d_t^{\tilde{\pi}}, A\sim\tilde{\pi}(\cdot\,;S)}[V_t^{\pi}(S)]\right).
\end{aligned}$$

对于第二个求和项，我们利用 $V_T^{\pi} \equiv 0$ 计算得：

$$\begin{aligned}
&\sum_{t=0}^{T-1} \gamma^t \left(\gamma\mathbb{E}_{S\sim d_t^{\pi}, A\sim\pi(\cdot\,;S)}\big[\mathbb{E}_{S'\sim p(\cdot\,;S,A)}[V_{t+1}^{\pi}(S')]\big] - \mathbb{E}_{S\sim d_t^{\pi}, A\sim\pi(\cdot\,;S)}[V_t^{\pi}(S)]\right) \\
=&\sum_{t=0}^{T-1} \gamma^t \left(\gamma\mathbb{E}^{\pi}[V_{t+1}^{\pi}(S_{t+1})] - \mathbb{E}^{\pi}[V_t^{\pi}(S_t)]\right) \\
=&\gamma^T\mathbb{E}^{\pi}[V_T^{\pi}(S_T)] - \mathbb{E}^{\pi}[V_0^{\pi}(S_0)] + \sum_{t=0}^{T-2} \gamma^{t+1}\mathbb{E}^{\pi}[V_{t+1}^{\pi}(S_{t+1})] - \sum_{t=1}^{T-1} \gamma^t\mathbb{E}^{\pi}[V_t^{\pi}(S_t)] \\
=&-\mathbb{E}^{\pi}[V_0^{\pi}(S_0)],
\end{aligned}$$

第三个求和同理。由于 $d_0^{\pi} = \mu = d_0^{\tilde{\pi}}$，我们有 $\mathbb{E}^{\pi}[V_0^{\pi}(S_0)] = \mathbb{E}^{\tilde{\pi}}[V_0^{\pi}(S_0)]$,，这意味着这些和相互抵消，从而完成证明。 □

利用性能差异恒等式，我们可以推导出以下关于真实策略梯度与（未裁剪的）替代目标之差的引理。接下来，我们使用 θ 和 θ_{old} 表示任意参数，因为这将在后续分析中使用。

引理 C.2。

$$\nabla_\theta J(\theta) - g_{PPO}(\theta, \theta_{old}) = \sum_{t=0}^{T-1} \gamma^t \nabla_\theta \left(\mathbb{E}^{\pi_\theta}[g_t^\theta(S_t)] - \mathbb{E}^{\pi_{\theta_{\text{old}}}}[g_t^\theta(S_t)] \right).$$

其中 $g_t^\theta(s) := \mathbb{E}_{A\sim\pi_\theta(\cdot;\ s)}[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,\ A)]$

证明。 该分解是性能差异引理和 Fubini定理的推论；后者允许我们对状态分布与动作分布进行分解。首先，根据性能差异引理和 Fubini定理

$$\begin{aligned}\nabla_\theta J(\theta) = \nabla_\theta \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_\mu^{\pi_\theta}\left[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t)\right] &= \nabla_\theta \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{S\sim d_t^{\pi_{\text{old}}}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot\,;\,S)}\left[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t)\right]\right] \\ &= \sum_{t=0}^{T-1} \gamma^t \nabla_\theta \mathbb{E}^{\pi_\theta}[g_t^\theta(S_t)].\end{aligned}$$

接下来,

$$\begin{aligned} g_{\text{PPO}}(\theta, \theta') &= \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_\mu^{\pi_{\theta_{\text{old}}}}\Big[\frac{\nabla_\theta \pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\text{old}}}(A_t\,;\,S_t)} \mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t)\Big] \\ &= \sum_{t=0}^{T-1} \gamma^t \nabla_\theta \mathbb{E}^{\pi_{\theta_{\text{old}}}}\Big[\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\text{old}}}(A_t\,;\,S_t)} \mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S_t, A_t)\Big] \\ &= \sum_{t=0}^{T-1} \gamma^t \nabla_\theta \mathbb{E}_{S\sim d_t^{\pi_{\theta_{\text{old}}}}}\left[\mathbb{E}_{A\sim\pi_\theta(\,\cdot\,;S)}[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(S, A)]\right], \end{aligned}$$

这里需要注意的是，去掉重要性比率乘积后，经 Fubini 分解，重要性比率只会对状态 S_t 下的动作重新加权。作差即可得到所述结论。 □

定理 C.3 (未裁剪替代目标梯度偏差)。 设 $\pi_{\theta_{\text{old}}}$ 是一个具有严格正权重的行为策略，并定义平均总变差距离

$$\text{M}_{\text{ean}}\text{TV}(\pi_{\theta_{\text{old}}}, \pi_\theta) := \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_\mu^{\pi_{\theta_{\text{old}}}}\left[\text{TV}\left(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t), \pi_\theta(\,\cdot\,;\,S_t)\right)\right]$$

以及最大总变差距离

$$\text{M}_{\text{ax}}\text{TV}(\pi_{\theta_{\text{old}}}, \pi_\theta) := \max_{t=0,\ldots,T-1} \mathbb{E}_\mu^{\pi_{\text{old}}}\left[\text{TV}\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t), \pi_\theta(\,\cdot\,;\,S_t)\big)\right].$$

在假设 *A.1*、*A.3* 下，有

$$\left\|\nabla_\theta J(\theta) - g_{PPO}(\theta, \theta_{old})\right\|_\infty \le \Pi_*\, R_*\, \min\left\{c_1\, \text{M}_{\text{ean}}\text{TV}(\pi_{\theta_{\text{old}}}, \pi_\theta), c_2\, \text{M}_{\text{ax}}\text{TV}(\pi_{\theta_{\text{old}}}, \pi_\theta)\right\},$$

with

$$c_1 \coloneqq \begin{cases} 8\frac{1-\gamma^T}{1-\gamma} T\big(\frac{\gamma}{(1-\gamma)^2} + \frac{\gamma}{1-\gamma}\big) \le \frac{16T\gamma}{(1-\gamma)^3} & : 0 < \gamma < 1 \\ 4T^4 & : \gamma = 1 \end{cases},$$

$$c_2 \coloneqq \begin{cases} 8\frac{1-\gamma^T}{1-\gamma} \frac{2\gamma - T(T+1)\gamma^{T-1} + 2(T^2-1)\gamma^T - T(T-1)\gamma^{T+1}}{(1-\gamma)^3} \le \frac{16\gamma}{(1-\gamma)^4} & : 0 < \gamma < 1 \\ 8T\frac{(T-1)T(T+1)}{3} \le \frac{8}{3}T^4 & : \gamma = 1 \end{cases}.$$

偏差界的表述看起来有些复杂，因为它同时涵盖了有限时域折扣、有限时域非折扣和无限时域折扣马尔可夫决策过程这几种情形。在我们的近端策略优化分析中，我们只研究有限时域情形，并使用平均总变差策略距离来界定梯度偏差。对于无限时域折扣马尔可夫决策过程，读者应使用最大总变差散度，其常数为有效时域 $\frac{1}{1-\gamma}$ 的四次方。 —

引理 C.4. 对于任意满足 $\big\|\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,\cdot)\big\|_\infty \le \mathbb{A}_{\max}$ 的 $s \in \mathcal{S}$，

$$\big|\mathbb{E}_{A\sim\pi_\theta(\cdot\,;\,s)}\big[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,A)\big]\big| \le 2\,\mathrm{TV}\big(\pi_\theta(\cdot\,;\,s), \pi_{\theta_{\text{old}}}(\cdot\,;\,s)\big)\mathbb{A}_{\max}.$$

证明。 由于 $\mathbb{E}_{A\sim\pi_{\theta_{\text{old}}}(\cdot\,;s)}[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,A)] = 0$，命题 4.5] 中的总变差距离不等式 [13, 推出

$$\begin{aligned}\big|\mathbb{E}_{A\sim\pi_\theta(\cdot\,;\,s)}\big[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,A)\big]\big| &= \big|\mathbb{E}_{A\sim\pi_\theta(\cdot\,;\,s)}\big[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,A)\big] - \mathbb{E}_{A\sim\pi_{\theta_{\text{old}}}(\cdot\,;\,s)}\big[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,A)\big]\big| \\ &\le 2\,\mathrm{TV}(\pi_\theta(\cdot\,;\,s), \pi_{\theta_{\text{old}}}(\cdot\,;\,s))\mathbb{A}_{\max}.\end{aligned}$$

□

定理 *C.3* 的证明。 我们定义 $g_t^\theta(s) \coloneqq \mathbb{E}_{A\sim\pi_\theta(\cdot\,;s)}[\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,A)]$，使得引理 C.2 给出

$$\nabla_\theta J(\theta) - g_{\text{PPO}}(\theta,\theta_{\text{old}}) = \sum_{t=0}^{T-1} \gamma^t \nabla_\theta \left(\mathbb{E}_{S\sim d_t^{\pi_\theta}}[g_t^\theta(S)] - \mathbb{E}_{S\sim d_t^{\pi_{\theta_{\text{old}}}}}[g_t^\theta(S)]\right).$$

对于 θ 的单个坐标 θ_j，我们首先计算

$$\partial_{\theta_j} d_t^{\pi_\theta}(s) = \sum_{\tau:s_t=s} \partial_{\theta_j}\mathbb{P}_t^{\pi_\theta}(\tau) = \sum_{\tau:s_t=s} \mathbb{P}_t^{\pi_\theta}(\tau)\,\partial_{\theta_j}\log\mathbb{P}_t^{\pi_\theta}(\tau) = \sum_{\tau:s_t=s} \mathbb{P}_t^{\pi_\theta}(\tau)\sum_{i=0}^{t-1}\partial_{\theta_j}\log\pi_\theta(a_i;s_i),$$

其中 $\tau : s_t = s$ 表示所有长度为 t 且以 $s_t = s$ 结尾的轨迹；转移概率的导数不会出现在最终表达式中，因为它们独立于 θ。我们还有如下估计

$$\left|\partial_{\theta_j} g_t^\theta(s)\right| = \left|\sum_{a\in\mathcal{A}} \partial_{\theta_j}\pi_\theta(a\,;\,s)\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,a)\right| = \left|\sum_{a\in\mathcal{A}} \pi_\theta(a\,;\,s)\partial_{\theta_j}\log\pi_\theta(a;s)\mathbb{A}_t^{\pi_{\theta_{\text{old}}}}(s,a)\big]\right| \le 2\Pi_* R_* \sum_{t=0}^{T-1}\gamma^t,$$

根据有界奖励假设，优势函数上界为 $2R_*\sum_{t=0}^{T-1}\gamma^t$。

结合上述结果与定理 C.4，并再次应用总变差距离不等式，我们得到

$$
\begin{aligned}
&\Big|\partial_{\theta_j}\Big(\mathbb{E}_{S\sim d_t^{\pi_\theta}}[g_t^\theta(S)] - \mathbb{E}_{S\sim d_t^{\pi_{\theta_{\text{old}}}}}[g_t^\theta(S)]\Big)\Big| \\
&= \Big|\sum_{s\in\mathcal{S}} \partial_{\theta_j}\Big(d_t^{\pi_\theta}(s) g_t^\theta(s)\Big) - \sum_{s\in\mathcal{S}} d_t^{\pi_{\theta_{\text{old}}}}(s)\,\partial_{\theta_j} g_t^\theta(s)\Big| \\
&= \Big|\sum_{s\in\mathcal{S}}\sum_{\tau:s_t=s} \mathbb{P}_t^{\pi_\theta}(\tau)\sum_{i=0}^{t-1}\partial_{\theta_j}\log\pi_\theta(a_i\,;\,s_i)\,g_t^\theta(s) + \sum_{s\in\mathcal{S}}(d_t^{\pi_\theta}(s) - d_t^{\pi_{\theta_{\text{old}}}}(s))\,\partial_{\theta_j} g_t^\theta(s)\Big| \\
&\le t\Pi_*\mathbb{E}^{\pi_\theta}\big[g_t^\theta(S_t)\big] + 4\Pi_* R_*\Big(\sum_{i=0}^{T-1}\gamma^i\Big)\mathrm{TV}\big(d_t^{\pi_{\theta_{\text{old}}}}, d_t^{\pi_\theta}\big) \\
&\le \underbrace{4\Pi_* R_*\Big(\sum_{i=0}^{T-1}\gamma^i\Big)}_{=:C}\Big(t\,\mathbb{E}^{\pi_\theta}\big[\mathrm{TV}\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t), \pi_\theta(\,\cdot\,;\,S_t)\big)\big] + \mathrm{TV}\big(d_t^{\pi_{\theta_{\text{old}}}}, d_t^{\pi_\theta}\big)\Big).
\end{aligned}
$$

因为我们希望 避免对 π_θ 取任何期望，我们再次使用总变差距离不等式来得到 t

$$
\mathbb{E}^{\pi_\theta}\big[\mathrm{TV}\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t), \pi_\theta(\,\cdot\,;\,S_t)\big)\big] \le \mathbb{E}^{\pi_{\theta_{\text{old}}}}\big[\mathrm{TV}\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t), \pi_\theta(\,\cdot\,;\,S_t)\big)\big] + 2\,\mathrm{TV}\big(d_t^{\pi_{\theta_{\text{old}}}}, d_t^{\pi_\theta}\big).
$$

这给出

$$
\begin{aligned}
\Big|\partial_{\theta_j}\Big(\mathbb{E}_{S\sim d_t^{\pi_\theta}}[g_t^\theta(S)] - \mathbb{E}_{S\sim d_t^{\pi_{\theta_{\text{old}}}}}[g_t^\theta(S)]\Big)\Big| \le C\Big(&t\,\mathbb{E}^{\pi_{\theta_{\text{old}}}}\big[\mathrm{TV}\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t), \pi_\theta(\,\cdot\,;\,S_t)\big)\big] \\
&+ \big(2t+1\big)\mathrm{TV}\big(d_t^{\pi_{\theta_{\text{old}}}}, d_t^{\pi_\theta}\big)\Big).
\end{aligned}
$$

应用定理 B.4 可得

$$
\begin{aligned}
&\big|\nabla_\theta J(\theta)_j - g_{\text{PPO}}(\theta,\theta_{\text{old}})_j\big| \\
&\le C\sum_{t=0}^{T-1}\gamma^t\Big(t\,\mathbb{E}^{\pi_{\theta_{\text{old}}}}\big[\mathrm{TV}\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t), \pi_\theta(\,\cdot\,;\,S_t)\big)\big] + (2t+1)\sum_{i=0}^{t-1}\mathbb{E}^{\pi_{\theta_{\text{old}}}}\big[\mathrm{TV}\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_i), \pi_\theta(\,\cdot\,;\,S_i)\big)\big]\Big).
\end{aligned}
\tag{11}
$$

现在，记作 $d_t := \mathbb{E}^{\pi_{\theta_{\text{old}}}}\big[$总变差$\big(\pi_{\theta_{\text{old}}}(\,\cdot\,;\,S_t)\;\pi_\theta(\,\cdot\,;\,S_t)\big]$，我们可以重新整理

$$
\begin{aligned}
\big|\nabla_\theta J(\theta)_j - g_{\text{PPO}}(\theta,\theta_{\text{old}})_j\big| &\le C\sum_{i=0}^{T-1} d_i\Big(i\gamma^i + \sum_{t=i+1}^{T-1}\gamma^t(2t+1)\Big) \\
&=: C\sum_{i=0}^{T-1} d_i w_i(\gamma) = 4\Pi_* R_*\sum_{i=0}^{T-1}\gamma^i\sum_{i=0}^{T-1} d_i w_i(\gamma).
\end{aligned}
$$

我们先作如下估计

$$
\sum_{i=0}^{T-1} d_i w_i(\gamma) \le \max_{i=0,\dots,T-1} w_i(\gamma)\sum_{i=0}^{T-1} d_i = \max_{i=0,\dots,T-1} w_i(\gamma)\,T\,\mathrm{M_{ean}TV}(\pi_{\theta_{\text{old}}}, \pi_\theta)
$$

现在需要为 $\max_{i=0,\dots,T-1} w_i(\gamma)$ 求一个上界。对于 $\gamma<1$，$x\mapsto x\gamma^x$ 在 $x=(-\log\gamma)^{-1}$ 处达到最大值 ~~$\exp(-1)\,(-\log\gamma)$~~ $\le \exp(-1)^{1-\gamma}$，因此我们有 $\max_{i=0,\dots,T-1} i\gamma^i \le \exp(-1)^{1-\gamma}$。此外，~~—~~

通过谨慎运用几何级数可得

$$\max_{i=0,\dots,T-1}\sum_{t=i+1}^{T-1}\gamma^t(2t+1)=\sum_{t=1}^{T-1}\gamma^t(2t+1)=2\frac{\gamma-T\gamma^T+(T-1)\gamma^{T+1}}{(1-\gamma)^2}+\frac{\gamma-\gamma^T}{1-\gamma}\le\frac{2\gamma}{(1-\gamma)^2}+\frac{\gamma}{1-\gamma}$$

结合这两个估计，我们得到 $\max_{i=0,\dots,T-1} w_i(\gamma) \le \exp\left(\frac{2\gamma}{(1-\gamma)^2}+\frac{\gamma}{1-\gamma}\right)$，这意味着在情形 $\gamma<1$ 下，断言中的常数为 c_1。对于 $\gamma=1$，由以下结果可得 c_1：

$$\max_{i=0,\dots,T-1} w_i(\gamma)=\max_{i=0,\dots,T-1} i+\sum_{t=i+1}^{T-1}(2t+1)=\max_{i=0,\dots,T-1} T^2-1-i^2-i\le T^2.$$

或者 对于 $\gamma<1$，仔细应用几何级数公式可得

$$\sum_{t=0}^{T-1}\gamma^t(t^2+t)=\frac{2\gamma-T(T+1)\gamma^{T-1}+2(T^2-1)\gamma^T-T(T-1)\gamma^{T+1}}{(1-\gamma)^3}\le\frac{2\gamma}{(1-\gamma)^3}$$

并且，对于 $\gamma=1$，有 $\sum_{t=0}^{T-1}(t^2+t)=\frac{(T-1)T(T+1)}{3}$，我们可以将此结果与公式（11）结合起来进行估计

$$\begin{aligned}\left|\nabla_\theta J(\theta)_j-g_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})_j\right|&\le C\sum_{t=0}^{T-1}\gamma^t\big(t\mathrm{M_{ax}TV}(\pi_{\theta_{\mathrm{old}}},\pi_\theta)+(2t+1)t\mathrm{M_{ax}TV}(\pi_{\theta_{\mathrm{old}}},\pi_\theta)\big)\\&=8\Pi_*R_*\mathrm{M_{ax}TV}(\pi_{\theta_{\mathrm{old}},\pi_\theta})\sum_{t=0}^{T-1}\gamma^t\sum_{t=0}^{T-1}\gamma^t(t^2+t)\\&\le\Pi_*R_*c_2\mathrm{M_{ax}TV}(\pi_{\theta_{\mathrm{old}}},\pi_\theta)\end{aligned}$$

其中，断言中的 c_2。 □

C.2 裁剪代理梯度偏差

本节将为未裁剪代理梯度与裁剪代理梯度之间的差异建立一个上界。该裁剪代理梯度仿照原始 PPO 论文 [30] 中引入的裁剪损失的结构。结合此上界与前一节推导的上界，我们可以界定裁剪代理梯度与真实策略梯度之间的距离。我们引入一种裁剪代理梯度的近似形式，将重要性比率偏离 1 过多的样本的贡献截断。与原始 PPO目标 [30], 相比，此处使用的截断在比率两侧是对称的，因此略微更为保守；详见下文的注释 C.5。

考虑以下代理梯度:

$$g^{\mathrm{clip}}_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}}):=\sum_{t=0}^{T-1}\gamma^t\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\Big[\frac{\nabla_\theta\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}-1\right|\le\epsilon}\mathbb{A}_t^{\pi_{\theta_{\mathrm{old}}}}(S_t,A_t)\Big]\tag{12}$$

注释 C.5。 请注意，(12) 是一个双侧截断梯度代理。相比之下，原始 PPO 裁剪目标 [30] 会根据优势函数的符号进行非对称裁剪，而 (12) 无论 $\mathbb{A}_t^{\pi_{\theta_{\mathrm{old}}}}$ 的符号如何，都会截断两侧；这简化了分析，但代价是更加保守。

本节的主要结果如下:

定理 C.6. 设$\pi_{\theta_{\text{old}}}$ 为具有严格正权重的行为策略。在假设 *A.1*、*A.3* 下，成立

$$\left\|g_{PPO}(\theta,\theta_{old})-g_{PPO}^{clip}(\theta,\theta_{old})\right\|_{\infty}\leq C\cdot \mathrm{M}_{ean}\mathrm{TV}(\pi_{\theta_{\text{old}}},\pi_{\theta})$$

对于 $C:=4\Pi_{*}R_{*}\Big(1+\frac{1}{\epsilon}\Big)\frac{T}{1-\gamma}$，其中 ϵ 为裁剪参数。

Using三角不等式，并结合定理 C.3 和定理 C.6，可得

$$\left\|\nabla_{\theta}J(\theta)-g_{\mathrm{PPO}}^{\mathrm{clip}}(\theta,\theta_{\mathrm{old}})\right\|_{\infty}\leq \text{const.}\cdot \mathrm{M}_{\mathrm{ean}}\mathrm{TV}(\pi_{\theta_{\text{old}}},\pi_{\theta}).$$

借助引理 A.5，用 $\frac{\Pi_{*}}{2}|\theta-\theta_{\mathrm{old}}|$ 估计平均总变差，最终给出正文中的主要偏差界:

定理 C.7（正文中的定理 4.2）**.** 设$\pi_{\theta_{\text{old}}}$ 为具有严格正权重的行为策略。在假设 *A.1*、*A.3* 下，成立

$$\left\|\nabla_{\theta}J(\theta)-g_{PPO}^{clip}(\theta,\theta_{old})\right\|_{\infty}\leq R\,|\theta-\theta_{old}|,$$

其中$R=\Pi_{*}^{2}R_{*}\Big(\frac{8T\gamma}{(1-\gamma)^{3}}+\frac{2T}{1-\gamma}\Big(1+\frac{1}{\epsilon}\Big)\Big)$ 是定理 *C.3* 和定理 *C.6* 中的常数之和乘以 $\frac{\Pi_{*}}{2}$.

为了证明定理 C.6，我们需要以下两个引理。

引理 C.8. 设 P 和 Q是定义在 $\mathcal{A}$上的两个离散概率分布。那么，

$$\mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}-1\right|\right]=2\,\mathrm{TV}(P,Q).$$

证明. 根据全变差距离的定义

$$\mathrm{TV}(P,Q)=\frac{1}{2}\sum_{a\in\mathcal{A}}|P(a)-Q(a)|.$$

因此，我们得到

$$\mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}-1\right|\right]=\sum_{a\in\mathcal{A}}P(a)\left|\frac{Q(a)}{P(a)}-1\right|=\sum_{a\in\mathcal{A}}|Q(a)-P(a)|=2\,\mathrm{TV}(P,Q).\quad\square$$

引理 C.9. 设 P 和 Q 是 $\mathcal{A}$ 上的两个离散概率分布。那么，

$$\mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}\right|\mathbb{1}_{\left|\frac{Q(A)}{P(A)}-1\right|>\epsilon}\right]\leq\left(2+\frac{2}{\epsilon}\right)\mathrm{TV}(P,Q).$$

证明. 只需在期望中应用三角不等式 $|\frac{Q}{(A)}{P(A)}| = |\frac{Q}{(A)}{P(A)} - 1 + 1| \le |\frac{Q}{(A)}{P(A)} - 1| + 1$，即可得到

$$\begin{aligned}
\mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}\right|\mathbb{1}_{\left|\frac{Q(A)}{P(A)}-1\right|>\epsilon}\right] &\le \mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}-1\right|\mathbb{1}_{\left|\frac{Q(A)}{P(A)}-1\right|>\epsilon}\right] + \mathbb{E}_{A\sim P}\left[\mathbb{1}_{\left|\frac{Q(A)}{P(A)}-1\right|>\epsilon}\right] \\
&\le \mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}-1\right|\right] + \mathbb{E}_{A\sim P}\left[\mathbb{1}_{\left|\frac{Q(A)}{P(A)}-1\right|>\epsilon}\right] \\
&= \mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}-1\right|\right] + \mathbb{P}_{A\sim P}\Big(\Big|\frac{Q(A)}{P(A)}-1\Big| > \epsilon\Big).
\end{aligned}$$

应用马尔可夫不等式,

$$\mathbb{P}_{A\sim P}\left(\left|\frac{Q(A)}{P(A)}-1\right| > \epsilon\right) \le \frac{\mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}-1\right|\right]}{\epsilon},$$

结合引理 C.8，可得最终结果

$$\mathbb{E}_{A\sim P}\left[\left|\frac{Q(A)}{P(A)}\right|\mathbb{1}_{\left|\frac{Q(A)}{P(A)}-1\right|>\epsilon}\right] \le 2\,\mathrm{TV}(P,Q) + \frac{2\,\mathrm{TV}(P,Q)}{\epsilon}. \qquad \square$$

现在，我们已具备证明定理 C.6 所需的全部要素。

定理 *C.6* 的证明。再次，我们计算单个坐标 $j \in \{1,\dots,d\}$ 的差值:

$$\begin{aligned}
&\left|g_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})_j - g^{\mathrm{clip}}_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})_j\right| \\
&= \Big|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\Big[\Big(\frac{\partial_{\theta_j}\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)} - \frac{\partial_{\theta_j}\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}-1\right|\le\epsilon}\Big)\mathbb{A}_t^{\pi_{\theta_{\mathrm{old}}}}(S_t,A_t)\Big]\Big| \\
&= \Big|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\Big[\frac{\partial_{\theta_j}\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}-1\right|>\epsilon}\mathbb{A}_t^{\pi_{\theta_{\mathrm{old}}}}(S_t,A_t)\Big]\Big| \\
&= \Big|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\Big[\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}-1\right|>\epsilon}\left(\partial_{\theta_j}\log\pi_\theta(A_t\,;\,S_t)\right)\mathbb{A}_t^{\pi_{\theta_{\mathrm{old}}}}(S_t,A_t)\Big]\Big| \\
&\le 2\Pi_* R_* \frac{1-\gamma^T}{1-\gamma}\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\Big[\Big|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\Big|\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}-1\right|>\epsilon}\Big],
\end{aligned}$$

其中我们使用假设 A.1、A.3 来界定 $\left|\partial_{\theta_j}\log\pi_\theta(A_t\,;\,S_t)\right| \le \Pi_*$ 以及 $\left|\mathbb{A}_t^{\pi_{\theta_{\mathrm{old}}}}(S_t,A_t)\right| \le 2R_*\frac{1-\gamma^T}{1-\gamma}$。现在，根据引理 C.9，对于固定的 $s\in\mathcal{S}$，我们得到

$$\mathbb{E}_{A_t\sim\pi_{\theta_{\mathrm{old}}}(\,\cdot\,;\,s)}\Big[\Big|\frac{\pi_\theta(A_t\,;\,s)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,s)}\Big|\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,s)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,s)}-1\right|>\epsilon}\Big] \le \Big(2+\frac{2}{\epsilon}\Big)\mathrm{TV}\left(\pi_{\theta_{\mathrm{old}}}(\cdot\,;\,s),\pi_\theta(\cdot\,;\,s)\right).$$

利用 Fubini定理对 S_t 的分布进行积分，可得

$$\begin{aligned}
\left|g_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})_j - g^{\mathrm{clip}}_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})_j\right| &= 2\Pi_* R_* \frac{1-\gamma^T}{1-\gamma}\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\left[\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\right|\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}-1\right|>\epsilon}\right]\\
&\le 2\Pi_* R_* \frac{1-\gamma^T}{1-\gamma}\left(2+\frac{2}{\epsilon}\right)\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\left[\mathrm{TV}\left(\pi_{\theta_{\mathrm{old}}}(\,\cdot\,;\,S_t),\pi_\theta(\,\cdot\,;\,S_t)\right)\right]\\
&\le 2\Pi_* R_* \frac{1-\gamma^T}{1-\gamma}\left(2+\frac{2}{\epsilon}\right)\frac{T}{T}\sum_{t=0}^{T-1}\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\left[\mathrm{TV}\left(\pi_{\theta_{\mathrm{old}}}(\,\cdot\,;\,S_t),\pi_\theta(\,\cdot\,;\,S_t)\right)\right]\\
&= 2\Pi_* R_* \frac{1-\gamma^T}{1-\gamma}\left(2+\frac{2}{\epsilon}\right)T\cdot \mathrm{M_{ean}TV}(\pi_{\theta_{\mathrm{old}}},\pi_\theta). \qquad \square
\end{aligned}$$

C.3 替代梯度有界

接下来，我们证明替代梯度一致有界。

命题 C.10 (替代梯度界)。 在假设 *A.1*、*A.3* 下，我们有 e

$$\left\|g^{clip}_{PPO}(\theta,\theta_{old})\right\|_\infty \le G := 2\,\Pi_* R_* \left(\frac{1-\gamma^T}{1-\gamma}\right)^2.$$

证明。 首先，回顾一下，对于有界奖励，真实优势函数 $\mathbb{A}^\pi_t$有界于 $2R_*\frac{1-\gamma^T}{1-\gamma}$。利用有界得分函数假设，即 $\|\nabla_\theta \log\,\pi_\theta(a\,;\,s)\|_\infty \le \Pi_*$，可得

$$\begin{aligned}
\left\|g^{\mathrm{clip}}_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})\right\|_\infty &= \left\|\sum_{t=0}^{T-1}\gamma^t\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\left[\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\,\mathbb{1}_{\left|\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}-1\right|\le\epsilon}\,\nabla_\theta\log\pi_\theta(A_t\,;\,S_t)\,\mathbb{A}^{\pi_{\theta_{\mathrm{old}}}}_t(S_t,A_t)\right]\right\|_\infty\\
&\le \sum_{t=0}^{T-1}\gamma^t\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\left[\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\,\|\nabla_\theta\log\pi_\theta(A_t\,;\,S_t)\|_\infty\left|\mathbb{A}^{\pi_{\theta_{\mathrm{old}}}}_t(S_t,A_t)\right|\right]\\
&\le \sum_{t=0}^{T-1}\gamma^t\Pi_* 2R_*\frac{1-\gamma^T}{1-\gamma}\,\mathbb{E}^{\pi_{\theta_{\mathrm{old}}}}\left[\frac{\pi_\theta(A_t\,;\,S_t)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,S_t)}\right].
\end{aligned}$$

此外，

$$\mathbb{E}_{A_t\sim\pi_{\theta_{\mathrm{old}}}(\,\cdot\,;s)}\left[\frac{\pi_\theta(A_t\,;\,s)}{\pi_{\theta_{\mathrm{old}}}(A_t\,;\,s)}\right] = \sum_{a\in\mathcal{A}}\pi_{\theta_{\mathrm{old}}}(a\,;\,s)\frac{\pi_\theta(a\,;\,s)}{\pi_{\theta_{\mathrm{old}}}(a\,;\,s)} = \sum_{a\in\mathcal{A}}\pi_\theta(a\,;\,s) = 1. \tag{13}$$

因此，先以 S_t 为条件，再将其积分消去，便得到

$$\left\|g^{\mathrm{clip}}_{\mathrm{PPO}}(\theta,\theta_{\mathrm{old}})\right\|_\infty \le 2\,\Pi_* R_*\frac{1-\gamma^T}{1-\gamma}\sum_{t=0}^{T-1}\gamma^t = 2\,\Pi_* R_*\left(\frac{1-\gamma^T}{1-\gamma}\right)^2.$$

$\square$

请注意，可以通过更精细的估计来界定裁剪替代梯度范数，例如对未裁剪概率 $\mathbb{P}^{\pi_{\theta_{\mathrm{old}}}}(|\ ^{\pi_\theta(A|S)}\,^{\pi_{\theta_{\mathrm{old}}}}_{(A|S)}\ -1|\le\epsilon)$进行界定。在强策略假设下，可以利用反集中不等式证明，当 θ 远离θ_{old}时，裁剪梯度范数趋于零。由于实践中的裁剪概率不会趋于零，我们采用较粗略的界。

D 收敛性证明

现在我们来证明收敛性，这一证明建立在前几节的预备工作之上。我们沿用 [18], 中提出的证明策略；该文在监督学习情形下分析了随机梯度下降的随机重排（RR），而我们将其思路拓展到强化学习框架中。为便于分析并统一合适的记号，我们对近端策略优化（PPO）的策略更新机制稍作改写。请注意，我们并不关注 PPO 的演员-评论家方面；也就是说，在随机情形下，我们假设可以获得有界且有偏的优势估计器（参见假设 A.2）。
从整体上看，PPO 算法可以描述如下。

- PPO 在每个周期 C开始时采样长度为 T 的新轨迹展开 n，并将样本展平为长度为 N 的状态-动作转移缓冲区$= nT$: 。该缓冲区还存储有偏的优势估计以及状态转移发生的相应时间，因为我们纳入了折扣。

- 在一个周期内，PPO 先执行一步优势演员-评论家，然后执行若干步裁剪重要性采样。

- 每个周期中的梯度步被划分为多个训练轮次。一个训练轮次包含 $m = \frac{N}{B}$ 个梯度步，其中每个梯度步都使用从转移缓冲区中抽取的 B 个状态转移（不放回）。开始一个训练轮次之前，会对转移缓冲区重新进行随机重排。

D.1 PPO 形式化

更正式地说，我们考虑以下算法。设定

- 周期数 C,
- 每个周期的训练轮次数 K ,
- 轨迹采样数 n,
- 状态转移批次大小 $N = nT$,
- 小批量大小 B，使得 $m := N/B \in \mathbb{N}$ 表示每个训练轮次的梯度步数
- 恒定学习率 $\eta > 0$。

对于每个周期 $c = 0, 1, \ldots\ldots, C-1$:

(i) 从 $\pi_{\theta_{c,\ 0,\ 0}}$ 采样一个包含 n 条长度为 T 的新鲜轨迹采样数据集（这就是 π_{old}），并使用这些轨迹计算（可能有偏差的）优势估计（例如，通过在真实价值函数下使用广义优势估计来计算）。将所得数据展平为转移缓冲区 $\{(s_c^i,\ a_c^i,\ r_c^i,\ \hat{\mathbb{A}}_c^i,\ t_c^i)\}_{i=0}^{N-1}$，大小为 N，其中 $(s_c^i,\ a_c^i,\ r_c^i,\ t_c^i)$ 遍历轨迹采样中遇到的所有状态-动作-奖励-时间四元组，$\hat{\ }\ \mathbb{A}_c^i$ 表示 $\mathbb{A}^{\pi_{\theta_c}}{}_{t_c^i}{}_{0\,0\,,\,(,\ a_c^i)}$ 的优势估计（假设其有偏差）。

(ii) 对于每个训练轮次 $e = 0, 1, \ldots\ldots, K-1$:

(a) 生成一个新的随机排列 $\sigma_{c,e} = (\sigma_{c,e,0}, \ldots, \sigma_{c,e,N-1})$，对转移缓冲区重新打乱。将其拆分为连续且互不重叠的小批次

$$\mathcal{B}_{c,e,k} := \{\sigma_{c,e,kB}, \ldots, \sigma_{c,e,(k+1)B-1}\}, \qquad k = 0, \ldots, m-1.$$

(b) 对训练轮次 $k = 0, \dots, m-1$ 中的每一步，计算小批量梯度估计量

$$\hat{g}_{c,e,k} := \hat{g}^{\text{clip}}(\theta_{c,e,k}, \theta_{c,0,0}; \mathcal{B}_{c,e,k}) := \frac{T}{B} \sum_{i \in \mathcal{B}_{c,e,k}} \gamma^{t_c^i} \frac{\nabla \pi_{\theta_{c,e,k}}(a_c^i \,;\, s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i \,;\, s_c^i))} \mathbb{1}_{\left| \frac{\pi_{\theta_{c,e,k}}(a_c^i \,;\, s_c^i))}{\pi_{c,0,0}(a_c^i \,;\, s_c^i)} - 1 \right| \le \epsilon} \hat{\mathbb{A}}_c^i$$

并更新

$$\theta_{c,e,t+1} = \theta_{c,e,t} + \eta \, \hat{g}_{c,e,k}.$$

(c) 设置 $\theta_{c,\ e+1,\ 0} := \theta_{c,\ e,\ m}$。

(iii) 设置 $\theta_{c+1,\ 0,\ 0} := \theta_{c,\ K,\ 0}$。

对于以下分析，我们使用裁剪代理目标

$$g_{\text{PPO}}^{\text{clip}}(\theta, \theta_{c,0,0}) := \sum_{t=0}^{T-1} \gamma^t \mathbb{E}^{\pi_{c,0,0}} \Big[\frac{\nabla \pi_\theta(A_t \,;\, S_t))}{\pi_{\theta_{c,0,0}}(A_t \,;\, S_t))} \mathbb{1}_{|\frac{\pi_\theta(A_t \,;\, S_t))}{\pi_{c,0,0}(A_t \,;\, S_t)} - 1| \le \epsilon} \mathbb{A}_t^{\pi_{\theta_{c,0,0}}}(S_t, A_t) \Big].$$

和未裁剪代理目标

$$g_{\text{PPO}}(\theta, \theta_{c,0,0}) := \sum_{t=0}^{T-1} \gamma^t \mathbb{E}^{\pi_{c,0,0}} \Big[\frac{\nabla \pi_\theta(A_t \,;\, S_t))}{\pi_{\theta_{c,0,0}}(A_t \,;\, S_t))} \mathbb{A}_t^{\pi_{\theta_{c,0,0}}}(S_t, A_t) \Big].$$

为了将符号与前文联系起来，只需回顾 $\pi_{\theta_{c,\ 0,\ 0}} =: \pi_{\theta_{\text{old}}}$。在每个周期内，我们定义裁剪代理目标的逐转移贡献

$$g_{\text{PPO}}^{(i),\text{clip}}(\theta, \theta_{c,0,0}) := T \gamma^{t_c^i} \frac{\nabla \pi_\theta(a_c^i \,;\, s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i \,;\, s_c^i))} \mathbb{1}_{|\frac{\pi_\theta(a_c^i \,;\, s_c^i))}{\pi_{c,0,0}(a_c^i \,;\, s_c^i)} - 1| \le \epsilon} \hat{\mathbb{A}}_c^i$$

以及未裁剪代理目标的逐转移贡献

$$g_{\text{PPO}}^{(i)}(\theta, \theta_{c,0,0}) := T \gamma^{t_c^i} \frac{\nabla \pi_\theta(a_c^i \,;\, s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i \,;\, s_c^i))} \hat{\mathbb{A}}_c^i$$

对于 $i = 0, \dots, N-1$。注意，在每个周期的第一步中，有 $g^{(i),\text{clip}}$ 近端策略优化 $(\theta_{c,0,0}, \theta_{c,0,0}) = g^{(i)}$ 近端策略优化$(\theta_{c,0,0}, \theta_{c,0,0})$。

D.2 确定性情形的证明，定理 5.1

我们首先分析确定性设置。这里，我们假设可以直接访问裁剪代理目标 g^{clip} 近端策略优化。对于长度为 K 的每个周期 $c = 0,...,C-1$，我们考虑迭代点

$$\begin{aligned} \theta_{c,e+1} &= \theta_{c,e} + \eta \, g_{\text{PPO}}^{\text{clip}}(\theta_{c,e}, \theta_{c,0}), \quad e = 0, \dots, K-1 \\ \theta_{c+1,0} &= \theta_{c,K} \end{aligned} \tag{14}$$

因此，在随机情形中，一次代理梯度步对应于一个训练轮次的小批量样本代理梯度步。

在确定性情形下，我们可以直接应用定理 4.2 中的偏差估计。这会得到更紧的误差界，并使我们能够证明，在许多现实场景中，PPO 相较于标准梯度上升具有优势。相比之下，对于下文讨论的随机情形，我们必须改为推导路径偏差界。具体推导将在接下来的小节中进行。

定理 *5.1* 的证明。下文中，我们将裁剪代理目标视为对精确梯度 ∇J 的有偏梯度近似，即定义

$$b_{c,e} := g_{\mathrm{PPO}}^{\mathrm{clip}}(\theta_{c,e}, \theta_{c,0}) - \nabla J(\theta_{c,e}).$$

我们将更新式 (14) 写作近似梯度上升方案

$$\theta_{c,e+1} = \theta_{c,e} + \eta(\nabla J(\theta_{c,e}) + b_{c,e}).$$

根据 J 的 L-光滑性，假设 $\eta \leq \frac{1}{L}$，则对于 $e = 0, \ldots\ldots, K-1$

$$\begin{aligned}
J(\theta_{c,e+1}) &\geq J(\theta_{c,e}) + \langle \nabla J(\theta_{c,e}), \theta_{c.e+1} - \theta_{c,e} \rangle - \frac{L}{2}|\theta_{c,e+1} - \theta_c|^2 \\
&= J(\theta_{c,e}) + \eta|\nabla J(\theta_{c,e})|^2 + \eta\langle \nabla J(\theta_{c,e}), b_{c,e} \rangle - \frac{L}{2}\eta^2|\nabla J(\theta_{c,e}) + b_{c,e}|^2 \\
&= J(\theta_{c,e}) + (1 - \frac{L\eta}{2})\eta|\nabla J(\theta_{c,e})|^2 - (1 - L\eta)\eta\langle \nabla J(\theta_{c,e}), b_{c,e} \rangle - \frac{L}{2}\eta^2|b_{c,e}|^2 \\
&\geq J(\theta_{c,e}) + \Big(1 - \frac{1}{2\delta} - \frac{L}{2}\Big(1 - \frac{1}{\delta}\Big)\eta\Big)\eta|\nabla J(\theta_{c,e})|^2 - \Big((1 - L\eta)\eta\frac{\delta}{2} + \frac{L}{2}\eta^2\Big)|b_{c,e}|^2,
\end{aligned}$$

其中，最后一个不等式对所有 $\delta > 0$ 均成立，这是杨氏不等式的结果。特别地，对于 $\delta = 1$，我们可推出

$$J(\theta_{c,e+1}) \geq J(\theta_{c,e}) + \frac{\eta}{2}|\nabla J(\theta_{c,e})|^2 - \frac{\eta}{2}|b_{c,e}|^2.$$

根据定理 4.2，我们可以通过以下方式控制偏差项

$$|b_{c,e}| \leq R|\theta_{c,e} - \theta_{c,0}|$$

其中某个 $R > 0$。此外，根据命题 C.10，存在 $G > 0$ 使得

$$|g_{\mathrm{PPO}}^{\mathrm{clip}}(\theta_{c,e}, \theta_{c,0})| \leq G$$

对于任意 $c = 0, \ldots\ldots, C-1$ 和 $e = 0, \ldots\ldots, K-1$。这意味着

$$|b_{c,e}| \leq R|\theta_{c,e} - \theta_{c,0}| \leq R\sum_{e'=0}^{e-1}|\theta_{c,e'+1} - \theta_{c,e'}| \leq \eta eRG$$

因此，

$$J(\theta_{c,e+1}) \geq J(\theta_{c,e}) + \frac{\eta}{2}|\nabla J(\theta_{c,e})|^2 - \frac{\eta^3}{2}e^2R^2G^2.$$

重新整理该不等式，并对所有 $c = 0, \ldots, C-1, e = 0, \ldots, K-1$ 求和，可得

$$\begin{aligned}
\frac{\eta}{2}\sum_{c=0}^{C-1}\sum_{e=0}^{K-1}|\nabla J(\theta_{c,e}|^2 &\leq \sum_{c=0}^{C-1}\sum_{e=0}^{K-1}\big(J(\theta_{c,e+1}) - J(\theta_{c,e})\big) + \sum_{c=0}^{C-1}\sum_{e=0}^{K-1}\frac{\eta^3}{2}e^2R^2G^2 \\
&\leq \Delta_0 + \frac{\eta^3}{12}C(K-1)K(2K-1)R^2G^2
\end{aligned}$$

其中我们应用了望远镜求和 $\sum_{c=0}^{C-1}\sum_{e=0}^{K-1}\big(J(\theta_{c,e+1}) - J(\theta_{c,e})\big) = J(\theta_{C,K}) - J(\theta_{0,0})$，而 $\Delta_0 := J_* - J(\theta_{0,0})$ 表示初始最优性差距。将两边除以 $\frac{\eta}{2CK}$，得到

$$\begin{aligned}\min_{0\le c\le C-1,\ 0\le e\le K-1} |\nabla J(\theta_{c,e})|^2 &\le \frac{1}{CK}\sum_{c=0}^{C-1}\sum_{e=0}^{K-1} |\nabla J(\theta_{c,e}|^2 \\ &\le \frac{2\Delta_0}{\eta CK} + \frac{1}{6}\eta^2(K-1)(2K-1)R^2G^2\,.\end{aligned}$$

□

特别是，在优化上界时

$$\frac{2\Delta_0}{\eta CK} + \frac{1}{6}\eta^2(K-1)(2K-1)R^2G^2$$

关于 η，我们得到

$$\eta^* = \min\Big(\frac{1}{L}, \Big(\frac{6\,\Delta_0}{CK(K-1)(2K-1)\,R^2G^2}\Big)^{\frac{1}{3}}\Big)$$

在 $\eta^* < \frac{1}{L}$ 的情况下，这给出了相应的上界

$$\min_{0\le c\le C-1,\ 0\le e\le K-1} |\nabla J(\theta_{c,e})|^2 \le \Big(\frac{9}{2}\,\frac{\Delta_0^2(K-1)(2K-1)\,R^2G^2}{(CK)^2}\Big)^{1/3}.$$

为简化常数，我们还优化了较弱的上界

$$f(\eta) := \frac{2\Delta_0}{\eta CK} + \frac{1}{3}\eta^2K^2R^2G^2$$

这给出 $\eta^* = \min(\frac{1}{L}, \frac{c}{K})$，其中

$$c = \Big(\frac{3\,\Delta_0}{CR^2G^2}\Big)^{\frac{1}{3}}. \tag{15}$$

对于 $\eta^* < \frac{1}{L}$，我们得到

$$f(\eta^*) = \left(\frac{3\Delta_0 RG}{C}\right)^{\frac{2}{3}}.$$

类似地，让我们针对周期长度 K 优化上界。首先，注意到 $K = 1$ 恢复了经典梯度上升速率

$$\min_{0\le c\le C-1,\ 0\le e\le K-1} |\nabla J(\theta_{c,e})|^2 \le \frac{2\Delta_0}{\eta CK}.$$

然而，在某些情况下，$K > 1$ 的表现优于梯度上升速率。暂且假设 $K \in (0, \infty)$是一个连续变量，并再次针对 $K > 1$ 优化简化后的（较弱）上界

$$f(K) := \frac{2\Delta_0}{\eta CK} + \frac{1}{3}\eta^2K^2R^2G^2$$

。此时，最优周期长度为$K^* = \frac{c}{\eta}$，其中 c 由式（15）给出。将其代回收敛速率后，再次得到

$$f(K^*) = \left(\frac{3\Delta_0 RG}{C}\right)^{\frac{2}{3}}.$$

因此，对所有$\eta \le \frac{1}{L}$，我们得到

$$\min_{0\le c\le C-1,\ 0\le e\le K-1} |\nabla J(\theta_{c,e})|^2 \le \min\left(\frac{2\Delta_0}{\eta CK}, f(\lfloor K^*\rfloor), f(\lceil K^*\rceil)\right).$$

总之，相对于标准梯度上升，在每个周期中纳入多个有偏梯度步（即选择 $K > 1$）会在 Δ_0 较大且参数 η、 R、 G和 C 较小的情形下产生更快的收敛。

D.3 随机情形的重要性质

设$(\mathcal{F}_c)_{c=0,\dots,C}$ 为当前周期之前由迭代点 θ 生成的典范滤过，即

$$\mathcal{F}_c = \sigma(\theta_{c',e,k} : c' = 0,\dots,c-1, e = 0,\dots,K-1, k = 0,\dots,m) \quad c = 0,\dots,C.$$

注意，$\theta_{c,0,0} = \theta_{c-1,K-1,m}$ 对所有 $c = 1,\dots,C$ 是 $\mathcal{F}_c$-可测的，并回顾我们有

$$\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0}) = \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i)}(\theta_{c,0,0},\theta_{c,0,0})$$

因为第一个周期步没有截断。

引理 D.1 (全批量方差)。 在假设 *A.2* 和 *A.3* 下，存在 $\sigma > 0$ 使得

$$\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{PPO}^{(i)}(\theta_{c,0,0},\theta_{c,0,0}) - g_{PPO}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \,\Big|\, \mathcal{F}_c\Big] \le 2\frac{\sigma^2}{N} + 2T^2\Pi_*^2\delta^2$$

其中 $\sigma^2 = \frac{1}{T}\Pi_*^2 A_*^2\left(\frac{1-\gamma^T}{1-\gamma}\right)^2$。

证明。 在全批量设置下，情况很简单。根据转移缓冲区的定义，所有转移估计量之和等于 n （相互独立的）轨迹采样之和。为清楚起见，我们先给出无偏情形（$\delta = 0$）的论证，此时有

$$\mathbb{E}\Big[\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0}) \,\Big|\, \mathcal{F}_c\Big] = \frac{1}{N}\sum_{i=0}^{N-1}\mathbb{E}\Big[g_{\mathrm{PPO}}^{(i)}(\theta_{c,0,0},\theta_{c,0,0}) \,\Big|\, \mathcal{F}_c\Big] = g_{\mathrm{PPO}}(\theta_{c,0,0},\theta_{c,0,0}). \quad (16)$$

根据马尔可夫性质，在周期起点处，我们可以写成

$$\begin{aligned}&\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i)}(\theta_{c,0,0},\theta_{c,0,0}) - g_{\mathrm{PPO}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \,\Big|\, \mathcal{F}_c\Big]\\ &= \frac{1}{N^2}\operatorname{Var}^{\pi_{\theta_{c,0,0}}}\Big[\sum_{j=0}^{n-1}\sum_{t=0}^{T-1} T\gamma^t \nabla\log\pi_{\theta_{c,0,0}}(A_t^j\,;\,S_t^j)\,\hat{\mathbb{A}}_t^j\Big],\end{aligned}$$

其中，$(S^1,\ A^1)$.....，$(S^n,\ A^n)$ 是在 $\pi_{\theta_{c,\ 0,\ 0}}$ 下马尔可夫决策过程的 n 个独立同分布副本，其优势估计为 $\hat{}\mathbb{A}^1$，.....，$\hat{}\mathbb{A}^n$。利用独立性和 $N = n \cdot T$，再结合得分函数与优势估计器的界，可以得到

$$
\begin{aligned}
&\frac{1}{N^2}\sum_{j=0}^{n-1}\mathrm{Var}^{\pi_{\theta_{c,0,0}}}\Big[\sum_{t=0}^{T-1}\gamma^t\nabla\log\pi_{\theta_{c,0,0}}(A_t^j\,;\,S_t^j)\,\hat{\mathbb{A}}_t^j\Big]\\
&=\frac{n}{N^2}\,\mathrm{Var}^{\pi_{\theta_{c,0,0}}}\Big[\sum_{t=0}^{T-1}\gamma^t\nabla\log\pi_{\theta_{c,0,0}}(A_t^1\,;\,S_t^1)\,\hat{\mathbb{A}}_t^1\Big]\\
&\le\frac{n}{N^2}\mathbb{E}^{\pi_{\theta_{c,0,0}}}\Big[\Big(\sum_{t=0}^{T-1}\gamma^t\nabla\log\pi_{\theta_{c,0,0}}(A_t^1\,;\,S_t^1)\,\hat{\mathbb{A}}_t^1\Big)^2\Big]\\
&\le\frac{n}{N^2}(\Pi_*A_*)^2\Big(\sum_{t=0}^{T-1}\gamma^t\Big)^2\\
&=\frac{1}{TN}\Pi_*^2A_*^2\Big(\frac{1-\gamma^T}{1-\gamma}\Big)^2.
\end{aligned}
$$

最后，由假设 A.2 和马尔可夫性质可得

$$
\mathbb{E}[|\mathbb{E}[\hat{\mathbb{A}}_t^j\mid\mathcal{F}_c,A_t^j,S_t^j]-\mathbb{A}_t^{\pi_{\theta_{c,0,0}}}(A_t^j,S_t^j)|^2\mid\mathcal{F}_c]=\mathbb{E}^{\pi_{\theta_{c,0,0}}}[|\mathbb{E}^{\pi_{\theta_{c,0,0}}}[\hat{\mathbb{A}}_t^j\mid A_t^j,S_t^j]-\mathbb{A}_t^{\pi_{\theta_{c,0,0}}}(A_t^j,S_t^j)|^2]\le\delta^2 \tag{17}
$$

因此

$$
\begin{aligned}
&\mathbb{E}[|T\gamma^t\nabla\log\pi_{\theta_{c,0,0}}(A_t^j\,;\,S_t^j)(\mathbb{E}[\hat{\mathbb{A}}_t^j\mid\mathcal{F}_c,A_t^j,S_t^j]-\mathbb{A}_t^{\pi_{\theta_{c,0,0}}}(A_t^j,S_t^j))|^2\mid\mathcal{F}_c]\\
&\le T^2\Pi_*^2\mathbb{E}[|\mathbb{E}[\hat{\mathbb{A}}_t^j\mid\mathcal{F}_c,A_t^j,S_t^j]-\mathbb{A}_t^{\pi_{\theta_{c,0,0}}}(A_t^j,S_t^j)|^2\mid\mathcal{F}_c]\\
&\le T^2\Pi_*^2\delta^2
\end{aligned}
$$

因此，利用 $(a+b)^2\le 2a^2+2b^2$，该结论成立，其中

$$
\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=0}^{N-1}g_{\mathrm{PPO}}^{(i)}(\theta_{c,0,0},\theta_{c,0,0})-g_{\mathrm{PPO}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2\,\Big|\,\mathcal{F}_c\Big]\le 2\frac{\sigma^2}{N}+2T^2\Pi_*^2\delta^2\,. \qquad \square
$$

引理 D.2 (有界漂移)。 在假设 *A.2* 和 *A.3* 下，对所有 (c,e,k)

$$
\big|\hat{g}^{clip}(\theta_{c,e,k},\theta_{c,0,0})\big|\le G,\qquad \textit{almost surely.}
$$

其中 $G=\ T(1+\epsilon)\,\Pi_*\,A_*$。

证明。 根据定义，

$$
\begin{aligned}
\big|\hat{g}^{\mathrm{clip}}(\theta_{c,e,k},\theta_{c,0,0})\big|&=\Big|\frac{T}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\gamma^{t_c^i}\frac{\nabla\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\mathbb{1}_{\left|\frac{\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}-1\right|\le\epsilon}\hat{\mathbb{A}}_c^i\Big|\\
&=\Big|\frac{T}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\gamma^{t_c^i}\frac{\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\mathbb{1}_{\left|\frac{\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}-1\right|\le\epsilon}\nabla\log\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)\hat{\mathbb{A}}_c^i\Big|\\
&\le\frac{T}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\gamma^{t_c^i}\frac{\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\mathbb{1}_{\left|\frac{\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}-1\right|\le\epsilon}\big|\nabla\log\pi_{\theta_{c,e,k}}(a_c^i\,;\,s_c^i)\big|\big|\hat{\mathbb{A}}_c^i\big|.
\end{aligned}
$$

此外，裁剪指示函数意味着 $\frac{\pi_{\theta_{c,e,k}}(a_c^i;\, s_c^i)}{\pi_{\theta_{c,0,0}}(a_c^i;\, s_c^i)} \leq (1+\epsilon)$ 几乎处处成立。结合对得分（参见假设 A.3 ）和优势估计（参见假设 A.2 ）的假设界限，即 $|\nabla \log \pi_{\theta_{c,e,k}}(a_c^i;\ s_c^i)| \leq \Pi_*$, 以及 $|\hat{\mathbb{A}}_c^i| \leq A_*$, 可得，几乎必然

$$\left|\hat{g}^{\text{clip}}(\theta_{c,e,k}, \theta_{c,0,0})\right| \leq \frac{T}{B} \sum_{i \in \mathcal{B}_{c,e,k}} \gamma^{t_c^i} (1+\epsilon)\, \Pi_*\, A_* = T(1+\epsilon)\, \Pi_*\, A_*.$$

□

引理 D.3 (路径级偏差分解)。 对于 $t = 0$、.....、$T-1$、 $s \in \mathcal{S}$、 $a \in \mathcal{A}$ 以及任意满足 $|\mathbb{A}| \leq A_*$ 的标量 $\mathbb{A}$，我们定义

$$\hat{g}_{PPO}^{clip}(\theta, \theta_{old}, t, a, s, \mathbb{A}) := T\gamma^t \frac{\nabla \pi_\theta(a\,;\, s)}{\pi_{\theta_{old}}(a\,;\, s)} \mathbb{1}_{|\frac{\pi_\theta(a\,;\, s))}{\pi_{old}(a\,;\, s)} - 1| \leq \epsilon} \mathbb{A}\,.$$

那么，根据 定理 *A.3* 和定理 *A.4*，对于任意 $t \leq T-1$、 $s \in \mathcal{S}$、 $a \in \mathcal{A}$ ，有 t

$$\begin{aligned}
&\left|\hat{g}_{PPO}^{clip}(\theta, \theta_{old}, t, a, s, \mathbb{A}) - \hat{g}_{PPO}^{clip}(\theta_{old}, \theta_{old}, t, a, s, \mathbb{A})\right| \\
&\qquad \leq B_1 |\theta - \theta_{old}| + B_2 \min(\epsilon, |r_{\theta, \theta_{old}}(s,a) - 1|) + B_3 \mathbb{1}_{|r_{\theta, \theta_{old}}(s,a) - 1| > \epsilon}\,,
\end{aligned}$$

其中 $r_{\theta,\ \theta_{old}}(s, a) := \frac{\pi_\theta(a\,;\ s)}{\pi_{old}(a\,;\ s)}$， $B_1 = TA_*L_s$，且 $B_2 = B_3 = TA_*\Pi_*$。

证明。 根据 $\hat{}g^{\text{clip}}$ 近端策略优化$(\theta,\ \theta_{\text{old}},\ t,\ a,\ s,\ \mathbb{A})$ 的定义以及事实 $|\mathbb{A}| \leq A_*$，我们有

$$\begin{aligned}
&\left|\hat{g}_{\text{PPO}}^{\text{clip}}(\theta, \theta_{\text{old}}, t, a, s, \mathbb{A}) - \hat{g}_{\text{PPO}}^{\text{clip}}(\theta_{\text{old}}, \theta_{\text{old}}, t, a, s, \mathbb{A})\right| \\
&\leq TA_*\gamma^t \left| r_{\theta,\theta_{\text{old}}}(s,a) \nabla \log \pi_\theta(s\,;\, a) \mathbb{1}_{|r_{\theta,\theta_{\text{old}}}(s,a) - 1| \leq \epsilon} - \nabla \log \pi_{\theta_{\text{old}}}(s\,;\, a) \right| \\
&\leq TA_* \left| r_{\theta_{\text{old}},\theta_{\text{old}}}(s,a) (\nabla \log \pi_{\theta_{\text{old}}}(s\,;\, a) - \nabla \log \pi_\theta(s\,;\, a)) \right| \big) \\
&\quad + TA_* \left| \nabla \log \pi_\theta(s\,;\, a) (r_{\theta,\theta_{\text{old}}}(s,a) \mathbb{1}_{|r_{\theta,\theta_{\text{old}}}(s,a) - 1| \leq \epsilon} - r_{\theta_{\text{old}},\theta_{\text{old}}}(s,a)) \right| \\
&= TA_* \left| \nabla \log \pi_{\theta_{\text{old}}}(s\,;\, a) - \nabla \log \pi_\theta(s\,;\, a) \right| \\
&\quad + TA_* \left| \nabla \log \pi_\theta(s\,;\, a) (r_{\theta,\theta_{\text{old}}}(s,a) \mathbb{1}_{|r_{\theta,\theta_{\text{old}}}(s,a) - 1| \leq \epsilon} - 1) \right| \\
&\leq TA_* L_s |\theta - \theta_{\text{old}}| + TA_* \Pi_* |r_{\theta,\theta_{\text{old}}}(s,a) \mathbb{1}_{|r_{\theta,\theta_{\text{old}}}(s,a) - 1| \leq \epsilon} - 1|,
\end{aligned}$$

这里我们用到了定理 A.3 和定理 A.4。最后，我们注意到

$$\left| r_{\theta,\theta_{\text{old}}}(s,a) \mathbb{1}_{|r_{\theta,\theta_{\text{old}}}(s,a) - 1| \leq \epsilon} - 1 \right| \leq \begin{cases} \min(\epsilon, |r_{\theta,\theta_{\text{old}}}(s,a) - 1|) & : r_{\theta,\theta_{\text{old}}}(s,a) \in [1-\epsilon, 1+\epsilon] \\ 1 & : r_{\theta,\theta_{\text{old}}}(s,a) \notin [1-\epsilon, 1+\epsilon] \end{cases},$$

由此便利用 $B_1 = TA_*L_s$, $B_2 = B_3 = TA_*\Pi_*$ 完成证明。 □

注 D.4. 我们将用到 $\min(\epsilon, r_{\theta,\theta_{\text{old}}}(s,a)) \leq \epsilon^{1-p} \cdot |r_{\theta,\theta_{\text{old}}}(s,a) - 1|^p$，其中 $p \in$ 为任意的 $(0,1)$。现在我们将更仔细地考察路径层面偏差分解中的上界。

引理 D.5 （利普希茨策略）。 在假设 *A.4* 下，我们有 $\theta \mapsto \pi_\theta$ 是一致利普希茨连续的，具体来说，

$$\left|\pi_\theta(a\,;\, s) - \pi_{\theta'}(a\,;\, s)\right| \leq \Pi_* \left|\theta - \theta'\right|, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, \theta, \theta' \in \mathbb{R}^d.$$

证明。 根据链式法则， $\nabla\pi_\theta(a;s) = \pi_\theta(a;s)\,\nabla\log\,\pi_\theta(a;s)$。注意， $0 \le \pi_\theta(a;\,s) \le 1$ 且$|\nabla\log(\pi_\theta(a;\,s))| \le \Pi_*$。 因此， $|\nabla\log\,\pi_\theta(a;\,s)| \le \Pi_*$，由中值定理可得利普希茨连续性。 □

我们估计裁剪概率，它出现在周期内路径级偏差分解的上界中（引理 D.5）。

引理 D.6 (有界权重)。 在假设 *A.4* 下，对于所有 $q \in (0,1)$

(i) 对于任意 (c,e) ，均有

$$\frac{1}{m}\sum_{k=0}^{m-1}\mathbb{P}\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\big|r_{\theta_{c,e,k},\theta_{c,0,0}}(a_c^i,s_c^i)-1\big|>\epsilon\,\big|\,\mathcal{F}_c\big)\le\frac{|\mathcal{A}|^q\,\Pi_*^q\,\sum_{k=0}^{m-1}\mathbb{E}[|\theta_{c,e,k}-\theta_{c,0,0}|^{\frac{q}{1-q}}\mid\mathcal{F}_c]^{1-q}}{\epsilon^q}.$$

类似地，对于任意 (c,e,k) ，均有

$$\frac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}\big[\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\big|r_{\theta_{c,e,k},\theta_{c,0,0}}(a_c^i,s_c^i)-1|\big)^q\,\big|\,\mathcal{F}_c\big]\le|\mathcal{A}|^q\Pi_*^q\sum_{k=0}^{m-1}\mathbb{E}\big[\big|\theta_{c,e,k}-\theta_{c,0,0}|^{\frac{q}{1-q}}\big|\,\mathcal{F}_c\big]^{1-q}.$$

(ii) 对于任意 (c,e) ，均有

$$\frac{1}{N}\sum_{i=0}^{N-1}\mathbb{P}\big(\big|r_{\theta_{c,e,k},\theta_{c,0,0}}(a_c^i,s_c^i)-1\big|>\epsilon\,\big|\,\mathcal{F}_c\big)\le\frac{|\mathcal{A}|^q\,\Pi_*^q\,\mathbb{E}[|\theta_{c,e,k}-\theta_{c,0,0}|^{\frac{q}{1-q}}\mid\mathcal{F}_c]^{1-q}}{\epsilon^q}.$$

类似地，对于任意 (c,e,k) ，均有

$$\frac{1}{N}\sum_{i=0}^{N-1}\mathbb{E}\big[\big|r_{\theta_{c,e,k},\theta_{c,0,0}}(a_c^i,s_c^i)-1|^q\,\big|\,\mathcal{F}_c\big]\le|\mathcal{A}|^q\Pi_*^q\mathbb{E}\big[\big|\theta_{c,e,k}-\theta_{c,0,0}|^{\frac{q}{1-q}}\big|\,\mathcal{F}_c\big]^{1-q}.$$

证明。 固定 $q\in(0,\,1)$，并令 $k\in\{0,\dots,m-1\}$ 为任意值。首先，我们应用马尔可夫不等式

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}|r_{\theta_{c,e,k},\theta_{c,0,0}}(a_c^i,s_c^i)-1|>\epsilon\mid\mathcal{F}_c\Big)&\le\frac{\mathbb{E}[\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}|r_{\theta_{c,e,k},\theta_{c,0,0}}(a_c^i,s_c^i)-1|\big)^q\mid\mathcal{F}_c]}{\epsilon^q}\\&=\frac{\mathbb{E}\Big[\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{|\pi_{\theta_{c,e,k}}(a_c^i;s_c^i)-\pi_{\theta_{c,0,0}}(a_c^i;s_c^i)|}{\pi_{\theta_{c,0,0}}(a_c^i;s_c^i)}\big)^q\mid\mathcal{F}_c\Big]}{\epsilon^q}.\end{aligned}$$

利用 D.5 中的策略 Lipschitz 性质，我们有

$$\begin{aligned}&\frac{\mathbb{E}\Big[\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{|\pi_{\theta_{c,e,k}}(a_c^i;s_c^i)-\pi_{\theta_{c,0,0}}(a_c^i;s_c^i)|}{\pi_{\theta_{c,0,0}}(a_c^i;s_c^i))}\big)^q\,\Big|\,\mathcal{F}_c\Big]}{\epsilon^q}\\\le&\frac{\mathbb{E}\Big[\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{\Pi_*|\theta_{c,e,k}-\theta_{c,0,0}|}{\pi_{\theta_{c,0,0}}(a_c^i;s_c^i)}\big)^q\,\Big|\,\mathcal{F}_c\Big]}{\epsilon^q}\\=&\,\Pi_*^q\frac{\mathbb{E}\Big[|\theta_{c,e,k}-\theta_{c,0,0}|^q\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i;s_c^i)}\big)^q\,\Big|\,\mathcal{F}_c\Big]}{\epsilon^q},\end{aligned}$$

其中我们利用了 $|\theta_{c,e,k}-\theta_{c,0,0}|$ 与 $B_{c,e,k}$ 相互独立。接下来，
我们应用（条件）Hölder不等式，得到

$$\mathbb{E}\Big[|\theta_{c,e,k}-\theta_{c,0,0}|^q\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\big)^q\mid\mathcal{F}_c\Big]$$

$$\le\mathbb{E}[|\theta_{c,e,k}-\theta_{c,0,0}|^{qs}\mid\mathcal{F}_c]^{1/s}\,\mathbb{E}[\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\big)^{\frac{qs}{s-1}}\mid\mathcal{F}_c]^{\frac{s-1}{s}},$$

其中 $s=1+\frac{q}{1-q}>1$，由此得到关系 $qs=\frac{q}{1-q}$，$\frac{1}{s}=1-q$ 和 $\frac{s-1}{s}=q$。因此，

$$\mathbb{E}\Big[|\theta_{c,e,k}-\theta_{c,0,0}|^q\big(\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\big)^q\mid\mathcal{F}_c\Big]$$

$$\le\mathbb{E}[|\theta_{c,e,k}-\theta_{c,0,0}|^{\frac{q}{1-q}}\mid\mathcal{F}_c]^{1-q}\,\mathbb{E}\Big[\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\mid\mathcal{F}_c\Big]^q,$$

最后，我们使用Jensen不等式，得到

$$\begin{aligned}\frac{1}{m}\sum_{k=1}^{m-1}\mathbb{E}\Big[\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\mid\mathcal{F}_c\Big]^q&\le\mathbb{E}\Big[\frac{1}{m}\sum_{k=1}^{m-1}\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\,\Big|\,\mathcal{F}_c\Big]^q\\&=\mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N-1}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\,\Big|\,\mathcal{F}_c\Big]^q.\end{aligned}$$

由于在给定 $\mathcal{F}_c$ 的条件下，$(s_c^i,\ a_c^i)_{i=1,\ \ldots,\ N-1}$ 是使用策略$\pi_{\theta_{c,\ 0,\ 0}}$运行马尔可夫决策过程的 n 个相互独立的运行，因此我们得到

$$\mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N-1}\frac{1}{\pi_{\theta_{c,0,0}}(a_c^i\,;\,s_c^i)}\,\Big|\,\mathcal{F}_c\Big]=\frac{1}{T}\sum_{t=0}^{T-1}\sum_{s\in\mathcal{S}}\mathbb{E}^{\pi_{\theta_{c,0,0}}}\Big[\mathbb{1}_{S_t=s}\frac{1}{\pi_{\theta_{c,0,0}}(A_t\,;\,S_t)}\Big]=|\mathcal{A}|.$$

其余三个结论可通过类似的论证得出。 □

引理 D.7 （L^2-第 个训练轮次的累积漂移控制）。 在假设 *A.2* 和 *A.3* 下，对所有（c,e,k）以及 $p>0$，都有

$$\mathbb{E}\big[|\theta_{c,e,0}-\theta_{c,0,0}|^p\,\big|\,\mathcal{F}_c\big]\le\eta^p e^p m^p G^p$$

and

$$\mathbb{E}\Big[\frac{1}{m}\sum_{k=0}^{m-1}|\theta_{c,e,k}-\theta_{c,0,0}|^p\,\Big|\,\mathcal{F}_c\Big]\le\eta^p(e+1)^p m^p G^p\,,$$

其中 $G=T\ (1+\epsilon)\ \Pi_*\,A_{*\circ}$

证明。 根据采用恒定学习率的 PPO 迭代定义（将 $e-1$已完成训练轮次的梯度与当前训练轮次的部分梯度相加），我们有

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{m}\sum_{k=0}^{m-1}|\theta_{c,e,k}-\theta_{c,0,0}|^p\,\Big|\,\mathcal{F}_c\Big]&=\frac{\eta^p}{m}\sum_{k=0}^{m-1}\mathbb{E}\Big[\Big|\sum_{e'=0}^{e-1}\sum_{k'=0}^{m-1}\hat{g}_{c,e',k'}+\sum_{k'=0}^{k-1}\hat{g}_{c,e,k'}\Big|^p\,\Big|\,\mathcal{F}_c\Big]\\&\le\eta^p(e+1)^p m^p G^p,\end{aligned}$$

这里用到了引理 D.2。第一个结论只需考虑后一求和中的第一项即可得出。

□

D.4 随机情形（近端策略优化）的证明，定理 6.2

我们首先证明固定周期内的上升性质。一个关键要素是命题 B.5 中所示的 J 的 L-光滑性。这部分证明受到 [18], 中研究的随机梯度下降设定的启发，我们研究一个训练轮次中所有迭代的综合上升效应。

引理 D.8 （每轮上升性质）。 设 $\eta \leq \frac{1}{Lm}$，则对于每个周期 $c = 0, \ldots, C-1$ 以及每个训练轮次 $e = 0, \ldots, K-1$， 以下结论几乎必然成立:

$$J(\theta_{c,e+1,0}) \geq J(\theta_{c,e,0}) + \frac{\eta m}{2}|\nabla J(\theta_{c,e,0})|^2 - \frac{\eta m}{2}\left|\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2.$$

证明。 根据上升引理，在 J 的 L-光滑性条件下（命题 B.5），我们有

$$\begin{aligned}
&J(\theta_{c,e+1,0}) \\
&\geq J(\theta_{c,e,0}) - \langle \nabla J(\theta_{c,e,0}), \theta_{c,e+1,0} - \theta_{c,e,0}\rangle - \frac{L}{2}|\theta_{c,e+1,0} - \theta_{c,e,0}|^2 \\
&= J(\theta_{c,e,0}) + \eta m \langle \nabla J(\theta_{c,e,0}), \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\rangle - \frac{\eta^2 m^2 L}{2}\left|\frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2 \\
&= J(\theta_{c,e,0}) + \frac{\eta m}{2}\Big(|\nabla J(\theta_{c,e,0})|^2 + |\frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}|^2 - |\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}|^2\Big) - \frac{\eta^2 m^2 L}{2}\left|\frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2 \\
&= J(\theta_{c,e,0}) + \frac{\eta m}{2}|\nabla J(\theta_{c,e,0})|^2 + \frac{\eta m}{2}(1 - L\eta m)\left|\frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2 - \frac{\eta m}{2}\left|\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2 \\
&\geq J(\theta_{c,e,0}) + \frac{\eta m}{2}\left|\nabla J(\theta_{c,e,0})\right|^2 - \frac{\eta m}{2}\left|\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2,
\end{aligned}$$

其中根据关于 η 的假设，我们使用了 $1 - L\eta m \geq 0$ 。 □

为了推导近端策略优化的收敛速率，我们还需要对以下量给出上界

$$\mathbb{E}\Big[\Big|\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\Big|^2 \,\Big|\, \mathcal{F}_c\Big].$$

为此，我们将其分解为

$$\begin{aligned}
\left|\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2 \leq\ & 2\left|\nabla J(\theta_{c,e,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0}, \theta_{c,0,0})\right|^2 \\
& + 2\left|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0}, \theta_{c,0,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\right|^2
\end{aligned} \tag{18}$$

并分别考虑这两项。

引理 D.9。 在假设 *A.2*、*A.3* 和 *A.4* 下，对于 $p, q \in (0,1)$ 以及任意 (c, e)，有

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\nabla J(\theta_{c,e,0}) - \frac{1}{N}\sum_{i=1}^{N-1} g_{PPO}^{(i),clip}(\theta_{c,e,0}, \theta_{c,0,0})\Big|^2 \,\Big|\, \mathcal{F}_c\Big] \\
&\le \frac{3\sigma^2}{N} + 3L^2\mathbb{E}\big[\big|\theta_{c,e,0} - \theta_{c,0,0}\big|^2 \,\big|\, \mathcal{F}_c\big] \\
&\quad + 9B_1^2\mathbb{E}[|\theta_{c,e,0} - \theta_{c,0,0}|^2 \mid \mathcal{F}_c] + 9B_2^2\epsilon^{2(1-p)}\frac{1}{N-1}\sum_{i=0}^{N-1}\mathbb{E}\big[|r_{\theta_{c,e,0},\theta_{c,0,0}}(a_c^i, s_c^i) - 1|^p \big|\, \mathcal{F}_c\big] \\
&\quad + 9B_3^2\frac{1}{N-1}\sum_{i=0}^{N-1}\mathbb{P}\big(|r_{\theta_{c,e,0},\theta_{c,0,0}}(a_c^i, s_c^i) - 1| > \epsilon \,\big|\, \mathcal{F}_c\big)\,.
\end{aligned}
$$

特别地，我们有

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\nabla J(\theta_{c,e,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{PPO}^{(i),clip}(\theta_{c,e,0}, \theta_{c,0,0})\Big|^2 \,\Big|\, \mathcal{F}_c\Big] \\
\le& 3\eta^2(3B_1^2 + L^2)K^2m^2G^2 + 9\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} \\
&+ \frac{9\eta^q B_3^2|\mathcal{A}|^q\Pi_*^q K^q m^q G^q}{\epsilon^q} + \frac{6\sigma^2}{N} + 6T^2\Pi_*^2\delta^2,
\end{aligned}
$$

其中常数定义于上述引理中。

证明。 我们进一步分解

$$
\begin{aligned}
\Big|\nabla J(\theta_{c,e,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0}, \theta_{c,0,0})\Big|^2 \le\ & 3\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0}, \theta_{c,0,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0}, \theta_{c,0,0})\Big|^2 \\
&+ 3\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0}, \theta_{c,0,0}) - \nabla J(\theta_{c,0,0})\Big|^2 \\
&+ 3\Big|\nabla J(\theta_{c,0,0}) - \nabla J(\theta_{c,e,0})\Big|^2.
\end{aligned}
$$

对于第二项，注意到在周期开始时，裁剪概率为 0，即 $g^{(i),\ \mathrm{clip}}$ 近端策略优化 $(\theta_{c,\ 0,\ 0},\ \theta_{c,\ 0,\ 0}) = g^{(i)}$ 近端策略优化$(\theta_{c,\ 0,\ 0},\ \theta_{c,\ 0,\ 0})$ 对所有 $i = 0,\ \ldots,\ N-1$ 均成立。因此，我们应用引理 D.1 来界定

$$
3\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0}, \theta_{c,0,0}) - \nabla J(\theta_{c,0,0})\Big|^2 \,\Big|\, \mathcal{F}_c\Big] \le \frac{6\sigma^2}{N} + 6T^2\Pi_*^2\delta^2\,.
$$

对于第三项，我们利用 L-光滑性来处理 J，

$$
3\mathbb{E}\big[\big|\nabla J(\theta_{c,0,0}) - \nabla J(\theta_{c,e,0})\big|^2 \mid \mathcal{F}_c\big] \le 3L^2\mathbb{E}[|\theta_{c,e,0} - \theta_{c,0,0}|^2 \mid \mathcal{F}_c]\,.
$$

根据引理 D.7，我们可以进一步将此表达式界定为

$$
3\mathbb{E}\big[\big|\nabla J(\theta_{c,0,0}) - \nabla J(\theta_{c,e,0})\big|^2 \,\big|\, \mathcal{F}_c\big] \le 3\eta^2L^2K^2m^2G^2\,.
$$

对于第一项，我们使用引理 D.3，其中抽象对象 $\mathbb{A}$ 由估计 $\hat{}\,\mathbb{A}_c^i$ 给出，并结合假设 A.2 以及对于任意 $p \in (0,\ 1)$，$\min(x,\ y) \le x^p y^{1-p}$ 这一事实，

$$
\begin{aligned}
&3\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \\
\le& 9B_1^2\frac{1}{N}\sum_{i=0}^{N-1}|\theta_{c,e,0}-\theta_{c,0,0}|^2 + 9B_2^2\epsilon^{2(1-p)}\frac{1}{N}\sum_{i=0}^{N-1}|r_{\theta_{c,e,0},\theta_{c,0,0}}(s_c^i,a_c^i)-1|^{2p} \\
&+ 9B_3^2\frac{1}{N}\sum_{i=0}^{N-1}\mathbb{1}_{|r_{\theta_{c,e,0},\theta_{c,0,0}}(s_c^i,a_c^i)-1|>\epsilon}\,.
\end{aligned}
$$

对 $\mathcal{F}_c$ 取条件期望，我们应用引理 D.6 可得

$$
\begin{aligned}
&\mathbb{E}\Big[3\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \,\Big|\, \mathcal{F}_c\Big] \\
&\le 9B_1^2\mathbb{E}\big[|\theta_{c,e,0}-\theta_{c,0,0}|^2 \,\big|\, \mathcal{F}_c\big] + 9B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}E[|\theta_{c,e,0}-\theta_{c,0,0}|^{\frac{2p}{1-2p}} \mid \mathcal{F}_c]^{1-2p} \\
&\quad + 9B_3^2\frac{|\mathcal{A}|^q\,\Pi_*^q\,\mathbb{E}[|\theta_{c,e,0}-\theta_{c,0,0}|^{\frac{q}{1-q}} \mid \mathcal{F}_c]^{1-q}}{\epsilon^q}\,.
\end{aligned}
$$

现在，我们可以应用引理 D.7 得到

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \,\Big|\, \mathcal{F}_c\Big] \\
&\le 9\eta^2B_1^2K^2m^2G^2 + 9\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} + \frac{9\eta^qB_3^2|\mathcal{A}|^q\Pi_*^qK^qm^qG^q}{\epsilon^q}\,.
\end{aligned}
$$

□

对于 (18) 中的第二项，我们证明如下上界。

引理 D.10. 在　　假设 *A.2*、*A.3* 和 *A.4* 下，对于 $p, q \in (0,\ 1)$ 以及任意 (c, e) 则有　　at

$$
\begin{aligned}
&\mathbb{E}\Big[\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{PPO}^{(i),clip}(\theta_{c,e,0},\theta_{c,0,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\Big|^2 \,\Big|\, \mathcal{F}_c\Big] \\
&\le 12\eta^2B_1^2K^2m^2G^2 + 12\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} + \frac{12\eta^qB_3^2|\mathcal{A}|^q\Pi_*^qK^qm^qG^q}{\epsilon^q},
\end{aligned}
$$

其中常数由上述引理定义。

证明。 我们首先应用Jensen不等式

$$
\begin{aligned}
&\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\Big|^2 \\
&= \Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,k},\theta_{c,0,0})\Big|^2 \\
&\le 2\Big|\frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \\
&\quad + 2\Big|\frac{1}{m}\sum_{k=1}^{m-1}\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,k},\theta_{c,0,0}) - \frac{1}{N}\sum_{i=0}^{N-1} g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \\
&\le \frac{2}{N}\sum_{i=0}^{N-1}\Big|g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2 \\
&\quad + \frac{2}{m}\sum_{k=0}^{m-1}\frac{1}{B}\sum_{i\in\mathcal{B}_{c,e,k}}\Big|g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,k},\theta_{c,0,0}) - g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\Big|^2
\end{aligned}
$$

接下来，我们应用引理 D.3，其中 $\mathbb{A}$ 由估计 $\hat{}\ \mathbb{A}_c^i$ 给出；结合假设 A.2，对 $\mathcal{F}_c$ 取条件期望，并应用引理 D.6，得到

$$
\begin{aligned}
&\frac{2}{N}\sum_{i=0}^{N-1}\mathbb{E}\big[\big|g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,0},\theta_{c,0,0}) - g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\big|^2 \,\big|\, \mathcal{F}_c\big] \\
&\le 6B_1^2\mathbb{E}\big[\big|\theta_{c,e,0}-\theta_{c,0,0}\big|^2\,\big|\,\mathcal{F}_c\big] + 6B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}E\big[\big|\theta_{c,e,0}-\theta_{c,0,0}\big|^{\frac{2p}{1-2p}}\,\big|\,\mathcal{F}_c\big]^{1-2p} \\
&\quad + 6B_3^2\frac{|\mathcal{A}|^q\,\Pi_*^q\,\mathbb{E}\big[\big|\theta_{c,e,0}-\theta_{c,0,0}\big|^{\frac{q}{1-q}}\,\big|\,\mathcal{F}_c\big]^{1-q}}{\epsilon^q} \\
&\le 6\eta^2B_1^2K^2m^2G^2 + 6\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} + \frac{6\eta^qB_3^2|\mathcal{A}|^q\Pi_*^qK^qm^qG^q}{\epsilon^q}\,.
\end{aligned}
$$

类似地，我们可以应用引理 D.6 来界定

$$
\begin{aligned}
&\mathbb{E}\big[\frac{2}{m}\sum_{k=0}^{m-1}\sum_{i\in\mathcal{B}_{c,e,k}}\big|g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,e,k},\theta_{c,0,0}) - g_{\mathrm{PPO}}^{(i),\mathrm{clip}}(\theta_{c,0,0},\theta_{c,0,0})\big|^2\,\big|\,\mathcal{F}_c\big] \\
&\le 6B_1^2\frac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}\big[\big|\theta_{c,e,k}-\theta_{c,0,0}\big|^2\,\big|\,\mathcal{F}_c\big] + 6B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}\frac{1}{m}\sum_{k=0}^{m-1}\mathbb{E}\big[\big|\theta_{c,e,k}-\theta_{c,0,0}\big|^{\frac{2p}{1-2p}}\,\big|\,\mathcal{F}_c\big]^{1-2p} \\
&\quad + 6B_3^2\frac{1}{m}\sum_{k=0}^{m-1}\frac{|\mathcal{A}|^q\,\Pi_*^q\,\mathbb{E}[|\theta_{c,e,k}-\theta_{c,0,0}|^{\frac{q}{1-q}}\,|\,\mathcal{F}_c]^{1-q}}{\epsilon^q} \\
&\le 6\eta^2B_1^2K^2m^2G^2 + 6\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} + \frac{6\eta^qB_3^2|\mathcal{A}|^q\Pi_*^qK^qm^qG^q}{\epsilon^q}\,.
\end{aligned}
$$

□

现在我们可以陈述并证明主要结果：对于在训练轮次开始时均匀选取的参数，其 L^2-梯度范数具有何种性质。这样的选择看似任意，但它是学习过程中 L^2-梯度范数最小值的一个上界；在较弱假设下，这一最小值通常是在随机梯度下降中研究的。

定理 D.11。 假设满足假设 *A.1*、*A.2*、*A.3*、*A.4*，并且学习率 η 小于 $\frac{1}{Lm}$。从 $\{\theta_{c,e,0} \mid c = 0, \dots, C-1,\ e = 0, \dots, K-1\}$中均匀抽取样本 $\tilde{\theta}$。那么，对于任意$p, q \in (0,1)$，有

$$\begin{aligned}
&\min_{c=0,\dots,C-1,\ e=0,\dots,K-1} \mathbb{E}\big[|\nabla J(\theta_{c,e,0})|^2\big] \\
&\leq \mathbb{E}\big[|\nabla J(\tilde{\theta})|^2\big] \\
&\leq \frac{2\Delta_0}{\eta CKm} + 6\eta^2(7B_1^2 + L^2)K^2m^2G^2 + 42\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} \\
&\quad + \frac{42\eta^q B_3^2|\mathcal{A}|^q\Pi_*^q K^q m^q G^q}{\epsilon^q} + \frac{12\sigma^2}{N} + 12T^2\Pi_*^2\delta^2\,,
\end{aligned}$$

其中常数定义见上述引理，且 $\Delta_0 := J_* - J(\theta_{0,0,0})$。

证明。 由引理 D.8 可知，由于 $\eta \leq \frac{1}{Lm}$，有 ——

$$\mathbb{E}\big[J(\theta_{c,e+1,0})\big] \geq \mathbb{E}\big[J(\theta_{c,e,0})\big] + \frac{\eta m}{2}\mathbb{E}\big[\big|\nabla J(\theta_{c,e,0})\big|^2\big] - \frac{\eta m}{2}\,\mathbb{E}\Big[\Big|\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\Big|^2\Big].$$

由引理 D.9 和引理 D.10 可知

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\nabla J(\theta_{c,e,0}) - \frac{1}{m}\sum_{k=0}^{m-1}\hat{g}_{c,e,k}\Big|^2\Big] \\
&\leq 6\eta^2(7B_1^2 + L^2)K^2m^2G^2 + 42\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} \\
&\quad + \frac{42\eta^q B_3^2|\mathcal{A}|^q\Pi_*^q K^q m^q G^q}{\epsilon^q} + \frac{12\sigma^2}{N} + 12T^2\Pi_*^2\delta^2\,.
\end{aligned}$$

~我们对 $c = 0, \dots, C-1$ 和 $e = 0, \dots, K-1$ 取平均，使得对于从训练轮次 $\{\theta_{c,e,0} \mid c = 0, \dots, C-1,\ e = 0, \dots, K-1\}$ 开始处均匀抽取的 θ，有

$$\begin{aligned}
&\min_{c=0,\dots,C-1,\ e=0,\dots,K-1} \mathbb{E}\big[|\nabla J(\theta_{c,e,0})|^2\big] \\
&\leq \frac{1}{CK}\sum_{c=0}^{C-1}\sum_{e=0}^{K-1}\mathbb{E}\big[|\nabla J(\theta_{c,e,0})|^2\big] \\
&\leq \frac{2\Delta_0}{\eta CKm} + 6\eta^2(7B_1^2 + L^2)K^2m^2G^2 + 42\eta^{2p}B_2^2\epsilon^{2(1-p)}|\mathcal{A}|^{2p}\Pi_*^{2p}K^{2p}m^{2p}G^{2p} \\
&\quad + \frac{42\eta^q B_3^2|\mathcal{A}|^q\Pi_*^q K^q m^q G^q}{\epsilon^q} + \frac{12\sigma^2}{N} + 12T^2\Pi_*^2\delta^2\,,
\end{aligned}$$

其中我们将等式两边都除以 $\frac{\eta m}{2}$，并使用了望远镜求和

$$\sum_{c=0}^{C-1}\sum_{e=0}^{K-1}\big(\mathbb{E}[J(\theta_{c,e+1,0}] - \mathbb{E}[J(\theta_{c,e,0}]\big) = \mathbb{E}[J(\theta_{C,K,0}) - J(\theta_{0,0,0})] \leq J_* - \mathbb{E}[J(\theta_{0,0,0})] = \Delta_0.$$

□

E 有限时间 GAE

E.1 时序差分误差、 k 步优势估计器与标准广义优势估计

对于不熟悉 GAE（适用于无限时域 MDP）的读者，本节汇总了最重要的定义。为了在演员-评论家框架中构造优势函数的估计器，GAE 依赖于时序差分（TD）误差的概念 [29]。给定一个价值函数近似 V（通常来自价值网络），时刻 t 的单步时序差分误差定义为

$$\delta_t := R_t + \gamma V(S_{t+1}) - V(S_t).$$

如果价值函数近似就是真实价值函数，那么时序差分误差就是优势函数的无偏估计量：

$$\mathbb{E}[\delta_t \mid S_t, A_t] = \mathbb{E}[R_t + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) \mid S_t, A_t] = Q^\pi(S_t, A_t) - V^\pi(S_t) = \mathbb{A}^\pi(S_t, A_t), \quad (19)$$

这归因于马尔可夫性质和贝尔曼方程。利用时序差分误差 [29]，定义 k步优势估计器，在使用 V进行自举之前，累积来自 k 个未来步的信息：

$$\hat{\mathbb{A}}_t^{(k)} := \sum_{\ell=0}^{k} \gamma^\ell \delta_{t+\ell} = \sum_{\ell=0}^{k} \gamma^\ell R_{t+\ell} + \gamma^{k+1} V(S_{t+k+1}) - V(S_t) \quad (20)$$

第二个等式来自伸缩和消去。较大的 k 会使随机回报带来的方差更大，同时减少自举造成的价值函数近似偏差；当$k \approx \infty$ 时，对应蒙特卡洛优势近似。反之，较小的 k 对应较低的方差，但函数近似偏差更大。

广义优势估计器是所有 k步优势估计器的指数混合。使用几何权重（$1-\lambda$）λ^k，对于 $\lambda \in$ （0,1），原始广义优势估计器定义如下：

$$\hat{\mathbb{A}}_t^{\infty} := (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \, \hat{\mathbb{A}}_t^{(k)}. \quad (21)$$

前置因子（$1-\lambda$）对几何权重进行归一化，使得 $\sum_{k\geq 0}(1-\lambda)\lambda^k = 1$。因此，式 (21) 是 k 步估计器的凸组合，其中较长时域对应的权重按指数衰减。超参数 λ 是一个连续参数，可在高方差和高偏差状态之间进行插值。式 (21) 的混合形式还可以等价地简洁表示为时序差分误差的折扣和。具体而言，代入 $\hat{\mathbb{A}}_t^{(k)} = \sum_{\ell=0}^{k} \gamma^\ell \delta_{t+\ell}$ 并交换求和顺序可得

$$\hat{\mathbb{A}}_t^{\infty} = \sum_{\ell=0}^{\infty} (\gamma\lambda)^\ell \, \delta_{t+\ell}. \quad (22)$$

备注 E.1（索引约定与 [29]）。 原始 GAE 论文 [29]将 k步优势函数定义为在时间 $t+k$进行自举。相比之下，我们在定义（20）中于时间 $t+k+1$进行自举。等价地，按[29]中使用的索引方式，我们的 k步估计器对应于（k+1）步估计器。这纯粹是记号上的调整，目的是使几何混合呈现为 $\sum_{k\geq 0} \lambda^k \hat{\mathbb{A}}_t^{(k)}$。

E.2 广义优势估计的尾部质量塌缩

由（21）和（22）定义的序列与无限时域马尔可夫决策过程有着内在联系。它们隐含地依赖于这样一个事实：$\delta_{t+\ell}$，因此马尔可夫决策过程，对所有未来时刻都有定义。然而，在实践中，GAE 估计器序列用于有限时间马尔可夫决策过程场景，例如近端策略优化的实现。本节将指出一种有限时间场景下的副作用，我们称之为尾部质量塌缩。接下来的几节将讨论有限时间场景下可避免尾部质量塌缩的 GAE 替代方案。

假设 T是一个有限时间范围，并且 $\tilde{\tau}:=\inf\{\, t \geq 0 \mid S_t$为终端，$\}$是终止时间（例如在Lunar Lander中着陆）。我们用 $\tau = \tilde{\tau} \wedge T$ 表示终止与 T的最小值，即一个回合的有效结束。例如，在近端策略优化中，人们收集轨迹采样直至结束 τ，然后使用回溯递归来计算优势估计量。在没有进一步论证的情况下，近端策略优化在实践中采用（22），并在终止后取消时序差分误差：

$$\hat{\mathbb{A}}_t := \sum_{\ell=0}^{\tau-t-1} (\gamma\lambda)^\ell \, \delta_{t+\ell}, \quad t \leq \tau. \tag{23}$$

根据 [30], ，我们将该估计器称为截断广义优势估计。（23）的形式特别有用，因为它给出

$$\hat{\mathbb{A}}_t = \delta_t + \gamma\lambda \sum_{\ell=1}^{\tau-t-1} (\gamma\lambda)^{\ell-1} \, \delta_{t+\ell} = \delta_t + \gamma\lambda \sum_{\ell=0}^{\tau-t-2} (\gamma\lambda)^{\ell} \, \delta_{t+1+\ell} = \delta_t + \gamma\lambda \hat{\mathbb{A}}_{t+1},$$

从而得到一种沿时间反向的迭代计算方案。对于收集到的直至 τ的轨迹展开，$\hat{}\,\hat{}\,\hat{}$可以直接使用终端条件 $\mathbb{A}_\tau^\infty:=0$回溯 $\mathbb{A}_t = \delta_t + \gamma\lambda\mathbb{A}_{t+1}$。
现在我们来讨论广义优势估计序列的有限时间截断所导致的尾部质量塌缩。这种缩放并不符合最初的目的。几何权重原本分布在 $\mathbb{N}$上，现在则被限制在 $\{0, ..., \tau\}$。 τ之后的全部权重都塌缩到可用的最长 k步估计器上，即最接近蒙特卡洛的那个。由此可见，将截断广义优势估计序列直接应用于有限长度的轨迹采样，会比最初预期具有更大的方差/更小的偏差。下面是一个形式化命题。

命题 E.2 （广义优势估计的尾部质量塌缩）。 固定 $t \in \{0, \ldots, \tau-1\}$，并假设广义优势估计$\hat{}$估计器序列 $\mathbb{A}_t^\infty$由(23)给出。那么，

$$\hat{\mathbb{A}}_t = \sum_{k=0}^{\tau-t-2} \underbrace{(1-\lambda)\lambda^k}_{GAE\ weights} \hat{\mathbb{A}}_t^{(k)} \; + \underbrace{\lambda^{\tau-t-1}}_{tail\text{-}mass\ collapse\ weight} \hat{\mathbb{A}}_t^{(\tau-t-1)},$$

约定空和等于零。

由于无穷多个权重塌缩为一个，我们将广义优势估计应用于有限时间设置时的这一特性称为广义优势估计尾部质量塌缩。正文中的图 2 展示了四种 t 取值下不同 k 步估计器的权重。大型蓝色原子体现了尾部质量塌缩。

证明。 固定 $t \in \{0, \ldots, \tau-1\}$。设 $\tilde{\mathbb{A}}_t := (1-\lambda) \sum_{k=0}^{\tau-t-2} \lambda^k \hat{}\,\mathbb{A}_t^{(k)} + \lambda^{\tau-t-1} \hat{}\,\mathbb{A}_t^{(\tau-t-1)}$。我们证明，$\tilde{\mathbb{A}}_t$ 的恒等式恰好得出 (23)。由于 $\hat{}\,\mathbb{A}_t^{(k)} = \sum_{\ell=0}^{k} \gamma^\ell \delta_{t+\ell}$，我们得到

$$\tilde{\mathbb{A}}_t = (1-\lambda) \sum_{k=0}^{\tau-t-2} \lambda^k \sum_{\ell=0}^{k} \gamma^\ell \delta_{t+\ell} \; + \; \lambda^{\tau-t-1} \sum_{\ell=0}^{\tau-t-1} \gamma^\ell \delta_{t+\ell}.$$

对于第一项，我们交换求和顺序，并使用以下标准公式

$$(1-\lambda)\sum_{k=0}^{\tau-t-2}\lambda^k\sum_{\ell=0}^{k}\gamma^\ell = (1-\lambda)\sum_{k=0}^{\tau-t-2}\lambda^k\sum_{\ell=0}^{\tau-t-2}\mathbb{1}_{\{\ell\le k\}}\gamma^\ell\delta_{t+\ell}\gamma^\ell\delta_{t+\ell} = (1-\lambda)\sum_{\ell=0}^{\tau-t-2}\gamma^\ell\delta_{t+\ell}\sum_{k=\ell}^{\tau-t-2}\lambda^k$$
$$= (1-\lambda)\sum_{\ell=0}^{\tau-t-2}\gamma^\ell\delta_{t+\ell}\,\lambda^\ell\sum_{j=0}^{\tau-t-\ell-2}\lambda^j = \sum_{\ell=0}^{\tau-t-2}\gamma^\ell\delta_{t+\ell}\,\lambda^\ell\big(1-\lambda^{\tau-t-\ell-1}\big)$$
$$= \sum_{\ell=0}^{\tau-t-2}(\gamma\lambda)^\ell\delta_{t+\ell} \;-\; \lambda^{\tau-t-1}\sum_{\ell=0}^{\tau-t-2}\gamma^\ell\delta_{t+\ell}.$$

˜将其代回 $\mathbb{A}_t$ 的公式，可得

$$\tilde{\mathbb{A}}_t = \sum_{\ell=0}^{\tau-t-2}(\gamma\lambda)^\ell\delta_{t+\ell} - \lambda^{\tau-t-1}\sum_{\ell=0}^{\tau-t-2}\gamma^\ell\delta_{t+\ell} + \lambda^{\tau-t-1}\sum_{\ell=0}^{\tau-t-1}\gamma^\ell\delta_{t+\ell}$$
$$= \sum_{\ell=0}^{\tau-t-2}(\gamma\lambda)^\ell\delta_{t+\ell} + \lambda^{\tau-t-1}\gamma^{\tau-t-1}\delta_{t+\tau-t-1} = \sum_{\ell=0}^{\tau-t-1}(\gamma\lambda)^\ell\delta_{t+\ell}.$$

这恰好就是式（23）。 □

因此，在有限时域（或会终止的）马尔可夫决策过程中使用标准广义优势估计的混合，会导致权重明显坍缩到最终非平凡估计量上。与此同时，广义优势估计最初的动机是，对 k 步估计量 [29] 进行类似 TD(λ) 的几何加权平均。

为缓解尾部质量坍缩，我们建议在归一化几何权重时考虑轨迹展开长度。我们不使用 $(1-\lambda)$ 进行归一化，而是自适应地使用 $\frac{1-\lambda}{1-\lambda^{T-t}}$ 或 $\frac{1-\lambda}{1-\lambda^{\tau-t}}$，从而使每个加和项都具有所需的指数权重。由此得到的反向归纳过程与广义优势估计相同，只是缩放因子不同。

E.3 固定时间 GAE

我们首先考察在轨迹时域 T 处进行确定性截断的影响。即使不存在提前终止，标准广义优势估计也会隐式地在无限范围的 k 上混合 k 步估计量，而在时间 t 之后收集的数据仅支持 $k \le T-t-1$ 的估计量。因此，一种自然的有限时间类比形式是将几何混合限制在可用范围内，并重新归一化权重，使其总和为一。

定义 E.3（固定时间 GAE）。取定 $\lambda \in (0,\ 1)$ 和时域 $T \in \mathbb{N}$。对于任意 $t \in \{0,\dots,T-1\}$，固定时间广义优势估计器定义为

$$\hat{\mathbb{A}}_t^T := \frac{1-\lambda}{1-\lambda^{T-t}}\sum_{k=0}^{T-t-1}\lambda^k\,\hat{\mathbb{A}}_t^{(k)}. \tag{24}$$

归一化因子 $\frac{1-\lambda}{1-\lambda^{T-t}}$ 确保几何权重之和为1，从而使ˆ $\mathbb{A}_t^T$ 成为一般可观测的 k步估计器的凸组合。该表述给出了与截断轨迹可用数据相一致的广义优势估计的固定时域类比形式。与广义优势估计类似，该估计器具有紧凑的TD求和表示，并得到可用于实际实现的递归公式。

命题 E.4（固定时间估计器的后向递归）。 对于 $t = T-1$，.....，0，我们有

$$\hat{\mathbb{A}}_t^T = \sum_{\ell=0}^{T-t-1} (\gamma\lambda)^\ell \frac{1-\lambda^{T-t-\ell}}{1-\lambda^{T-t}} \delta_{t+\ell}.$$

此外，如果我们设定 $\hat{\mathbb{A}}_T^T = 0$ ，该估计器满足以下递归公式:

$$\hat{\mathbb{A}}_t^T = \delta_t + \gamma\lambda \underbrace{\frac{1-\lambda^{T-t-1}}{1-\lambda^{T-t}}}_{new\ additional\ factor} \hat{\mathbb{A}}_{t+1}^T, \qquad t = T-1, \ldots, 0.$$

证明。固定 $t \in \{0, \ldots, T-1\}$.

$$\begin{aligned}
\hat{A}_t^T :&= \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{T-t-1} \lambda^k \hat{A}_t^{(k)} = \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{T-t-1} \lambda^k \sum_{\ell=0}^{T-t-1} \mathbb{1}_{\{\ell \le k\}} \gamma^\ell \delta_{t+\ell} \\
&= \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{\ell=0}^{T-t-1} \gamma^\ell \delta_{t+\ell} \sum_{k=\ell}^{T-t-1} \lambda^k = \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{\ell=0}^{T-t-1} \gamma^\ell \delta_{t+\ell}\, \lambda^l \sum_{k=0}^{T-t-l-1} \lambda^k \\
&= \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{\ell=0}^{T-t-1} \gamma^\ell \delta_{t+\ell}\, \lambda^l \frac{1-\lambda^{T-t-l}}{1-\lambda} = \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{\ell=0}^{T-t-1} \gamma^\ell \delta_{t+\ell} \frac{\lambda^l - \lambda^{T-t}}{1-\lambda} \\
&= \sum_{\ell=0}^{T-t-1} \frac{\lambda^\ell - \lambda^{T-t}}{1-\lambda^{T-t}} \gamma^\ell \delta_{t+\ell} = \sum_{\ell=0}^{T-t-1} \frac{1-\lambda^{T-t-\ell}}{1-\lambda^{T-t}} \delta_{t+\ell} (\gamma\lambda)^\ell.
\end{aligned}$$

因此，我们可以利用上述方程，通过对 t进行归纳法来得到递归公式。基础情形很容易得出，因为 $\hat{\mathbb{A}}_T^T = 0$；对于 $t = T-1$，，我们有 $\hat{\mathbb{A}}_{T-1}^T = \delta_{T-1}$。对于一般的 $t \in \{0, \ldots, T-2\}$，上述公式给出

$$\hat{\mathbb{A}}_t^T = \delta_t + \sum_{\ell=1}^{T-t-1} \frac{\lambda^\ell - \lambda^{T-t}}{1-\lambda^{T-t}} \gamma^\ell \delta_{t+\ell} = \delta_t + \sum_{\ell=0}^{T-t-2} \frac{\lambda^{\ell+1} - \lambda^{T-t}}{1-\lambda^{T-t}} \gamma^{\ell+1} \delta_{t+1+\ell} = \delta_t + \frac{1-\lambda^{T-t-1}}{1-\lambda^{T-t}} \lambda\gamma \hat{\mathbb{A}}_{t+1}^T \square$$

与无限时域广义优势估计类似，在使用策略 π 的真实价值函数 V_t^π来计算时序差分误差的理想化情形下，估计器 $\hat{\mathbb{A}}_t^T$对于时变优势 $\mathbb{A}_t^\pi(S_t, A_t)$ 仍然是无偏的。

命题E.5。 假设在时序差分误差中使用 π 的真实价值函数 V_t^π:

$$\delta_t := R_t + \gamma V_t^\pi(S_{t+1}) - V_t^\pi(S_t), \qquad \gamma \in (0, 1).$$

那么，对于任意 $t \in \{0, \ldots T-1\}$:

$$\mathbb{E}\Big[\hat{\mathbb{A}}_t^T \mid S_t, A_t\Big] = \mathbb{A}_t^\pi(S_t, A_t)$$

证明。 对于固定的起始时刻 t，回顾一下，固定时间估计器定义为 k 步估计器的几何加权平均。对于任意 $k \le T-t-1$，根据贝尔曼方程，我们有

$$\mathbb{E}\left[\hat{\mathbb{A}}_t^{(k)} \mid S_t, A_t\right] = \mathbb{E}\Big[\sum_{\ell=0}^{k} \gamma^\ell R_{t+\ell} + \gamma^{k+1} V_t^\pi(S_{t+k+1}) - V_t^\pi(S_t) \mid S_t, A_t\Big] = \mathbb{A}_t^\pi(S_t, A_t).$$

由于固定时间估计器是 k 步估计器的归一化几何加权平均，利用上述恒等式可得

$$\mathbb{E}[\hat{\mathbb{A}}_t^T \mid S_t, A_t] = \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{T-t-1} \lambda^k \, \mathbb{E}[\hat{\mathbb{A}}_t^{(k)} \mid S_t, A_t] = \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{T-t-1} \lambda^k \, \mathbb{A}_t^\pi(S_t, A_t) = \mathbb{A}_t^\pi(S_t, A_t),$$

因为几何权重之和为 1。 □

到目前为止，我们将固定时间估计器 $\hat{}\,\mathbb{A}_t^T$ 引入为一种合理的方法，用于处理轨迹时域 T 处的截断。现在，我们分析当回合在 T 之前终止时会发生什么。在这种情况下，我们有 $\tilde{\tau} < T$，因此 $\tau < T$。采用通常的终止状态约定，这意味着所有超出终止时间的 k 步估计器都与最后一个非平凡估计器一致，因此固定时间几何混合会将剩余的尾部质量隐式地重新分配给该最终估计。下面的命题将这种权重坍缩精确表述出来。

命题 E.6（固定时间 GAE 的权重坍缩）。 假设对于所有 $\tau \le u \le T$，都有

$$S_u = S_{\tilde{\tau}}, \qquad R_u = 0 \quad \textit{and} \quad V(S_{\tilde{\tau}}) = 0.$$

那么，固定时间 *GAE* 估计器 (24) 可分解为

$$\hat{\mathbb{A}}_t^T = \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{\tau-t-2} \lambda^k \hat{\mathbb{A}}_t^{(k)} \; + \; \frac{\lambda^{\tau-t-1} - \lambda^{T-t}}{1-\lambda^{T-t}} \, \hat{\mathbb{A}}_t^{(\tau-t-1)} \tag{25}$$

对于任意 $t \in \{0, \dots, \tau-1\}$。

证明。 固定 $t \in \{0, \dots, \tau-1\}$。根据假设，对于任意 $u \ge \tau$，我们有 $R_u = 0$，且 $S_u = S_\tau$，并且 $V_u(S_\tau) = 0$。因此，对于所有 $u \ge \tau$，

$$\delta_u = R_u + \gamma V_{u+1}(S_{u+1}) - V_u(S_u) = 0 + \gamma \cdot 0 - 0 = 0.$$

因此，对于任意 $\tau - t \le k \le T - t$，

$$\hat{\mathbb{A}}_t^{(k)} = \sum_{\ell=0}^{k} \gamma^\ell \delta_{t+\ell} = \sum_{\ell=0}^{\tau-t-1} \gamma^\ell \delta_{t+\ell} = \hat{\mathbb{A}}_t^{(\tau-t-1)},$$

这再次表明，所有需要超出 τ 范围信息的 k 步优势函数估计器，都与最后一个非平凡估计器 $\hat{}\,\mathbb{A}_t^{(\tau-t-1)}$. 相同。因此，在最后一个可观测索引 $\tau - t - 1$ 处拆分几何和，可得

$$\begin{aligned} \sum_{k=0}^{T-t-1} \lambda^k \hat{\mathbb{A}}_t^{(k)} &= \sum_{k=0}^{\tau-t-1} \lambda^k \hat{\mathbb{A}}_t^{(k)} \; + \sum_{k=\tau-t}^{T-t-1} \lambda^k \hat{\mathbb{A}}_t^{(k)} \\ &= \sum_{k=0}^{\tau-t-1} \lambda^k \hat{\mathbb{A}}_t^{(k)} \; + \; \hat{\mathbb{A}}_t^{(\tau-t-1)} \sum_{k=\tau-t}^{T-t-1} \lambda^k. \end{aligned}$$

尾和是一个几何级数：

$$\sum_{k=\tau-t}^{T-t-1} \lambda^k = \lambda^{\tau-t} \sum_{j=0}^{T-\tau-1} \lambda^j = \lambda^{\tau-t} \frac{1-\lambda^{T-\tau}}{1-\lambda}.$$

乘以前置因子 $(1-\lambda)/(1-\lambda^{T-t})$ 得到

$$\begin{aligned}
\hat{\mathbb{A}}_t^T &= \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{\tau-t-1} \lambda^k \hat{\mathbb{A}}_t^{(k)} \;+\; \frac{1-\lambda}{1-\lambda^{T-t}} \lambda^{\tau-t} \frac{1-\lambda^{T-\tau}}{1-\lambda} \hat{\mathbb{A}}_t^{(\tau-t-1)} \\
&= \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{\tau-t-1} \lambda^k \hat{\mathbb{A}}_t^{(k)} \;+\; \frac{\lambda^{\tau-t}-\lambda^{T-t}}{1-\lambda^{T-t}} \hat{\mathbb{A}}_t^{(\tau-t-1)} \\
&= \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{\tau-t-2} \lambda^k \hat{\mathbb{A}}_t^{(k)} \;+\; \Big(\frac{1-\lambda}{1-\lambda^{T-t}} \lambda^{\tau-t-1} + \frac{\lambda^{\tau-t}-\lambda^{T-t}}{1-\lambda^{T-t}} \Big) \hat{\mathbb{A}}_t^{(\tau-t-1)} \\
&= \frac{1-\lambda}{1-\lambda^{T-t}} \sum_{k=0}^{\tau-t-2} \lambda^k \hat{\mathbb{A}}_t^{(k)} \;+\; \frac{\lambda^{\tau-t-1}-\lambda^{T-t}}{1-\lambda^{T-t}} \hat{\mathbb{A}}_t^{(\tau-t-1)},
\end{aligned}$$

这正是式（25）。 □

如果在一个回合结束之前发生终止，即 $\tau < T$，式（25）表明，固定时间估计器不再是对真正不同的 k 步估计器进行纯粹的几何加权平均。相反，原本分配给不可观测索引 $\tau - t \le k \le T - t$ 的几何尾部权重，实际上会重新分配给最后一个非平凡估计 $\hat{\mathbb{A}}_t^{(\tau-t-1)}$，从而再次导致该项出现类似的权重坍缩效应，不过其程度弱于采用标准估计量时的情况。终止越早（即 $\tau - t$ 越小）、λ 越大，相应的尾部系数就越大，因此 $\hat{\mathbb{A}}_t^T$ 会越集中于 $\hat{\mathbb{A}}_t^{(\tau-t-1)}$，而不是将权重分配到 k 的观测范围内。

不过，这促使我们采用一种终止自适应变体：在有效终点 τ 截断几何混合并进行相应的归一化，使估计器仅取决于截至时间 $\tau - 1$ 所观测到的奖励和时序差分误差。

E.4 终止时间 GAE

如上所述，我们将几何加权平均限制在终止前实际可用的步数范围内。这会产生一个依赖于随机时域的估计器，该时域由回合的终止时间给出。对于任意 $t \in \{0, \dots, \tau - 1\}$，只有带 $k \le \tau - t - 1$ 的 k 步估计器才能被已观测的轨迹片段完全支撑。因此，我们定义如下重新归一化的几何混合。

定义 E.7 (终止时间 GAE). 对于任意 $t \in \{0, \dots, \tau - 1\}$，终止时间 GAE 估计器定义为

$$\hat{\mathbb{A}}_t^\tau := \frac{1-\lambda}{1-\lambda^{\tau-t}} \sum_{k=0}^{\tau-t-1} \lambda^k \, \hat{\mathbb{A}}_t^{(k)}. \tag{26}$$

$\hat{}$ 根据构造，$\mathbb{A}_t^\tau$ 仅使用截至有效终点 τ 的信息。它仅取决于奖励 $\{R_t, \dots, R_{\tau-1}\}$ 以及沿状态 $\{S_t, \dots, S_\tau\}$ 的价值函数评估。当 $\tau = T$ 时，该估计器与固定时间估计器 $\hat{\mathbb{A}}_t^T$ 一致。当 $\tau < T$ 时，它会自动适应较短的可用轨迹长度，并避免索引 $k \ge \tau - t$ 到 $k = \tau - t - 1$ 发生质量坍缩。

命题 E.8 (终止时间 GAE 的反向递归**)**。 对于任意 $t \in \{0, \dots, \tau - 1\}$，终止时间估计器都具有 *TD* 求和表示

$$\hat{\mathbb{A}}_t^\tau = \sum_{\ell=0}^{\tau-t-1} (\gamma\lambda)^\ell \, \frac{1-\lambda^{\tau-t-\ell}}{1-\lambda^{\tau-t}} \, \delta_{t+\ell}. \tag{27}$$

此外，若设定 $\mathbb{A}_\tau^\tau = 0$，则 $\mathbb{A}_t^\tau$ 满足以下后向递归

$$\hat{\mathbb{A}}_t^\tau = \delta_t + \gamma\lambda \, \frac{1-\lambda^{\tau-t-1}}{1-\lambda^{\tau-t}} \, \hat{\mathbb{A}}_{t+1}^\tau, \qquad t = \tau - 1, \dots, 0. \tag{28}$$

证明。 将确定性时域 T 替换为（随机）有效时域 τ，并沿每条路径进行推导，证明方式与命题 E.4 的证明类似。 □

算法 2 给出了终止时间 GAE 的伪代码。

E.5 估计器之间的关系与偏差-方差权衡

现在，我们利用命题 E.2 和 E.6 来说明这三种估计器之间的关系，然后讨论一些关于它们的偏差-方差权衡的启发式方法。

命题 E.9 (标准、固定时刻方法与终止时间 GAE 之间的关系**).** 在与 *E.2* 和 *E.6* 相同的假设下，对于 $t \in \{0, \dots, \tau - 1\}$，以下等式成立：

$$\hat{\mathbb{A}}_t = (1-\lambda^{\tau-t}) \, \hat{\mathbb{A}}_t^\tau \; + \; \lambda^{\tau-t} \, \hat{\mathbb{A}}_t^{(\tau-t-1)}, \tag{29}$$

$$\hat{\mathbb{A}}_t^T = \frac{1-\lambda^{\tau-t}}{1-\lambda^{T-t}} \, \hat{\mathbb{A}}_t^\tau \; + \; \frac{\lambda^{\tau-t} - \lambda^{T-t}}{1-\lambda^{T-t}} \, \hat{\mathbb{A}}_t^{(\tau-t-1)}. \tag{30}$$

特别地，对于 $\lambda \in (0,\ 1)$ 和 $\tau \le T$，$\hat{\mathbb{A}}_t$ 与 $\hat{\mathbb{A}}_t^T$ 都是 $\hat{\mathbb{A}}_t^\tau$ 和最后一个非平凡 k 步估计器 $\hat{\mathbb{A}}_t^{(\tau-t-1)}$ 的凸组合。

证明。 我们从标准估计量 $\mathbb{A}_t$ 与终止时间估计器 $\hat{\mathbb{A}}_t^\tau$ 之间的关系入手。根据命题 7.1，我们有

$$\hat{\mathbb{A}}_t = (1-\lambda) \sum_{k=0}^{\tau-t-2} \lambda^k \, \hat{\mathbb{A}}_t^{(k)} \; + \; \lambda^{\tau-t-1} \, \hat{\mathbb{A}}_t^{(\tau-t-1)}. \tag{31}$$

By 根据 $\mathbb{A}_t^\tau$ 的定义，我们可以将直到 $\tau - t - 2$ 的部分几何级数改写为 a s

$$\hat{\mathbb{A}}_t^\tau = \frac{1-\lambda}{1-\lambda^{\tau-t}} \left(\sum_{k=0}^{\tau-t-2} \lambda^k \, \hat{\mathbb{A}}_t^{(k)} + \lambda^{\tau-t-1} \hat{\mathbb{A}}_t^{(\tau-t-1)} \right),$$

因此

$$(1-\lambda) \sum_{k=0}^{\tau-t-2} \lambda^k \, \hat{\mathbb{A}}_t^{(k)} = (1-\lambda^{\tau-t}) \, \hat{\mathbb{A}}_t^\tau \; - \; (1-\lambda)\lambda^{\tau-t-1} \, \hat{\mathbb{A}}_t^{(\tau-t-1)}. \tag{32}$$

将（32）代入（31）可得

$$\begin{aligned}\hat{\mathbb{A}}_t &= \Big((1-\lambda^{\tau-t})\,\hat{\mathbb{A}}_t^{\tau} - (1-\lambda)\lambda^{\tau-t-1}\hat{\mathbb{A}}_t^{(\tau-t-1)}\Big) \;+\; \lambda^{\tau-t-1}\hat{\mathbb{A}}_t^{(\tau-t-1)} \\ &= (1-\lambda^{\tau-t})\,\hat{\mathbb{A}}_t^{\tau} \;+\; \lambda^{\tau-t-1}\big(1-(1-\lambda)\big)\hat{\mathbb{A}}_t^{(\tau-t-1)} \\ &= (1-\lambda^{\tau-t})\,\hat{\mathbb{A}}_t^{\tau} \;+\; \lambda^{\tau-t}\hat{A}_t^{(\tau-t-1)},\end{aligned}$$

这就证明了（29）。类似地，对于固定时间估计器 $\hat{}A_t^T$与终止时刻估计器 $\hat{}\ \mathbb{A}_t^{\tau}$之间的关系，根据命题 E.6 可得

$$\hat{\mathbb{A}}_t^T = \frac{1-\lambda}{1-\lambda^{T-t}}\sum_{k=0}^{\tau-t-2}\lambda^k\,\hat{\mathbb{A}}_t^{(k)} \;+\; \frac{\lambda^{\tau-t-1}-\lambda^{T-t}}{1-\lambda^{T-t}}\,\hat{\mathbb{A}}_t^{(\tau-t-1)}.$$

将（32）乘以（$1-\lambda^{T-t}$）$^{-1}$并代入，可得

$$\begin{aligned}\hat{\mathbb{A}}_t^T &= \frac{1}{1-\lambda^{T-t}}\Big((1-\lambda^{\tau-t})\,\hat{\mathbb{A}}_t^{\tau} - (1-\lambda)\lambda^{\tau-t-1}\hat{\mathbb{A}}_t^{(\tau-t-1)}\Big) \;+\; \frac{\lambda^{\tau-t-1}-\lambda^{T-t}}{1-\lambda^{T-t}}\,\hat{\mathbb{A}}_t^{(\tau-t-1)} \\ &= \frac{1-\lambda^{\tau-t}}{1-\lambda^{T-t}}\,\hat{\mathbb{A}}_t^{\tau} \;+\; \frac{-(1-\lambda)\lambda^{\tau-t-1}+\lambda^{\tau-t-1}-\lambda^{T-t}}{1-\lambda^{T-t}}\,\hat{\mathbb{A}}_t^{(\tau-t-1)} \\ &= \frac{1-\lambda^{\tau-t}}{1-\lambda^{T-t}}\,\hat{\mathbb{A}}_t^{\tau} \;+\; \frac{\lambda^{\tau-t}-\lambda^{T-t}}{1-\lambda^{T-t}}\,\hat{\mathbb{A}}_t^{(\tau-t-1)}.\end{aligned}$$

□

命题 E.9 明确指出，一旦发生提前终止 $\{\tau < T\}$，$\hat{}\ \mathbb{A}_t$ 和$\hat{}\ \mathbb{A}_t^T$都可以被视为终止时间混合估计量 $\mathbb{A}_t^{\tau}$的重新加权扩展。（29）和（30）中的第二个分量始终是最大的非平凡 k 步估计器$\hat{}\mathbb{A}_t^{(\tau-t-1)}$，即在自举之前使用最长可用前瞻的估计器。该项通常表现出最小的$\hat{}$自举偏差（因为它对 V 的依赖最少），但同时具有最大的方差，因为它累加了最长的（折扣）时序差分误差之和。

终止时间混合估计器 $\mathbb{A}_t^{\tau}$避免将观测范围之外的尾部赋予任何额外权重：它只对数据在有效终点 τ之前所支持的、真正不同的 k步估计器进行平均。因此，从方差角度来看，这是最保守的选择，并且可以预期，在保持γ, λ固定的情况下，它在三者中的方差最小。在事件 $\{\tau = T\}$（轨迹展开期间未发生终止）发生时，终止时间估计器与固定时间估计器按定义一致，$\hat{}\ \mathbb{A}_t^{\tau} = \hat{}\mathbb{A}_t^T$。

当 $\tau < T$时，固定时间估计器 $\hat{}\mathbb{A}_t^T$仍通过式(30)中的系数 $\frac{\lambda^{\tau-t}-\lambda^{T-t}}{1-\lambda^{T-t}}$，将额外的几何权重分配给最后一个非平凡项。与 $\hat{}\ \mathbb{A}_t^{\tau}$相比，这加大了对 $\hat{}\mathbb{A}_t^{(\tau-t-1)}$ 的重视程度，直观上会降低偏差，但增加方差。标准估计量 $\hat{}\ \mathbb{A}_t$ 最能体现这一效应：它在式 (29) 中将全部尾部质量 $\lambda^{\tau-t}$ 分配给 $\hat{}\mathbb{A}_t^{(\tau-t-1)}$，因此预期自举偏差最小，但方差最大。

最后，当 $\tau - t$ 较小时，各估计器之间的差异最为显著，即时间 t 之后可用的有效轨迹后缀较短时（要么是因为回合非常短，要么是因为 t 接近 τ）。在这一情形下，即使 λ 取中等数值，也会产生相当大的相对尾部权重，而(29)-(30)中的凸组合可能显著不同。我们在E.9小节中，在关于时序差分误差的简化假设下进一步明确量化了这一效应。

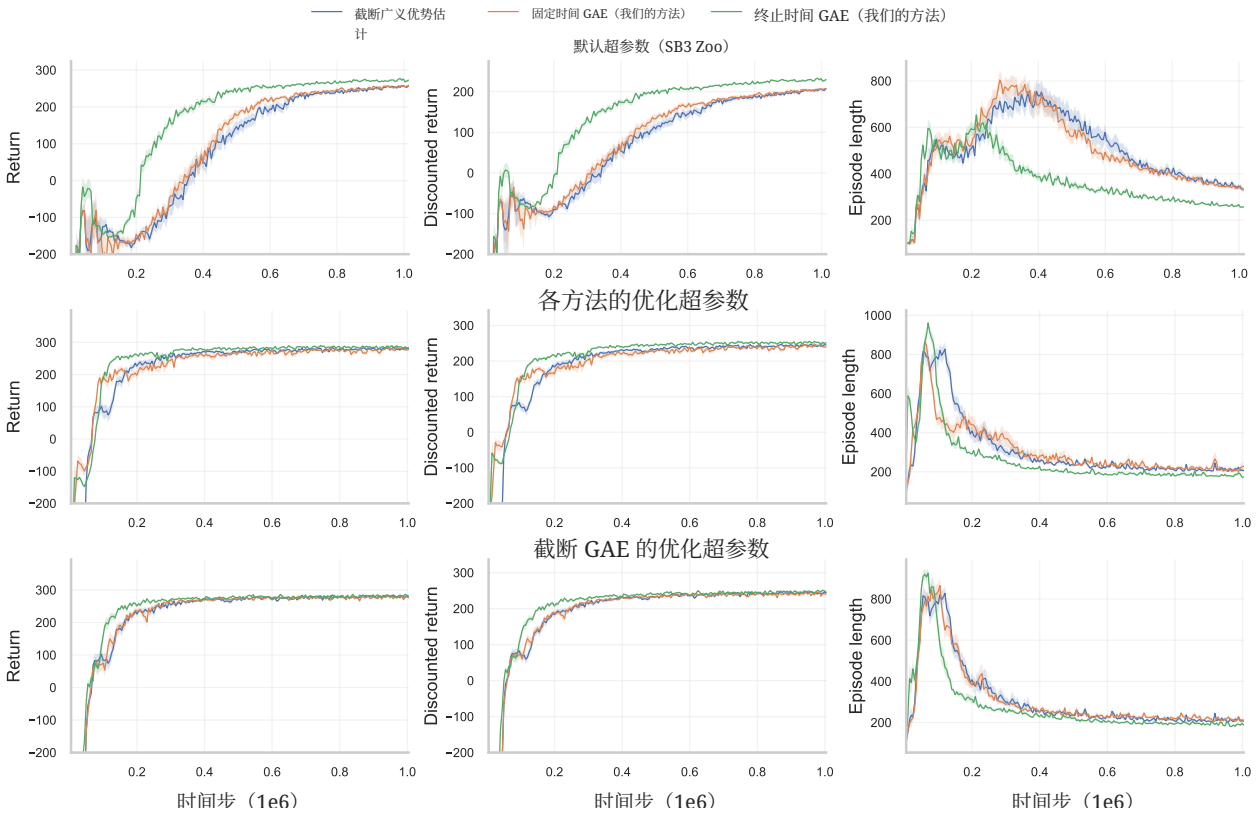


图 5：三种超参数（HP）设置下 LunarLander-v3 的评估学习曲线。各列分别报告评估回报（奖励总和）、折扣评估回报（折扣奖励总和）和评估回合长度。曲线为 20 个随机种子的平均值，阴影区域表示随机种子的标准误。各行对应不同的超参数设置：上：Stable-Baselines3 Zoo 中的默认 PPO 超参数。中：通过超参数优化（100 次试验，每次试验使用 3 个随机种子）分别为每种方法找到的最佳超参数。下：所有方法均使用截断广义优势估计的超参数优化所得的超参数进行评估（所有方法使用相同的超参数）。

E.6 实现说明

尽管本文重点关注近端策略优化的数学基础，我们仍进行了一些实验，以凸显清晰表述的实用性，尤其是无限时间GAE估计量的有限时间用法。我们使用Stable-Baselines3 [24]在LunarLander-v3上进行了实验。

我们的理论定义将 $(S_t, A_t, R_t)_{t\geq 0}$ 视为一个随机过程，并将优势估计量定义为随机变量。然而，在实现中，我们只能获得该过程的有限实现，即在当前策略下采样得到的一组有序转移。因此，轨迹缓冲区存储一组有限的有序实现，我们用单一的全局索引$n \in \{0, \dots, N-1\}$对其进行索引，即使缓冲区包含多个回合：

$$\mathcal{B} = \{(s_n, a_n, r_n, v_n, t_n, d_n)_{n=0}^{N-1}\},$$

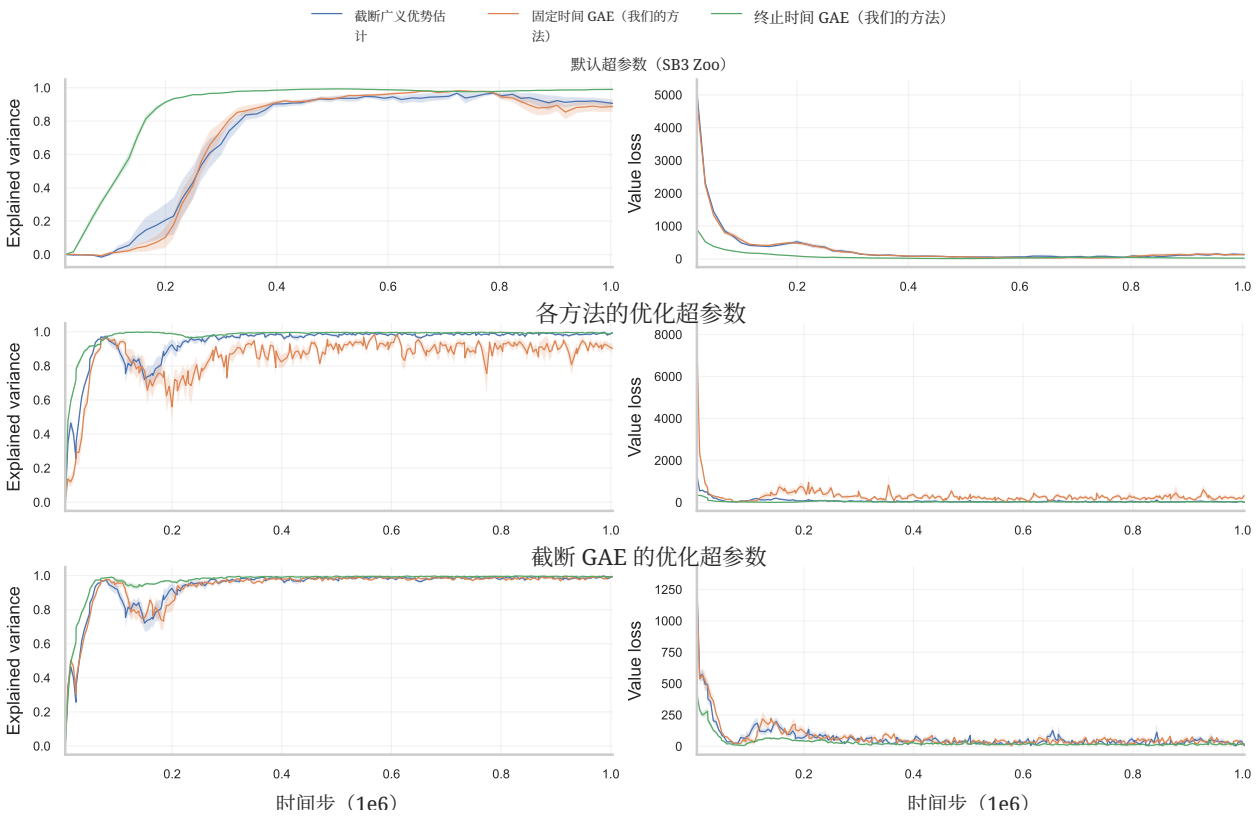


图 6：使用截断 GAE 和我们修改后的 GAE 变体进行 PPO 训练时，LunarLander-v3 上的训练诊断。各行对应不同的超参数（HP）设置：上：Stable-Baselines3 Zoo 中的默认 PPO 超参数。中：通过超参数优化（100 次试验，每次试验使用 3 个随机种子）为每种方法分别找到的最佳超参数。下：所有方法均使用通过截断 GAE 超参数优化得到的超参数进行评估（所有方法的超参数相同）。各列显示价值函数拟合的解释方差（左）和价值损失（右）。曲线为 20 个随机种子的平均值，阴影区域表示随机种子的标准误。

其中 n 是全局缓冲区索引（跨越多次轨迹采样）。这里 $s_n \in \mathcal{S}$ 表示状态， $a_n \in \mathcal{A}$ˆ表示动作， r_n 表示观测奖励， $v_n = V\left(s_n\right)$ 表示所存储的与时间无关的价值估计，而$t_n \in \{0, \ldots, T\}$ 表示状态转移 n的回合内时间戳。此外， $d_n \in \{0,1\}$ 是结束掩码，用于指示下一个缓冲区条目是否属于同一回合，即若状态转移 $n+1$ 延续状态转移 n 的回合，则为 $d_n = 1$ ；若在 n 与 n+1 之间出现回合边界（要么因为回合在 n终止，要么因为轨迹展开被截断且新回合从 n+1开始），则为 $d_n = 0$ 。

因此，除了标准 PPO 缓冲区外，唯一还需要的信息是每个状态转移对应的回合内时间戳 t_n 。这是必要的，因为命题 E.4 和 E.8 中的递归权重取决于固定时间估计器所对应的剩余距离 T，以及终止时间估计器所对应的剩余距离 τ 。

在此基础上　　在 s 个命题下，两种估计器都可以通过一次反向扫描计算得出

over 缓冲区， $n = N-1$、.....、0。定义缓冲区上的单步时序差分残差 by

$$\delta_n := r_n + \gamma\, d_n\, v_{n+1} - v_n,$$

其中， v_{n+1} 被视为下一个状态的自举值。然后，在算法上，我们从后向前遍历缓冲区， $n = N-1$、.....、0，并为当前状态转移所属的回合ˆ片段维护一个有效结束时刻 τ_{eff}。接着，为了估计优势 $\mathbb{A}_n$ ，我们使用以下更新方案:

$$\hat{\mathbb{A}}_n = \delta_n + \gamma\lambda\, \frac{1-\lambda^{\tau_{\text{eff}}-t_n-1}}{1-\lambda^{\tau_{\text{eff}}-t_n}}\, d_n\, \hat{\mathbb{A}}_{n+1}. \tag{33}$$

在固定时间情形下， $\tau_{\text{eff}} \equiv T$ 是常数。在终止时间情形下，每当反向扫描跨越回合边界时，也就是说，每当 $d_n = 0$时，都会更新 τ_{eff}： 此时，将 τ_{eff} 设为该回合片段的有效结束时刻；在缓冲区时间中，这对应于该片段最后一个状态转移的 $\tau_{\text{eff}} = t_n + 1$ 。掩码 d_n 确保递归在回合边界处重置（因为ˆ$d_n = 0$ 意味着 $\mathbb{A}_n = \delta_n$），因此不同回合之间不会在存储于 $\mathcal{B}$中的数据间传播信息。ˆ最后，价值函数回归所用的回报目标按点计算为 $G_n = v_n + \mathbb{A}_n$。

E.7 Lunar Lander 实验

我们在 LunarLander-v3 环境中，通过固定折扣因子 $\gamma = 0.999$，对 Stable-Baselines3 (SB3) 的 PPO 实现 [24] （采用标准截断广义优势估计）与我们的广义优势估计有限时间变体进行实证比较。整个过程中，我们保持 PPO 算法和架构不变，只修改优势估计（从而改变由此产生的价值目标），因此观察到的差异可归因于估计器。

所有比较均在三种超参数（HP）设置下进行。首先，我们使用 SB3-Zoo 默认的 PPO超参数运行每种方法。其次，针对每种广义优势估计方法（截断、固定时刻、终止时间）分别进行超参数优化（HPO），使用 TPE采样器进行 100 次试验，且不进行剪枝。每次试验均使用 3 个随机种子进行评估，优化目标为汇总这 3 个随机种子结果后的最终折扣评估回报。随后，将每种方法找到的最佳 HP 配置用于绘制学习曲线。第三，为了检验鲁棒性，并排除性能提升纯粹源于调参改进的可能性，我们还使用截断广义优势估计 HPO 得到的 HP 配置评估所有方法（即所有估计器共用一套 HP 配置）。这样便可在默认开箱即用设置、每种方法各自可达到的最佳调参结果，以及共享的基线调参配置下进行受控比较。表1汇总了超参数搜索空间，以及 HPO 为每种优势估计器找到的最佳配置。

训练期间，我们每经过 5000 个环境步数就暂停学习一次，并使用 5 个回合评估当前策略。评估指标见图 5，训练诊断见图 6。每种方法和每种 HP 方案的曲线均取 20 个独立训练随机种子的平均值。阴影区域表示随机种子间的标准误。

总体而言，终止时间估计器始终呈现最快的学习动态和最短的回合长度，表明其能够更快、更可靠地完成着陆。在 SB3-Zoo 默认的 PPO超参数下，各估计器之间的性能差距最为明显。在这一设置中，终止时间估计器的学习速度显著更快：它更早达到较高回报，并在整个训练过程中实现更短的回合长度。与截断广义优势估计相比，固定时刻方法有时能改善早期学习表现，但通常在回报提升速度或回合长度的持续缩短方面，都无法达到终止时间估计器变体的水平。

超参数优化的超参数搜索空间		标准	固定时间终止时间	
轨迹展开				
n_步	2^p, $p \sim$ 均匀分布$\{5, \ldots, 12\}$	256	256	256
GAE				
广义优势估计_lambda	$\lambda = 1 - \epsilon, \quad \epsilon \sim$ 对数均匀分布$[10^{-4}, 10^{-1}]$	0.94748	0.99989	0.95827
优化				
批次大小_	$2^p, \quad p \sim$ 均匀分布$\{4, \ldots, 10\}$	128	256	16
学习_率	对数均匀分布$[10^{-5}, 2 \cdot 10^{-3}]$	1.037×10^{-4}	5.967×10^{-4}	1.687×10^{-4}
n_训练轮次	分类分布$\{1, 5, 10, 20\}$	20	20	10
最大_梯度_范数	均匀分布$[0.3,\ 2.0]$	0.3589	1.3625	1.8185
PPO目标 / 正则化				
截取_范围	分类分布$\{0.1, 0.2, 0.3, 0.4\}$	0.3	0.2	0.1
熵系数_	对数均匀分布$[10^{-8},\ 10^{-1}]$	1.323×10^{-5}	2.324×10^{-4}	5.482×10^{-3}
目标_kl	对数均匀分布$[10^{-3},\ 5.0]$	3.06	3.86×10^{-2}	1.75
网络				
网络架构_	分类分布{`[64]`, `[64,64]`, `[256,256]`}	`[256,256]`	`[256,256]`	`[256,256]`
激活函数_	分类分布{双曲正切函数, 修正线性单元}	ReLU	ReLU	ReLU

表1：用于基于树结构 Parzen 估计器的超参数优化的超参数搜索空间（100 次试验，每次试验 3 个随机种子）以及为每个估计器找到的最佳配置。折扣因子固定为 $\gamma = 0.999$。分类分布$\{\cdot\}$表示分类采样（在所列值上均匀采样），均匀分布$[a, b]$ 表示在 $[a, b]$ 上进行连续均匀采样，均匀分布$\{a, \ldots, b\}$ 表示在 $\{a, \ldots, b\}$ 上进行均匀均匀采样，而对数均匀分布$[a, b]$表示在 $[a, b]$ 上进行对数均匀采样。

在针对每种估计器分别优化超参数后，定性排名仍大致相似。终止时间估计器的评估回报增长速度仍然最快，并且达到的回合长度最短。固定时刻方法在初期表现出大幅改进，但其学习曲线随后变得与截断广义优势估计变体相似。

当所有方法都使用通过优化截断广义优势估计获得的超参数进行评估时，终止时间估计器仍然能够更快地学习。具体而言，与截断广义优势估计和固定时间 GAE 相比，回报和着陆速度（回合长度）都更早得到改善。这表明，观察到的收益并非仅仅是逐方法调优造成的假象，而是反映了终止自适应重新归一化所带来的更稳健的学习行为。

图 6 展示了训练期间的价值函数解释方差和价值损失，这些是评论家估计的诊断指标。在默认超参数下，终止时间估计器的解释方差显著早于其他估计器接近 1，并且其价值损失始终明显更小。一种合理的解释是，终止时间重新归一化降低了优势标签的方差ˆ，进而降低了用于评论家回归的价值目标 $G_n = v_n + \mathbb{A}_n$ 的方差。从这个意义上说，评论家面对的是一个条件更好的监督学习问题，标签噪声也更少，因此价值拟合能够更快稳定下来，进而为策略更新提供更可靠的优势估计。

使用截断优化的超参数配置时，所有方法的解释方差通常会在训练早期上升，随后短暂下降，然后再次上升。值得注意的是，这一下降发生在评估回合长度急剧缩短的同一时期左右，这表明训练机制发生了变化：策略从粗略控制转变为持续成功

着陆。这种转变可能会导致访问状态分布和回报结构发生显著变化，从而暂时降低评论家拟合效果。在此阶段，终止时间估计器的解释方差降幅明显较小，且价值损失也更低，这与回归目标稳定性有所提升的情况一致。

在采用各方法分别优化后的超参数时，终止时间 GAE 与截断广义优势估计呈现出大体相似的评论家诊断指标，而固定时刻方法的解释方差可能明显更低。一个可能的原因是，针对固定时刻方法的超参数优化选择了更接近 1 的 λ 值，这会加强对长时域成分的侧重，并可能增大优势与价值目标的方差，从而使评论家拟合更加困难，即便策略在初期提升较快。

摘要。 在所有超参数设置下，我们的终止时间 GAE 都提高了 LunarLander-v3 上的学习速度和着陆效率。训练诊断表明，这些提升伴随着评论家学习更快、更稳定（解释方差更高，价值损失更低），这与以下假设相符：终止自适应重新归一化可降低有限时域截断和提前终止所引起的方差。

E.8 连续控制实验

鉴于我们在 Lunar Lander 上的实证评估表明，所提出的终止时间 GAE 权重修正能够在终止效应显著的环境中带来明显提升，我们进一步评估这些改进能否推广至连续控制任务。由于有效时域较大时，固定时间 GAE 与标准截断 GAE 的差异很小，我们将重点放在实证影响最大的变体上，并在 MuJoCo [36] 基准测试 Ant-v4 上，将终止时间 GAE 与标准截断 GAE 进行比较。

结果如图 7 所示。所有运行均采用 SB3-Zoo [24] 中的默认近端策略优化超参数。学习曲线表明，终止时间 GAE 仍具有竞争力，与截断广义优势估计相比，还可能提高学习速度。需要强调的是，这项 Ant-v4 研究只是在默认超参数和较短训练时间下进行的最低限度连续控制验证。更全面地评估 MuJoCo 任务和不同调参方案，将留待未来工作开展。

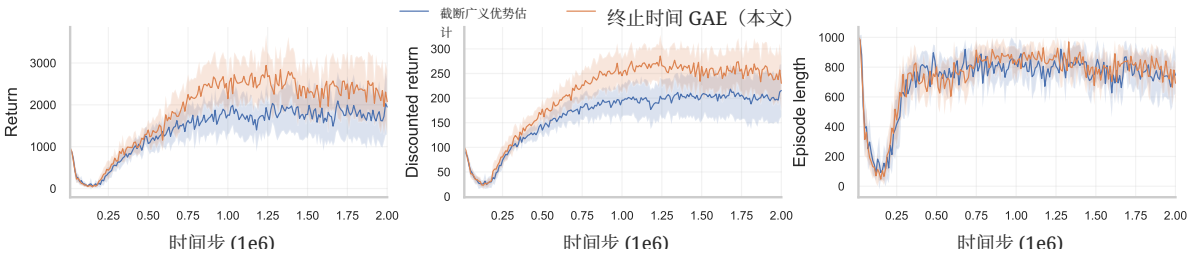


图 7：Ant-v4 评估学习曲线，对比在 Stable-Baselines3 Zoo 默认 PPO超参数下的截断 GAE 与我们的终止时间 GAE。各列分别展示评估回报（奖励总和）、折扣评估回报（折扣奖励总和）和评估回合长度。曲线为 10 个随机种子的平均值，阴影区域表示随机种子间的标准误。

E.9 玩具模型：独立同分布的时序差分误差下的协方差结构 s

本节介绍一个经过刻意简化的玩具模型，旨在分离由广义优势估计中指数衰减的时序差分误差聚合所引发的方差机制。我们假设时序差分误差是中心化独立同分布的，并推导所得优势序列协方差结构的闭式表达式。在这一受控设定下，我们通过对比截断广义优势估计与有限时间估计器的协方差函数，进而比较它们随时间产生的方差模式。这里有意采用较强的假设，以便所有量都能得到闭式表达式并绘制成图。

固定有限时域 $T \in \mathbb{N}$和参数 $\gamma,\ \lambda \in (0,\ 1)$。我们将时序差分误差 $(\delta_t)_{t=0}^{T-1}$ 视为驱动广义优势估计（GAE）的随机输入，并研究由此产生的优势估计在各个时刻所具有的协方差结构。我们采用以下独立性模型。

假设 E.10 （中心化独立同分布的时序差分误差）。 序列 $(\delta_t)_{t=0}^{T-1}$是独立同分布的，且

$$\mathbb{E}[\delta_t] = 0, \qquad \mathrm{Var}(\delta_t) = \sigma_\delta^2 < \infty.$$

考察 标准（有限时域截断）广义优势估计（GAE）所产生的优势序列 ，

$$\hat{\mathbb{A}}_t = \sum_{j=0}^{T-t-1} (\gamma\lambda)^j \, \delta_{t+j}, \quad t \in \{0, \dots, T-1\}. \tag{34}$$

并将其与我们的有限时间重归一化变体进行比较（此处为固定时域情形 $\tau = T$），

$$\hat{\mathbb{A}}_t^T = \sum_{j=0}^{T-t-1} (\gamma\lambda)^j \, \frac{1-\lambda^{T-t-j}}{1-\lambda^{T-t}} \, \delta_{t+j}, , \quad t \in \{0, \dots, T-1\}. \tag{35}$$

我们的目标是比较这两个估计器所引起的时间依赖性。为此，在假设 E.10 下，我们推导出它们的协方差函数的闭式表达式，并通过热力图展示由此得到的协方差矩阵及其差异。

关键步骤是进行重叠分解。由于时序差分误差相互独立，只有共享的时序差分误差项才会对协方差产生贡献。由此可得闭式公式，并可通过简单的支配性论证比较有限时间估计器与截断广义优势估计。

引理 E.11 (截断广义优势估计的协方差函数**)。** 在假设 *E.10* 下，由 $\hat{}(\mathbb{A}_t)_{t=0,\dots,T-1}$ 根据 (34) 给出的序列满足：对于任意 $t \in \{0, \dots, T-1\}$ 和 $k \in \{0, \dots, T-t-1\}$:

$$\mathrm{Cov}[\hat{\mathbb{A}}_t, \hat{\mathbb{A}}_{t+k}] = \sigma_\delta^2 (\gamma\lambda)^k \frac{1-(\gamma\lambda)^{2(T-t-k)}}{1-(\gamma\lambda)^2}. \tag{36}$$

证明。 由于根据假设 E.10，TD 误差被假定为中心化的，我们有

$$\begin{aligned}
\mathrm{Cov}[\hat{\mathbb{A}}_t, \hat{\mathbb{A}}_{t+k}] &= \mathrm{Cov}\left[\sum_{l=0}^{T-t-1} (\gamma\lambda)^l \, \delta_{t+l}, \ \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^m \, \delta_{t+k+m}\right] \\
&= \mathbb{E}\left[\left(\sum_{l=0}^{T-t-1} (\gamma\lambda)^l \, \delta_{t+l}\right)\left(\sum_{m=0}^{T-t-k-1} (\gamma\lambda)^m \, \delta_{t+k+m}\right)\right].
\end{aligned}$$

此外，由于 $(\delta_t)^{t\in\{0,\dots,T-1\}}$ 是独立的， $\mathbb{E}\left[\delta_{t+l}\,\delta_{t+k+m}\right]\neq 0$，仅当 $l=k+m$。因此，

$$\begin{aligned}\operatorname{Cov}[\hat{\mathbb{A}}_t,\hat{\mathbb{A}}_{t+k}] &= \sum_{l=0}^{T-t-1}\sum_{m=0}^{T-t-k-1}(\gamma\lambda)^{l+m}\,\mathbb{E}\left[\delta_{t+l}\,\delta_{t+k+m}\right]\\ &= \sum_{m=0}^{T-t-k-1}(\gamma\lambda)^{k+m+m}\,\sigma_\delta^2\\ &= \sigma_\delta^2(\gamma\lambda)^k\sum_{m=0}^{T-t-k-1}(\gamma\lambda)^{2m}\\ &= \sigma_\delta^2(\gamma\lambda)^k\frac{1-(\gamma\lambda)^{2(T-t-k)}}{1-(\gamma\lambda)^2}.\end{aligned}$$

□

接下来，我们计算由有限时间 GAE 给出的两两协方差估计量的公式。

引理 E.12（有限时域广义优势估计的协方差函数）。 在假设 *E.10* 下，由（$\hat{}\mathbb{A}_t^T$）$_{t=0,\ \dots,\ T-1}$ 给出的序列满足（35）对于任意 $t\in\{0,\dots,T-1\}$ 和 $k\in\{0,\dots,T-t-1\}$:

$$\begin{aligned}\operatorname{Cov}[\hat{\mathbb{A}}_t^T,\hat{\mathbb{A}}_{t+k}^T] = \frac{\sigma_\delta^2(\gamma\lambda)^k}{(1-\lambda^{T-t})(1-\lambda^{T-t-k})}\Bigg[&\frac{1-(\gamma\lambda)^{2(T-t-k)}}{1-(\gamma\lambda)^2}-2\lambda^{T-t-k}\frac{1-(\gamma^2\lambda)^{T-t-k}}{1-\gamma^2\lambda}\\ &+\lambda^{2(T-t-k)}\frac{1-\gamma^{2(T-t-k)}}{1-\gamma^2}\Bigg].\end{aligned}\tag{37}$$

证明。 根据（35），我们得到时序差分误差表示

$$\hat{\mathbb{A}}_t^T=\sum_{\ell=0}^{T-t-1}w_\ell^t\,\delta_{t+\ell},\qquad w_\ell^t:=(\gamma\lambda)^\ell\frac{1-\lambda^{T-t-\ell}}{1-\lambda^{T-t}}.$$

由于时序差分误差根据假设 E.10 进行了中心化，我们得到

$$\begin{aligned}\operatorname{Cov}[\hat{\mathbb{A}}_t^T,\hat{\mathbb{A}}_{t+k}^T] &= \operatorname{Cov}\left[\sum_{\ell=0}^{T-t-1}w_\ell^t\delta_{t+\ell},\ \sum_{m=0}^{T-t-k-1}w_m^{t+k}\delta_{t+k+m}\right]\\ &= \mathbb{E}\left[\left(\sum_{\ell=0}^{T-t-1}w_\ell^t\delta_{t+\ell}\right)\left(\sum_{m=0}^{T-t-k-1}w_m^{t+k}\delta_{t+k+m}\right)\right]\end{aligned}$$

再次，除非 $\ell=k+m$，此时它等于 σ_δ^2。因此，

$$\begin{aligned}\operatorname{Cov}[\hat{\mathbb{A}}_t^T,\hat{\mathbb{A}}_{t+k}^T] &= \sum_{\ell=0}^{T-t-1}\sum_{m=0}^{T-t-k-1}w_\ell^t\,w_m^{t+k}\,\mathbb{E}[\delta_{t+\ell}\,\delta_{t+k+m}]\\ &= \sigma_\delta^2\sum_{m=0}^{T-t-k-1}w_{k+m}^t\,w_m^{t+k}.\end{aligned}$$

代入显式权重可得

$$
\begin{aligned}
\mathrm{Cov}[\hat{\mathbb{A}}_t^T, \hat{\mathbb{A}}_{t+k}^T] &= \sigma_\delta^2 \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{k+m} \frac{1-\lambda^{T-t-k-m}}{1-\lambda^{T-t}} (\gamma\lambda)^m \frac{1-\lambda^{T-t-k-m}}{1-\lambda^{T-t-k}} \\
&= \frac{\sigma_\delta^2(\gamma\lambda)^k}{(1-\lambda^{T-t})(1-\lambda^{T-t-k})} \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m} (1-\lambda^{T-t-k-m})^2 \\
&= \frac{\sigma_\delta^2(\gamma\lambda)^k}{(1-\lambda^{T-t})(1-\lambda^{T-t-k})} \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m} \Big(1 - 2\lambda^{T-t-k-m} + \lambda^{2(T-t-k-m)}\Big) \\
&= \frac{\sigma_\delta^2(\gamma\lambda)^k}{(1-\lambda^{T-t})(1-\lambda^{T-t-k})} \Bigg[\underbrace{\sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m}}_{=:S_1} - 2 \underbrace{\sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m}\lambda^{T-t-k-m}}_{=:S_2} \\
&\qquad + \underbrace{\sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m}\lambda^{2(T-t-k-m)}}_{=:S_3} \Bigg],
\end{aligned}
$$

with

$$
\begin{aligned}
S_1 &= \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m} = \frac{1-(\gamma\lambda)^{2(T-t-k)}}{1-(\gamma\lambda)^2}, \\
S_2 &= \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m}\lambda^{T-t-k-m} = \lambda^{T-t-k} \sum_{m=0}^{T-t-k-1} \gamma^{2m}\lambda^m = \lambda^{T-t-k} \frac{1-(\gamma^2\lambda)^{T-t-k}}{1-\gamma^2\lambda}, \\
S_3 &= \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{2m}\lambda^{2(T-t-k-m)} = \lambda^{2(T-t-k)} \sum_{m=0}^{T-t-k-1} \gamma^{2m} = \lambda^{2(T-t-k)} \frac{1-\gamma^{2(T-t-k)}}{1-\gamma^2}.
\end{aligned}
$$

将 S_1, S_2, S_3 代入协方差表达式，恰好得到（37）。 □

引理 E.13（截断 GAE 协方差更大）。设 $(\hat{\mathbb{A}}_t)_{t=0}^{T-1}$ 与 $(\hat{\mathbb{A}}_t^T)_{t=0}^{T-1}$ 为 (34) 和 (35) 给出的序列。在假设 *E.10* 下，对任意 $t \in \{0, \dots, T-1\}$ 与 $k \in \{0, \dots, T-t-1\}$:

$$
0 \le \mathrm{Cov}[\hat{\mathbb{A}}_t^T, \hat{\mathbb{A}}_{t+k}^T] \;\le\; \mathrm{Cov}[\hat{\mathbb{A}}_t, \hat{\mathbb{A}}_{t+k}].
$$

图 8 中的协方差热图展示了在简化假设下，截断有限时域广义优势估计和我们的有限时间（重新归一化）广义优势估计变体所产生的协方差矩阵，从中可以了解方差行为的结构性成因。在所有 (T, λ) 配置下，有限时间方法的方差和协方差均一致较小，即 $\mathrm{Cov}[\hat{\mathbb{A}}_s^T, \hat{\mathbb{A}}_t^T] \le \mathrm{Cov}[\hat{\mathbb{A}}_s, \hat{\mathbb{A}}_t]$ 逐元素成立。这与引理 E.13 的支配结果一致；该结果可由将每个优势估计表示为独立同分布的时序差分误差的加权和得出：固定时刻方法会额外根据时域对较晚的时序差分误差进行衰减，乘数因子不超过 1，因此只会降低二阶矩。改变 λ 可以揭示指数衰减的时序差分误差聚合如何产生时间相关性。随着 λ 增大，有效权重 $(\gamma\lambda)^k$ 随时间偏移 $k = |t-s|$ 的衰减变慢，因此不同时刻的优势估计会共享更大比例的

共同的时序差分误差项。在热力图中，这表现为对角线周围的协方差带变宽：当 λ 较大时，较大的时间间隔之间仍存在显著协方差；而当 λ 较小时，协方差则集中在对角线附近。

引理 *E.13* 的证明。固定 $0 \le t \le T-1$ 和 $0 \le k \le T-t-1$。根据有限时间估计器的时序差分误差表示（见引理 E.12 的证明），我们有

$$\mathrm{Cov}[\hat{\mathbb{A}}_t^T, \hat{\mathbb{A}}_{t+k}^T] = \sigma_\delta^2 \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{k+m} \frac{1-\lambda^{T-t-k-m}}{1-\lambda^{T-t}} (\gamma\lambda)^m \frac{1-\lambda^{T-t-k-m}}{1-\lambda^{T-t-k}}.$$

所有被加数均为非负数，因此 $\mathrm{Cov}[\hat{}\mathbb{A}_s^T, \hat{}\mathbb{A}_t^T] \ge 0$。此外，权重有界，

$$(\gamma\lambda)^{k+m} \frac{1-\lambda^{T-t-k-m}}{1-\lambda^{T-t}} (\gamma\lambda)^m \frac{1-\lambda^{T-t-k-m}}{1-\lambda^{T-t-k}} \le (\gamma\lambda)^{k+m}(\gamma\lambda)^m = (\gamma\lambda)^{k+2m}$$

因此，

$$\mathrm{Cov}[\hat{\mathbb{A}}_t^T, \hat{\mathbb{A}}_{t+k}^T] \le \sigma_\delta^2 \sum_{m=0}^{T-t-k-1} (\gamma\lambda)^{k+2m} = \mathrm{Cov}[\hat{\mathbb{A}}_t, \hat{\mathbb{A}}_{t+k}],$$

其中，最后一个等式由截断广义优势估计的显式重叠公式得出（参见引理 E.11 的证明）。

□

截断方法与固定时刻方法的协方差之间的差异主要集中在轨迹展开末端附近（矩阵的右上区域）。这种局部性直接源于固定时重新加权：它用重新归一化方案取代标准几何加权，使接近时域末端的时序差分误差按 $(1-\lambda^{L-j})/(1-\lambda^L)$ 形式的因子降低权重，其中 $L = T-t$ 是剩余时域。当 t 远离终端边界（ L 较大）时，在大部分相关时序差分误差上，这些因子都接近 1，因此矩阵主体部分的协方差结构与截断广义优势估计相符。当 t 接近边界（ L 较小）时，较晚出现的时序差分误差会受到显著抑制，从而使协方差大幅降低，这一现象在右上角最为明显。随着 T 增大，接近时域末端的索引所占比例缩小，因此固定时刻方法与截断方法存在实质差异的区域相对变小，尽管逐元素支配关系仍然成立。

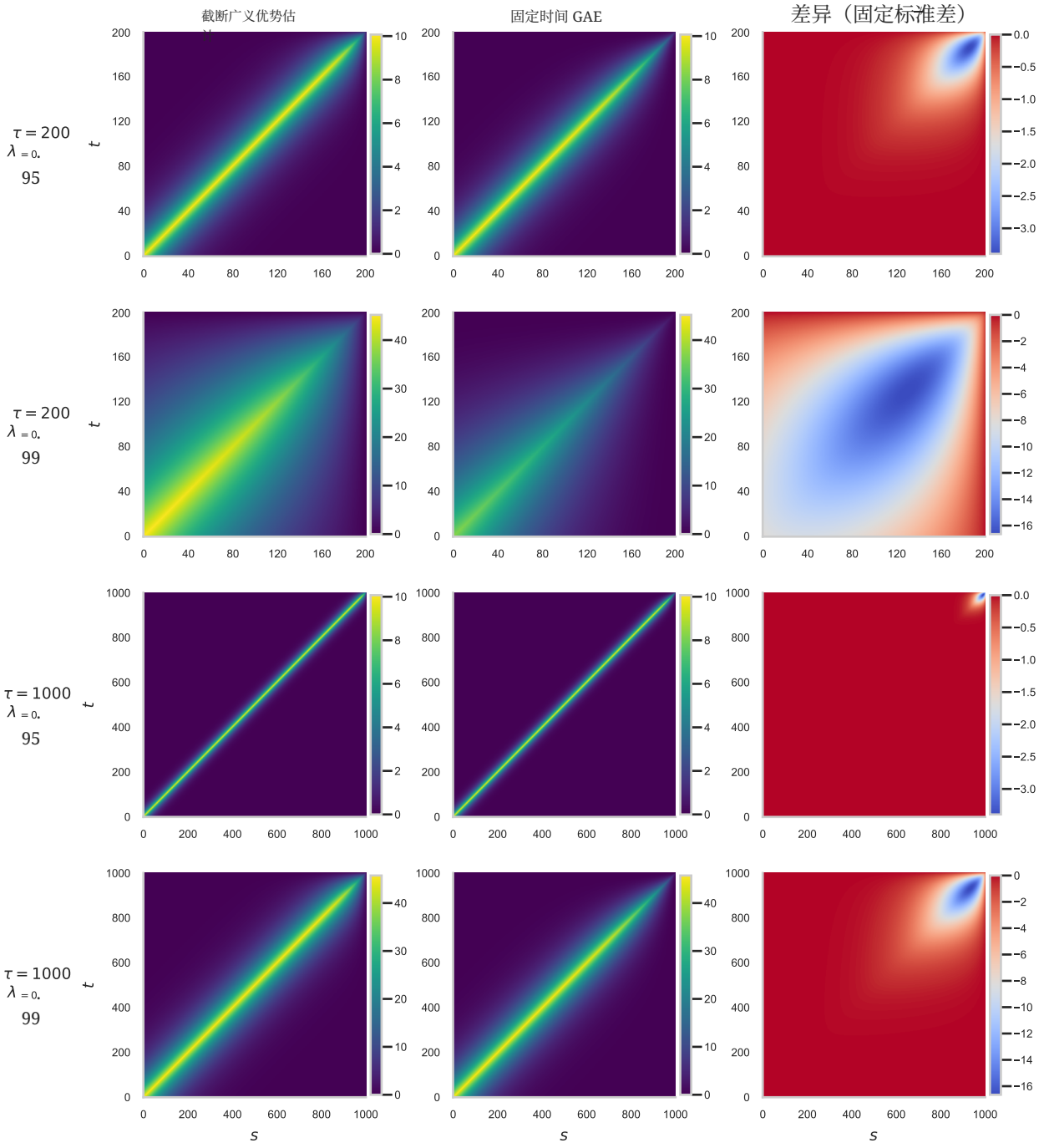


图 8：固定 $\gamma = 0.999$和 $\sigma_\delta^2 = 1.0$.下，截断广义优势估计与固定时间 GAE 优势的协方差结构。对于每种 $(T,\ \lambda)$ 配置（各行），我们绘制截断广义优势估计（左）和有限时间 GAE（中）的 Cov[$\hat{\mathbb{A}}_s,\ \hat{\mathbb{A}}_t$]，以及逐项差值（右）。正如权重支配性质所预测的，固定时协方差均匀地更小，且最大降幅出现在时域末端附近。