arXiv:2302.01734v2 [cs.LG] 8 Nov 2023

随机策略梯度方法：Fisher 非退化策略的样本复杂度改进

Ilyas Fatkhullin [1] **Anas Barakat** [1] **Anastasia Kireeva** [2] **Niao He** [1]

摘要

近年来，策略梯度（PG）方法在实证上的显著成功推动了其理论基础的发展。尽管人们为设计高效的随机 PG 类算法付出了巨大努力，但对这些算法收敛至全局最优策略的理解仍然有限。本文针对一类广义的 Fisher 非退化参数化策略，提出了改进的全局收敛保证，从而能够处理连续状态—动作空间的情形。首先，我们提出一种带隐式梯度传输的归一化策略梯度方法（N-PG-IGT），并推导出该方法寻找全局 ε-最优策略所需的 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ 样本复杂度。与此前已知的 $\tilde{\mathcal{O}}(\varepsilon^{-3})$ 复杂度相比，该算法无需使用重要性采样或二阶信息，且每次迭代仅采样一条轨迹。其次，我们进一步通过考虑一种经海森-向量积校正增强的海森矩阵辅助递归策略梯度（(N)-HARPG）算法，将这一复杂度提升至 $\tilde{\mathcal{O}}(\varepsilon^{-2})$。值得注意的是，这两种算法均（$i$）简单易实现：采用单循环结构，无需大量轨迹批次，且每次迭代最多采样两条轨迹；（ii）计算和内存效率高：每次迭代无需执行开销昂贵的子程序，且所需内存与参数维度呈线性关系。

1. 引言

追溯至 Williams (1992)；Sutton 等人(1999)，策略梯度（PG）方法是最受欢迎且最高效的强化学习（RL）算法类别之一。结合深度神经网络，现代策略梯度算法（Silver 等人，2014；Schulman 等人，2015；2017）已在若干涉及大型乃至连续状态空间和动作空间的挑战性任务中取得了令人瞩目的实证成果。

在这类方法中，为了学习参数化策略，智能体会在策略参数空间中执行随机梯度上升，以最大化其期望回报 $J(\theta)$，其中 $\theta \in \mathbb{R}^d$ 是策略参数化的参数。鉴于策略优化目标具有非凹性，文献中的若干研究着重于策略参数收敛至 ε-近似一阶驻点（FOSP），即 $\|\nabla J(\theta)\| \leq \varepsilon$。这一性能指标使我们能够将策略梯度方法的样本复杂度定义为：找到期望回报函数的一个 ε-近似 FOSP 所需的随机轨迹模拟（或样本）数量。具体而言，采用与非凸平滑优化中的随机梯度下降类似的分析方法（Ghadimi 和 Lan，2013），研究表明，原始策略梯度（REINFORCE）的样本复杂度为 $\tilde{\mathcal{O}}(\varepsilon^{-4})$（例如，参见 Yuan 等人（2022b））。过去几年，受随机优化中方差缩减技术发展的推动，大量研究致力于设计策略梯度方法的方差缩减变体，以提高样本效率。与经典监督学习不同，强化学习中的采样分布会因策略更新而呈现非平稳性。因此，要将方差缩减技术应用于策略梯度方法，必须谨慎处理这种非平稳性。大多数研究通过引入重要性采样权重来纠正分布偏移。然而，使用重要性采样机制通常需要一个难以验证的强假设，即重要性采样权重的方差有界。近期，Shen 等人（2019）和 Salehkaleybar 等人（2022）提出了一种有望替代重要性采样的方法，该方法利用二阶信息来纠正策略更新造成的分布偏移。在样本复杂度方面，这两种方法均保证使用 $\tilde{\mathcal{O}}(\varepsilon^{-3})$ 个样本找到一个 ε-FOSP。近期研究表明，对于随机优化问题，即便能够获得更高阶的随机预言机和光滑性，该复杂度仍是最优且无法进一步改进的（Arjevani 等人，2020）。此

[1]苏黎世联邦理工学院计算机科学系，瑞士 [2]苏黎世联邦理工学院数学系，瑞士。通信作者：I.F. <ilyas.fn979@gmail.com>。

论文集：40th 国际机器学习会议，美国夏威夷檀香山。PMLR 202，2023。版权所有 © 2023 作者。

因此，需要利用策略优化问题的特殊结构来改进策略梯度方法的样本复杂度。

	原始策略梯度	N-PG-IGT	(N)-HARPG
FOSP	$\tilde{\mathcal{O}}(\varepsilon^{-4})$	$\tilde{\mathcal{O}}(\varepsilon^{-3.5})$ (new)[1]	$\tilde{\mathcal{O}}(\varepsilon^{-3})$
全局	$\tilde{\mathcal{O}}(\varepsilon^{-3})$	$\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ （新）	$\tilde{\mathcal{O}}(\varepsilon^{-2})$ （新）

[1] 基于隐式梯度传输的方法此前已在随机优化中用于寻找 FOSP（Cutkosky 和 Mehta，2020年），但我们尚未发现其在强化学习中的应用。

表1：计算效率较高的 策略梯度方法复杂度结果汇总。表格第一行列出了找到一个 ε-FOSP 所需的样本数，第二行则列出了在假设 ε下找到3 和4-全局最优解所需的样本数。

近来，一系列研究致力于建立全局最优收敛率，并推导达到期望回报函数的 ε近似全局最优解所需的样本复杂度（即 $J^* - J(\theta) \leq \varepsilon$，其中 J^*为最优回报），而非仅仅达到 ε近似 FOSP（Agarwal 等人，2021；Fazel 等人，2018年；Fatkhullin 和 Polyak，2021；Bhandari 和 Russo，2019年；Mei 等人，2020年；Zhang 等人，2020b；2021b）。这方面日益丰富的研究利用了策略优化问题的额外结构，具体表现为梯度支配（Agarwal 等人，2021）、Łojasiewicz 型不等式（Mei 等人，2020年；2021b）或隐式凸性（Zhang 等人，2021a）。正如这些结构性质先前在确定性优化中（Lojasiewicz，1963年；Polyak，1963年；Kurdyka，1998年）（例如，另见 Attouch 和 Bolte（2009年））以及后来在随机优化中（Fontaine 等人，2021；Fatkhullin 等人，2022年；Scaman 等人，2022年；Masiha 等人，2022）所利用的那样，这些性质保证非凹的策略优化目标不存在次优驻点，并使得推导函数值的收敛速率成为可能。这些研究大多假设可以访问精确的策略梯度，以证明快速收敛速率（Mei 等人，2020年；2021b；Xiao，2022）。

相比之下，具有实际重要性的随机梯度问题，直到最近才有少数研究予以探讨（Liu 等人，2020年；Ding 等人，2022年；Yuan 等人，2022b；Masiha 等人，2022年）。这一系列研究确立了期望回报的全局收敛保证（但存在由策略参数化引起的偏差）。其分析中的一个关键工具是：对于费希尔非退化参数化策略这一类策略，期望回报函数满足宽松弱梯度支配不等式；高斯策略是其中的典型例子。基于这一结构性质，Liu 等人（2020年）；Ding 等人（2022年）；Yuan 等人（2022b）实现了一个$\tilde{\mathcal{O}}(\varepsilon^{-3})$全局最优性速率，但存在由策略参数化导致的偏差误差项。

最近，Masiha 等人（2022年）通过分析一种随机二阶方法，进一步将这一复杂度提升至 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$。然而，他们的算法需要近似求解一个三次子问题，这要求在每次迭代中额外执行 $\tilde{\mathcal{O}}(\varepsilon^{-3})$ 次运算，并带来额外的计算负担。由这些最新结果引出的一个自然问题是：

> 是否可以使用计算效率高的策略梯度算法，改进一般 *FND* 参数化策略的 $\tilde{\mathcal{O}}(\varepsilon^{-3})$ 样本复杂度？

在这项工作中，我们对此给出了肯定的回答。我们的贡献总结如下。

1.1. 贡献总结

- 首先，我们提出了一种带隐式梯度传输的归一化动量策略梯度方法（N-PG-IGT），并建立了 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ 全局收敛样本复杂度。有趣的是，该算法无需重要性采样机制，不使用海森矩阵估计等曲率信息，并且每次迭代只采样一条轨迹。

- 接下来，我们考虑海森矩阵辅助动量策略梯度方法的两种变体：(*i*) 带归一化的N-HARPG，以及(*ii*) 不带归一化的HARPG 。我们对这两种变体进行了更精细的收敛分析，并证明它们能够实现改进后的全局收敛，样本复杂度为 $\tilde{\mathcal{O}}(\varepsilon^{-2})$。与大多数先前研究不同，我们的分析不要求重要性采样权重有界这一强假设，因为HARPG不使用重要性采样。此外，这些算法采用单循环结构，每次迭代只需采样两条轨迹，并且可以高效实现，其内存和计算复杂度与海森矩阵无关方法相近。

我们强调，与该领域以往相关工作相比，我们改进后的样本复杂度是在相同（或限制更少）的假设下推导得出的（Liu 等人，2020年；Ding 等人，2021；Yuan 等人，2022b；Masiha 等人，2022年）。我们请读者参阅表1 和2，以了解我们的贡献在相关文献中的定位。

1.2. 相关工作

我们简要回顾与随机策略梯度方法全局收敛结果最密切相关的文献。关于方差缩减策略梯度方法的一阶驻点性以及精确策略梯度方法的全局最优性，我们将在附录 D 中进行更详尽的文献回顾。为方便读者，我们请读者参阅表 2，了解相关工作和本文主要贡献的概况。

近期只有少数研究为随机策略梯度的情形提供了全局收敛保证（Liu 等人，2020年；Zhang 等人，2021a；Ding 等人，2021；2022年；Yuan 等人，2022b；Masiha 等人，2022年）。

我们设定下的已有结果。对于*Fisher*-非退化参数化策略类，Liu 等人（2020年）提出了随机方差缩减方法和自然策略梯度方法，并证明其寻找 $\tilde{\mathcal{O}}(\varepsilon^{-3})$ 全局最优策略的样本复杂度为 ε（由于策略参数化带来的相容函数逼近误差除外）。随后，Ding 等人（2022年）使用单循环、有限批量且基于动量的策略梯度算法，达到了相同的样本复杂度。然而，上述工作要求作出重要性采样权重方差有界这一强假设。最近，Yuan 等人（2022b）表明，在类似假设下，对于相同的策略类，随机原始策略梯度也具有相同的 $\tilde{\mathcal{O}}(\varepsilon^{-3})$ 样本复杂度。近期，Masiha 等人（2022年）提出随机三次正则化方法（SCRN），在类似假设下实现了 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ 的样本复杂度。尽管如此，该方法在每次迭代中都需要调用开销很大的子程序来求解三次正则化子问题，并且需要使用大量样本批次来估计梯度和海森矩阵（批量大小取决于维度 d 和所需精度 ε，参见其中的引理 11）。Masiha 等人（2022年）进一步将复杂度改进至 $\tilde{\mathcal{O}}(\varepsilon^{-2})$，方法是针对随机优化考虑 SCRN 的方差缩减变体（参见其中的算法 2）。不过，该算法同样需要较大的批量。此外，将该方法应用于强化学习并非易事，因为这需要为随机梯度和海森矩阵引入重要性采样权重，并假设它们的方差有界，以处理强化学习中固有的非平稳性。

表格型策略和 Softmax 策略的全局样本复杂度。在表格型设定中，Lan（2022）证明，随机策略镜像下降方法的样本复杂度为 $\tilde{\mathcal{O}}(\varepsilon^{-2})$。然而，他们的分析仅适用于状态—动作空间为离散且不使用参数化的表格型设定。在 Softmax 策略参数化下，Zhang 等人（2021a）通过采用批次大小较大的双循环方差缩减策略梯度算法，使用 $\tilde{\mathcal{O}}(\varepsilon^{-2})$ 个样本获得了全局 ε-最优策略。有趣的是，该算法利用 softmax 策略参数化的特殊结构实现梯度截断机制，从而放宽了对重要性采样权重方差有界的假设。值得注意的是，Zhang 等人（2021a）仅关注 softmax 策略参数化，并采用一个难以验证的过度参数化假设（假设 5.11）来利用隐式凸性。就在最近，在完成本文准备工作之际，我们发现了 Alfano 和 Rebeschini（2022）以及 Yuan 等人（2022a）的研究，这些工作分析了对数线性策略下自然策略梯度及其 Q 版本（Q-NPG）的收敛性。具体而言，Yuan 等人（2022a）借助兼容函数逼近框架，实现了 $\tilde{\mathcal{O}}(\varepsilon^{-2})$ 的样本复杂度，使用框架。此外，在每次迭代中，NPG 和Q-NPG 都需要将随机梯度下降（SGD）作为子程序，以求解回归问题并确定更新方向，同时还依赖复杂的采样子程序（算法3、4，第28–30页）。与上述所有针对离散有限状态—动作空间、采用直接参数化和 softmax 参数化的研究相辅相成，我们通过 FND 策略类分析连续状态—动作空间的情形；这是一个本质上不同的问题设定。

2. 预备知识

2.1. 问题表述

我们考虑一个离散时间折扣马尔可夫决策过程（MDP）（Puterman，2014）$\mathcal{M}=(\mathcal{S},\mathcal{A},\mathcal{P},r,\rho,\gamma)$，其中 $\mathcal{S}$ 和 $\mathcal{A}$ 分别是（可能连续的）状态空间和动作空间，$\mathcal{P}$：$\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}_+$ 是状态转移密度，r：$\mathcal{S}\times\mathcal{A}\to[-r_{\max},r_{\max}]$ 是奖励函数，其有界于 $r_{\max}>0$，ρ 是初始状态分布，$\gamma\in$ (0,1)是折扣因子。对于每个参数 $\theta\in\mathbb{R}^d$和状态 $s\in\mathcal{S}$，参数化策略均定义为动作空间 $\mathcal{A}$ 上的概率密度函数 $\pi_\theta(\cdot|s)$。在每个时间步 $t\in\mathbb{N}$，当处于状态 $s_t\in\mathcal{S}$时，强化学习智能体根据分布 $\pi_\theta(\cdot|s_t)$选择动作 $a_t\in\mathcal{A}$，观察到奖励 $r(s_t,a_t)$，并且环境根据 $\mathcal{P}(\cdot|s_t,a_t)$ 转移到区域 $\mathcal{S}'\subset\mathcal{S}$内的状态 s_{t+1}。

强化学习智能体的目标是找到一个策略 π_θ，使下文定义的期望回报最大化：

$$J(\theta)\stackrel{\text{def}}{=}\mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{+\infty}\gamma^t r(s_t,a_t)\right], \tag{1}$$

其中，期望是相对于马尔可夫链 $(s_t,\ a_t)_{t\in\mathbb{N}}$ [1] 智能体只能访问由马尔可夫决策过程生成的有限长度轨迹 H，而状态转移核 $\mathcal{P}$ 和奖励函数 $r(\cdot,\ \cdot)$ 均未知。由初始分布 ρ 和策略 π_θ 在长度为 $H\geq 1$ 的轨迹空间上诱导的概率分布具有如下密度：

$$p(\tau|\pi_\theta)\stackrel{\text{def}}{=}\rho(s_0)\,\pi_\theta(a_0|s_0)\prod_{t=1}^{H-1}\mathcal{P}(s_t|s_{t-1},a_{t-1})\,\pi_\theta(a_t|s_t),$$

其中 $\tau=(s_0,\ a_0,\ \cdots,\ s_{H-1},\ a_{H-1})$。我们还将给定时域 H 下的截断期望回报定义为 $J_H(\theta)$ 定义为 $=\mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{H-1}\gamma^t r(s_t,\ a_t)\right]$.

[1]在本文余下部分中，若上下文清楚，我们通常会省略该期望中的下标。

算法	样本复杂度*	最后一次迭代	无重要性采样	无二阶平滑。	无批量	复杂度。高效
原始策略梯度 (Yuan 等人, 2022b)	$\frac{1}{(1-\gamma)^5}\varepsilon^{-3}$	✔	✔	✔	✔	✔
Natural-PG (Liu 等人，2020年)	$\frac{1}{(1-\gamma)^6}\varepsilon^{-3}$	✗	✔	✔	✗	✗
SRVR-PG (Liu 等人，2020年)	$\frac{1+W(1-\gamma)^3}{(1-\gamma)^{5.5}}\varepsilon^{-3}$	✗	✗	✔	✗	✔
STORM-PG-F (Ding 等人，2022年)	$\frac{1+W^{3/2}}{(1-\gamma)^{12}}\varepsilon^{-3}$	✗	✗	✔	✔	✔
SCRN (Masiha 等人，2022年)	$\frac{1}{(1-\gamma)^{7.5}}\varepsilon^{-2.5}$	✔	✔	✗	✗	✗
N-PG-IGT 本文	$\frac{1}{7.5}\varepsilon^{-2.5}$	✔	✔	✗	✔	✔
(N)-HARPG 本文	$\frac{1}{(1-\gamma)^6}\varepsilon^{-2}$	✔	✔	✔	✔	✔

表 2：FND 策略全局最优性收敛相关工作的概览。“样本复杂度”报告达到目标精度所需的样本轨迹数量（忽略对数因子），即当保证针对“最后一次迭代”时，达到 $J^*-\mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon+\frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}$ 所需的样本复杂度；当保证针对“最佳”迭代时，达到 $\min_{t=0,\cdots,T} J^*-\mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon+\frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}$ 所需的样本复杂度（ $\varepsilon_{\text{bias}}$ 的定义见假设 4）。“无 IS”表示该方法不使用重要性采样（IS）。如果算法使用 IS，则还需假设重要性采样权重的方差有界，即 $\mathbb{V}\left(\frac{p}{p}(\tau|\theta)\,(\tau|\theta')\right) \leq W, \forall\,\theta,\theta' \in \mathbb{R}^d$。“无二阶光滑性”表示不需要策略参数化的二阶光滑性（假设 2）。“无批量”表示算法不要求使用小批量。如果某方法在每次迭代中仅需 $\mathcal{O}(d)$ 次算术运算，我们称其为“计算高效”。相比之下，SCRN 和 Natural-PG 需要额外的子程序，参见 Masiha 等人（2022年，备注 6）和 Liu 等人（2020年，定理 4.9）。* 样本复杂度中隐含了对常数 $M_g, M_h, l_2, r_{max}, \mu_F$ 以及 ε^{-1} 中多对数因子的依赖。

2.2. 一阶平稳性与全局最优性

策略优化问题是可微且平滑的（我们将在后文明确一些正则性条件），但通常并非凹函数。因此，若干先前研究（综述见附录）专注于证明收敛到期望回报函数 J 的 ε-近似 FOSP，即收敛到满足以下条件的参数 θ

$\|\nabla J(\theta)\| \leq \varepsilon$. 本文关注于寻找满足以下条件的参数 θ，使得 $J^*-J(\theta) \leq \varepsilon+\mathcal{O}(\sqrt{\varepsilon_{\text{偏差}}})$，其中 J^* 是最优期望回报（最大化 $J(\pi)$），而 $\mathcal{O}(\sqrt{\varepsilon_{\text{偏差}}})$ 误差项源于策略参数化的近似能力（见假设4）。

2.3. 策略梯度与海森矩阵

在策略参数化满足温和的正则性假设时，可以证明，截断期望回报函数 J_H 关于策略参数 θ 的梯度以及在 θ 处的海森矩阵（见（Shen 等人，2019年））可写为:

$$\begin{aligned}\nabla J_H(\theta) &= \mathbb{E}_{\rho,\pi_\theta}[g(\tau,\theta)], \qquad (2)\\ g(\tau,\theta) &= \sum_{t=0}^{H-1}\left(\sum_{h=t}^{H-1}\gamma^h r(s_h,a_h)\right)\nabla\log\pi_\theta(a_t|s_t),\end{aligned}$$

$$\begin{aligned}\nabla^2 J_H(\theta) &= \mathbb{E}_{\rho,\pi_\theta}[B(\tau,\theta)], \qquad (3)\\ B(\tau,\theta) &\overset{\text{def}}{=} \nabla\Phi(\tau,\theta)\,\nabla\log p(\tau|\pi_\theta)^T+\nabla^2\Phi(\tau,\theta)\end{aligned}$$

with

$$\Phi(\tau,\theta) \overset{\text{def}}{=} \sum_{t=0}^{H-1}\left(\sum_{h=t}^{H-1}\gamma^h r(s_h,a_h)\right)\log\pi_\theta(a_t|s_t)$$

并且 τ 是从策略为 H的马尔可夫决策过程采样得到的长度为 π_θ 的轨迹，初始分布为 ρ。注意，所有导数均是关于 θ 的，而由于时域长度无限，无法计算 J 的完整梯度。因此，梯度 $\nabla J_H(\theta)$ 的无偏估计量为 $g(\tau,\theta)$，其定义见方程2。

3. 归一化动量策略梯度算法

本节介绍我们的两种基于动量的策略梯度算法: N-PG-IGT和N-HARPG。这两种方法具有两个关键的算法特征：归一化和动量。此外，N-PG-IGT 采用前瞻机制来利用二阶光滑性，而N-HARPG则采用带海森修正的方差缩减，以提高样本复杂度。

算法 1N-PG-IGT
(带隐式梯度传输的归一化策略梯度)

输入: $\theta_0, \theta_1, d_0, T, \{\eta_t\}_{t\geq 0}, \{\gamma_t\}_{t\geq 0}$
对于 $t = 1, \ldots, T-1$ **做**
  $\tilde{\theta}_t = \theta_t + \frac{1-\eta_t}{\eta_t}(\theta_t - \theta_{t-1})$
  $\tilde{\tau}_t \sim p(\cdot|\pi_{\tilde{\theta}_t})$
  $d_t = (1-\eta_t)d_{t-1} + \eta_t g(\tilde{\tau}_t, \tilde{\theta}_t)$
  $\theta_{t+1} = \theta_t + \gamma_t \frac{d_t}{\|d_t\|}$
</style}结束循环
返回 θ_T

3.1. 带隐式梯度传输的归一化策略梯度

我们首先介绍一种用于策略优化的策略梯度算法，称为带隐式梯度传输的归一化策略梯度（简称 N-PG-IGT）。我们提出了N-PG-IGT 算法，用于解决我们的强化学习策略优化问题。该算法的灵感来自近期提出的NIGT 方法，该方法是在随机优化的背景下提出的（从固定分布中获取独立同分布样本）（Cutkosky 和 Mehta，2020年；Arnold 等人，2019年）。

动机。 在介绍 N-PG-IGT算法之前，我们先说明一些动机。当能够获取精确的策略梯度时，简单的归一化策略梯度得益于策略优化目标的非均匀光滑性，具有几何全局收敛速率（Mei 等人，2021b）。因此，自然需要研究在更现实的情形下该算法的性质：此时只能利用从马尔可夫决策过程（MDP）中采样得到的样本计算随机策略梯度。在这种随机情形下，我们对该方法的理解仍然有限。因此，讨论随机优化中关于归一化随机梯度方法的已知结果颇有启发。如 Cutkosky 和 Mehta (2020年，第 2 节）所述，简单归一化随机梯度方法的问题在于，归一化前的梯度估计误差可能大于真实梯度。为解决这一问题，文献中的若干工作提出降低随机梯度的方差：一种做法是使用非常大的批次大小（例如参见（Hazan 等人， 2015））， 另一种做法则是不依赖大批量，而是像近期工作（Cutkosky 和 Mehta， 2020年）建议的那样使用动量。回到我们的策略优化问题，我们研究了带动量的归一化策略梯度在全局最优性方面的样本复杂度，超越了一阶驻点性的收敛速率。然而，该算法的结果仅与标准策略梯度的表现相当。我们将在附录 I中讨论这些结果，并正式介绍算法 3 ，同时在定理 8中证明其收敛保证。有了这些动机，我们现在可以介绍N-PG-IGT 算法。该算法同样采用带归一化的动量，并将帮助我们推导出更优的收敛速率。

N-PG-IGT（算法 1）。该算法沿着基于动量的归一化随机策略梯度方向，在策略参数空间中执行归一化梯度上升。与简单的归一化动量策略梯度算法不同，方向 d_t 是使用˜随机策略梯度 $g(\tilde{\tau}_t, \theta_t)$而非 $g(\tau_t, \theta_t)$˜和 $\tau_t \sim p(\cdot|\pi_{\theta_t})$计算得到的。这里，辅助序列 θ_t 对于加快收敛至关重要。注意，其更新规则（步骤 3）˜可以改写为 $\theta_t = \eta_t\theta_t + (1-\eta_t)\theta_{t-1}$ ，这表明˜使用 θ_t 可以视为执行一个前瞻步，即根据先前的迭代点 θ_t 和 θ_{t-1} 进行外推。有关该方案的更多直观解释，请参见（Cutkosky 和 Mehta，2020年，第 3 节）。请注意，该算法没有采用任何重要性采样机制来处理我们的强化学习问题中的非平稳性（这也是强化学习区别于标准随机优化之处）。算法在每次迭代时仅采样一条轨迹，无需使用大批量。

3.2. 海森矩阵辅助递归策略梯度

现在，我们介绍两种海森矩阵辅助策略梯度算法的变体，即海森矩阵辅助递归策略梯度（HARPG）及其归一化版本（N-HARPG）。这里说明一下，N-HARPG 此前由 Salehkaleybar 等人(2022年)提出，当时称为SHARP [2]。

(N)-HARPG （**见算法2**）。该算法基于一种与STORM 方差缩减方法（Cutkosky 和 Orabona，2019年）相似的思路，并采用二阶信息（Tran 和 Cutkosky，2021），而不是连续随机梯度之间的差值。有趣的是，这使得算法无需使用重要性采样，也无需通过不可验证的假设来控制重要性采样权重。为计算算法第 7 步中的更新方向 d_t ，将二阶修正项 $(1-\eta_t)\, v_t$ 加到动量随机梯度 $(1-\eta_t)\, d_{t-1} + \eta_t g(\tau_t, \theta_t)$ 。关键在于，第 3 至第 6 步中的均匀采样过程保证 v_t 是 $\nabla J(\theta_t) - \nabla J(\theta_{t-1})$ 的无偏估计量；这一项让人联想到原始STORM中的对应项。算法的最后一步（第 8 步）使用方向 d_t ，通过归一化或不归一化的方式更新参数 θ_t .

海森-向量实现（步骤6）。我们强调，定义校正项的海森-向量积 v_t 可以通过以下方式轻松高效地实现，计算复杂度为 $\mathcal{O}(Hd)$。注意，对于任意向量 $u \in \mathbb{R}^d$、任意轨迹 τ 以及任意参数 $\theta \in \mathbb{R}^d$，我们可以将海森-向量积写为

[2]我们为该算法采用了不同的名称，因为我们还研究了未进行归一化的算法（即 HARPG），并且在分析中为归一化算法使用了不同的步长和动量参数。

算法 2(N)-HARPG
（（归一化）-海森矩阵辅助递归策略梯度）

输入：$\theta_0, \theta_1, d_0, T, \{\eta_t\}_{t\geq 0}$，$\{\gamma_t\}_{t\geq 0}$
对于 $t = 1, \ldots, T-1$ **执行**
  $q_t \sim \mathcal{U}([0,1])$
  $\hat{\theta}_t = q_t\theta_t + (1-q_t)\theta_{t-1}$
  $\tau_t \sim p(\cdot|\pi_{\theta_t})$; $\hat{\tau}_t \sim p(\cdot|\pi_{\hat{\theta}_t})$
  $\hat{v}_t = B(\hat{\tau}_t, \theta_t)(\theta_t - \theta_{t-1})$
  $d_t = (1-\eta_t)(d_{t-1} + v_t) + \eta_t g(\tau_t, \theta_t)$
  $\theta_{t+1} = \begin{cases} \theta_t + \gamma_t d_t & \text{(HARPG)} \\ \theta_t + \gamma_t \frac{d_t}{\|d_t\|} & \text{(N-HARPG)} \end{cases}$
结束循环
返回 θ_T

，即 $B(\tau, \theta)\, u$ 利用方程 3 可得：

$$B(\tau,\theta)\, u = \langle \nabla \log p(\tau|\pi_\theta), u\rangle\, g(\tau,\theta) + \nabla \langle g(\tau,\theta), u\rangle .$$

第二项可以通过对标量值 $\langle g(\tau, \theta)\ u\rangle$ 进行自动微分来轻松计算，而不必像以往工作中那样使用有限差分近似（Shen 等人，2019年；Huang 等人，2020年）。无需计算和ˆ存储完整的海森矩阵估计 $B(\hat{\tau}_t, \theta_t)$ 。利用策略海森矩阵中的曲率信息不会增加 $\mathcal{O}(Hd)$ 每次迭代的计算成本。

4. 全局收敛分析

本节将证明，我们的算法达到全局最优性所需的样本复杂度，即在期望意义下达到一个 ε-近似全局最优解所需的样本数，同时考虑由策略参数化导致的偏差。为了提供此类全局最优性保证，我们利用非凹策略优化问题的一项结构性质：期望回报 J满足放宽的弱梯度支配条件。

4.1. 假设

我们的第一个假设是关于策略参数化正则性的标准假设。[3]

假设 1 (策略参数化正则性). 对于每个 $s, a \in \mathcal{S}\times\mathcal{A}$，函数 $\theta \mapsto \pi_\theta(a|s)$ 为正且连续可微。此外，

(i) 存在 $M_g > 0$，使得对于每个 $\theta \in \mathbb{R}^d, s \in \mathcal{S}, a \in \mathcal{A}$，$\|\nabla \log \pi_\theta(a|s)\| \leq M_g$.

(ii) 函数 $\theta \mapsto \pi_\theta(a|s)$ 对每个 s、$a \in \mathcal{S}\times\mathcal{A}$ 都是二阶连续可微的，并且存在 $M_h > 0$，使得对于每个 $\theta \in \mathbb{R}^d$ 以及每个 $s \in \mathcal{S}$、$a \in \mathcal{A}$、$\|\nabla^2$ 对数 $\pi_\theta(a|s)\| \leq M_h$.

[3]详见附录 B，其中对我们所有假设进行了更详细的讨论。

在此假设下，我们汇总了文献中的一些已知结果（例如，参见（Papini 等人，2018年；Shen 等人，2019年；Huang 等人，2020年；Xu 等人，2020a；Yuan 等人，2022b）），这些结果涉及目标函数 J及其随机策略梯度的性质，并将在下一个命题中给出。

命题 1。在假设 *1*-（*i*）下，

(i) Function J is L_g-smooth with $L_g = \frac{r_{\max}(M_g^2 + M_h)}{(1-\gamma)^2}$.

(ii) $\mathbb{E}\left[\|g(\tau|\theta) - \nabla J_H(\theta)\|^2\right] \leq \sigma_g^2$ *with* $\sigma_g^2 = \frac{r_{\max}^2 M_g^2}{(1-\gamma)^3}$.

(iii) $\mathbb{E}\left[\|B(\tau|\theta) - \nabla^2 J_H(\theta)\|^2\right] \leq \sigma_h^2$ *with* $\sigma_h^2 = \frac{r_{\max}^2(H^2 M_g^4 + M_h^2)}{(1-\gamma)^4}$ *if Assumption 1-(ii) also holds.*

条目（i）和（ii）见 Yuan 等人（2022b，引理 4.2、4.4）；条目（iii）见 Shen 等人（2019年，引理 4.1）。以下关于策略参数化二阶正则性的假设，将有助于建立我们的一些结果。其他一些研究也采用了这一假设（例如，参见（Zhang 等人，2020b；Masiha 等人，2022年；Yang 等人，2021））。

假设 2. 对于每个$(s, a) \in \mathcal{S}\times\mathcal{A}$，函数 $\theta \mapsto \pi_\theta(a|s)$ 为正值且二阶连续可微，并且存在 $l_2 > 0$，使得对于任意 θ、$\theta' \in \mathbb{R}^d$ 以及任意 $s \in \mathcal{S}$、$a \in \mathcal{A}$，

$$\|\nabla^2 \log \pi_\theta(a|s) - \nabla^2 \log \pi_{\theta'}(a|s)\| \leq l_2 \|\theta - \theta'\| .$$

在假设 1 的正则性基础上，结合假设 2，期望回报 J 还将具有 Lipschitz 海森矩阵。注意，例如，高斯策略可以同时满足这两个假设。

引理 1 (Zhang 等人 (2020b))。设假设 *1*和 *2*成立。那么对于任意 θ，$\theta' \in \mathbb{R}^d$，$\|\nabla^2 J(\theta) - \nabla^2 J(\theta')\| \leq L_h\|\theta - \theta'\|$，其中 $L_h = \mathcal{O}\left((1-\gamma)^{-3}\right)$。

本文聚焦于 Fisher 非退化参数化策略这一一般类别，以进行收敛分析。为陈述我们的假设，我们将策略 π_θ 在某个 $\theta \in \mathbb{R}^d$ 和初始分布 ρ 下诱导的 Fisher 信息矩阵 $F_\rho(\theta)$ 定义如下：

$$F_\rho(\theta) \overset{\text{def}}{=} \mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}[\nabla \log \pi_\theta(a|s) \nabla \log \pi_\theta(a|s)^\top],$$

其中，$d_\rho^{\pi_\theta}(\cdot)$ 定义 $= (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\rho,\ \pi_\theta}(s_t \in \cdot)$ 是折扣状态访问度量。

假设3（Fisher 非退化策略）。存在 $\mu_F > 0$ 使得对所有 $\theta \in \mathbb{R}^d$，$F_\rho(\theta) \succcurlyeq \mu_F \cdot I_d$.

当参数化均值 $\mu_\theta(s)$ 的雅可比矩阵具有满行秩，且协方差矩阵 Σ 固定时，高斯策略 $\pi_\theta(\cdot|s) = \mathcal{N}(\mu_\theta(s), \Sigma)$ 满足假设3。更广泛的指数族参数化策略的一个子类也满足该假设（更多详情见 Ding 等人（2022年，第8节）），某些神经策略也满足该假设（另见 Liu et al.（2020年，B.2节））。这一假设在文献中很常见（Zhang et al., 2020b；Liu et al., 2020；Ding 等人，2022；Yuan 等人，2022b；Masiha 等人，2022）。我们特别指出，本文将推导的全局收敛结果未必适用于 softmax 参数化，因为当它趋近于确定性策略时，就不再满足假设3。

沿用先前的工作（Agarwal 等人，2021；Liu 等人，2020年；Ding 等人，2022年；Yuan 等人，2022b），我们通过兼容函数逼近框架（Sutton 等人，1999；Agarwal 等人，2021）引入一个假设，以反映我们的策略参数化类的表达能力。

假设 4。存在 $\varepsilon_{bias} \geq 0$，使得对每个 $\theta \in \mathbb{R}^d$，迁移[4] 误差满足：

$$\mathbb{E}[(A^{\pi_\theta}(s,a) - (1-\gamma)w^*(\theta)^\top \nabla \log \pi_\theta(a|s))^2] \leq \varepsilon_{bias},$$

其中 A^{π_θ} 是优势函数[5]，$w^*(\theta) \overset{\text{定义为}}{=} F_\rho(\theta)^\dagger \nabla J(\theta)$，其中 $F_\rho(\theta)^\dagger$ 是矩阵 $F_\rho(\theta)$ 的伪逆，期望取自 $s \sim d_\rho^{\pi^*}$，$a \sim \pi^*(\cdot|s)$，其中 π^* 是一个最优策略（最大化 $J(\theta)$）。

常见的假设 4 指出，策略参数化 π_θ 能够通过得分函数 $\nabla \log \pi_\theta$ 近似优势函数 A^{π_θ}。重要的是，我们明确指出，对于一种未涵盖所有随机策略集合的策略参数化，$\varepsilon_{\text{bias}}$ 为正；而对于丰富的神经策略，π_θ 较小（$\varepsilon_{\text{bias}}$ Wang 等人，Wang etal.，2020年）。有关假设4的更多细节，请参阅前述文献。

结合假设 3 和 4，并遵循 Ding 等人 (2022年) 的推导，我们可以得到以下放宽的弱梯度支配不等式。

引理 2（放宽的弱梯度支配，（Ding 等人，2022年））。设假设 *1*-（*i*）、*3* 和 *4* 成立。则

$$\forall \theta \in \mathbb{R}^d, \quad \varepsilon' + \|\nabla J(\theta)\| \geq \sqrt{2\mu}\,(J^* - J(\theta)),$$

where $\varepsilon' = \frac{\mu_F \sqrt{\varepsilon_{bias}}}{M_g(1-\gamma)}$ *and* $\mu = \frac{\mu_F^2}{2M_g^2}$.

这一结果是分析我们算法收敛性的关键工具。更多详情及附录 C 中的证明概要，请参阅此处。

[4]也就是说，相容函数逼近误差是在由 $d_\rho^{\pi^*}$ 和 π^*诱导的偏移分布下评估的。[5] 其定义将在附录中给出，此处仅会用到它。

4.2. N-PG-IGT 的收敛分析

我们的第一个主要结果刻画了N-PG-IGT算法的样本复杂度，该算法能够生成一个 ε-全局最优策略，但受策略参数化的近似能力所导致的偏差误差项影响。[6]

定理1（**N-PG-IGT 的全局收敛**N-PG-IGT）。设假设 *1* 至 *4* 成立。令 $\gamma_t = \frac{6M_g}{\mu_F}$ $(t+2)$ $\eta_t = \left(\frac{2}{t+2}\right)^{4/5}$，且 $H = \frac{(1-\gamma)}{}$ $^{-1}\log(T+1)$。那么，对于每个整数 $T \geq 1$，*N-PG-IGT* 的输出 θ_T 满足

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{\sigma_g + L_h}{(T+1)^{2/5}}\right) + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma},$$

其中 σ_g，L_h 分别定义于命题 *1* 和引理 *1*。因此，对于每个 $\varepsilon > 0$，输出 θ_T 满足 $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + \sqrt{\varepsilon_{bias}}/(1-\gamma)$ 对于 $T = \mathcal{O}(\varepsilon^{-2.5})$，即总样本复杂度的量级为 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$。

我们的分析N-PG-IGT 关键依赖于 $J(\theta)$的二阶光滑性，这源自引理 1。 N-PG-IGT 对于 Fisher 非退化策略，其样本复杂度比已知最佳的（随机）原始策略梯度（Yuan 等人，2022b）和 STORM- PG-F （Ding 等人，2022年）低 $\mathcal{O}(\varepsilon^{-0.5})$ 倍。近期提出的 SCRN 算法（Masiha etal.，2022年）在类似的假设条件下也能达到 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ 的样本复杂度。然而，该算法需要一个用于近似求解三次正则化子问题的子程序。此外，SCRN 需要较大的批次大小，并利用二阶信息。相比之下，N-PG-IGT 采用单循环，每次迭代只需采样一条轨迹，且实现时无需任何二阶信息。

4.3. (N)-HARPG 的收敛分析(N)-HARPG

本节进一步改进全局收敛至 $\tilde{\mathcal{O}}(\varepsilon^{-2})$ 的样本复杂度。我们分别针对 HARPG 及其归一化版本N-HARPG给出这一结果。

定理 *2*（*HARPG* 的全局收敛）。设假设 *1*、*3* 和 *4* 成立。令 $\gamma_t = \gamma_0 \eta_t^{1/2}$，$\eta_t = \frac{2}{t+2}$ *with* $\gamma_0 = min\left\{\frac{}{8\sqrt{6}}\right.$ $^{1}($ ~~$\frac{}{L_g+\sigma_g+D_h\gamma^H}$~~ $\left.\frac{\sqrt{2M_g}}{\sqrt{3\sigma_g\mu_F}}\right\}$, ~~$H = (1-\gamma)^{-1}log(T+1)$~~。则对于每个整数 $T \geq 1$，*HARPG* 的输出 θ_T（见算法 *2*）满足

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{\sigma_g + L_g + \sigma_h}{(T+1)^{1/2}}\right) + \frac{2\sqrt{\varepsilon_{bias}}}{1-\gamma},$$

其中 σ_g、σ_h、L_g 在命题 *1* 中定义，D_h 在引理 *4* 中定义（见附录 *E*）。

[6]我们还在定理 1 和 2（假设 1 和 2）下，在附录 F 的定理 5 中分析了该方法在更广泛的策略参数化类别下收敛至 FOSP 的性质。

定理 3（**N-HARPG 的全局收敛**N-HARPG）。假设 *1*、*3* 和 *4* 成立。令 $\gamma_t = \frac{6M_g}{\mu_F (t+2)}$ 以及 $\eta_t = \frac{2}{t+2}$，$H = (1-\gamma)^{-1}\log(T+1)$。那么，对于每个整数 $T \geq 1$，*N-HARPG* 的输出（见算法 *2*）满足

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{\sigma_g + L_g + \sigma_h}{(T+1)^{1/2}}\right) + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma},$$

其中 σ_g，σ_h，L_g 在命题 *1* 中定义。

推论 1. 在定理 *2*（定理 *3*）的设定下，对于每个 $\varepsilon > 0$，(N)-HARPG 的输出 θ_T 满足 $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + 2\sqrt{\varepsilon_{bias}}/(1-\gamma)$ 对于 $T = \mathcal{O}(\varepsilon^{-2})$，即总样本复杂度的阶为 $\tilde{\mathcal{O}}(\varepsilon^{-2})$。

值得注意的是，尽管 (N)-HARPG使用海森信息，我们的分析并不要求 $J(\theta)$ 满足二阶光滑性，因此不需要假设 2。这是通过在该方法的步骤 3–5 中额外采用均匀采样实现的。

5. 实验

我们在具有连续状态-动作空间的基准强化学习任务上对这些方法的性能进行了实证测试。我们使用对角高斯策略类，动作从以下分布中采样

$$a = \mu_\theta(s) + \sigma_\theta(s) \odot z, \quad z \sim \mathcal{N}(0, I),$$

对于每个 $\theta \in \mathbb{R}^d$、$s \in \mathcal{S} = \mathbb{R}^p$、$a \in \mathcal{A} = \mathbb{R}^q$，其中$\mathcal{N}(0, I)$ 是多元正态分布，$\odot$表示两个向量的逐元素乘积，$\mu_\theta(s)$ 和 $\sigma_\theta(s)$ 由全连接神经网络参数化。我们基于 garage 库（garage 贡献者，2019年）中原始策略梯度（REINFORCE）的实现来实现这些算法，并在常用的 MuJoCo 环境中测试这些方法。本节展示 Humanoid 和 Reacher 环境上的结果，其他任务的结果见附录 A。对于所有方法，每次迭代均采样 20 条轨迹，且每条轨迹的最大长度为 $H = 500$。为确保所有方法每次迭代的计算成本相同，在 (N)-HARPG 中，我们采样一半轨迹来计算随机梯度，另一半用于估计海森-向量积。为了公平比较不同方法，我们所有实验都从相同的初始策略 π_{θ_0} 开始，其中 θ_0 为随机初始化。对于所有方法，我们使用时变步长 γ_t 和动量参数 η_t，并根据各方法的理论选择这些参数，详见附录 A。

5.1. 与调优后的初始步长比较

在第一个实验中，我们比较了在 Humanoid 任务上分别调优初始步长的方法。我们为每种算法运行 13 个不同的初始步长 γ_0，其范围为 $[10^{-3}, 4]$，并选取在 平均奖励 最高的（5 次独立运行的结果按每种算法和初始步长取平均） $2 \cdot 10^7$ 次系统探测（即状态转移）后的结果。在图 1 （左）中，我们展示了中位数，以及 $1/4$ 和 $3/4$ 分位数（阴影区域）随系统探测次数变化的平均奖励。我们发现，N-PG-IGT 优于其他方法，并能迅速将奖励提升至450以上。与原始策略梯度相比，它能更好地稳定在500左右。令人惊讶的是，理论上更优的(N)-HARPG 方法在我们采用调优步长的实验中，表现始终较差。我们认为，这种现象可能是因为 (N)-HARPG 中采样的随机海森向量积在实际中的方差远高于随机梯度，这使我们无法选用更大的步长，详见注释1。

5.2. 对初始步长的鲁棒性

在本实验中，我们测试了这些方法在 Reacher 环境中对初始步长选择的鲁棒性。图1 （右）报告了达到平均奖励 超过 -11 分所需的系统探测次数，该次数取决于初始步长 γ_0。我们基于 5 次独立运行绘制中位数以及 $1/4$ 和 $3/4$ 分位数。我们观察到，对于较小的初始步长 γ_0，（方差缩减）HARPG 比原始策略梯度更快达到目标精度。相比之下，N-PG-IGT 比其他方法能容忍更大的步长，并且对初始步长的鲁棒性更强。

备注1. 此前有研究报告称，包括N-HARPG （在SHARP 中称为（*Salehkaleybar* 等人，*2022*年））在内的方差缩减方法，无需调优步长便优于原始策略梯度 （*Shen* 等人，*2019*年；*Huang* 等人，*2020*年）。对于较小且未经调优的步长，我们也观察到类似现象，见附录*A*中的图*4*。

6. 结论

在本研究中，我们改进了现有方法针对 Fisher非退化参数化策略的全局最优性样本复杂度 $\tilde{\mathcal{O}}(\varepsilon^{-3})$。我们提出了两种基于动量的单循环策略梯度方法，分别实现了 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ 和 $\tilde{\mathcal{O}}(\varepsilon^{-2})$ 样本复杂度，且每次迭代仅使用一条或两条轨迹。值得注意的是，与大多数先前研究不同，我们的分析不依赖于重要性采样权重方差有界这一不切实际的假设。为达到最佳的 $\tilde{\mathcal{O}}(\varepsilon^{-2})$ 样本复杂度，我们的分析依赖于将结构性弱梯度支配条件与方差缩减策略梯度的设计巧妙结合。未来的研究方向包括放宽策略正则性条件（Zhang 等人，2022年），并研究在非紧空间和重尾策略参数化条件下，用于连续控制的更高效策略搜索方法的收敛性，以提升探索能力；

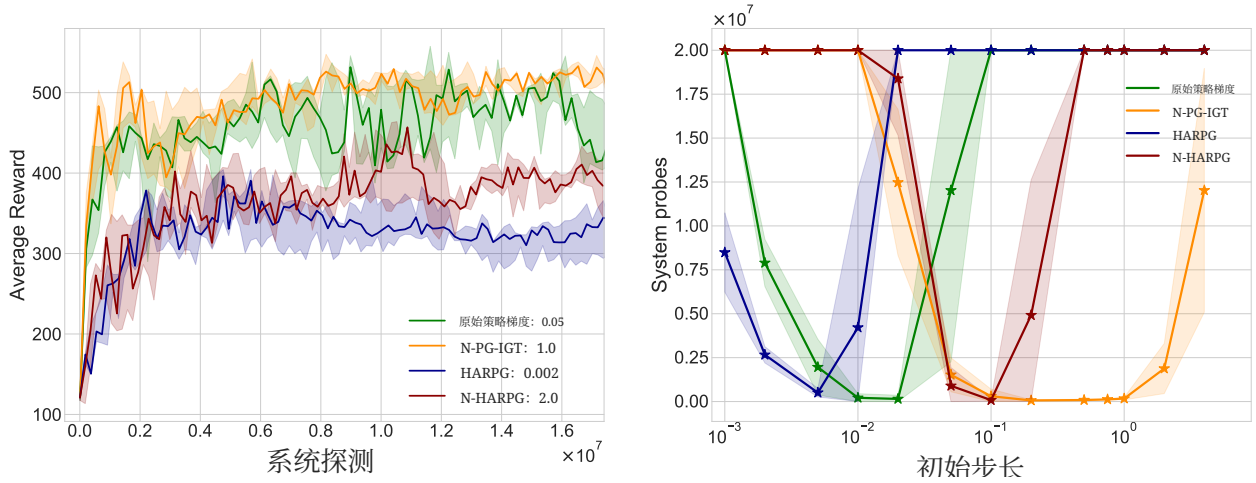

图1：左：原始策略梯度、N-PG-IGT、HARPG、N-HARPG 在 Humanoid 环境中的表现。右：所提方法在 Reacher 环境中对初始步长的鲁棒性。

鉴于近期在这一方向上的研究成果（Bedi 等人，2021），进一步提升探索能力。

致谢

本研究获得了 ETH AI 中心博士奖学金、ETH 数据科学基金会（ETH-FDS）以及由苏黎世联邦理工学院基金会资助的 ETH 研究资助的支持。

参考文献

AlekhAgarwal、Sham M. Kakade、Jason D. Lee 和 Gaurav Mahajan。《策略梯度方法理论：最优性、近似与分布偏移》。机器学习研究期刊，22(98):1–76，2021。

Carlo Alfano 和 Patrick Rebeschini。对数线性策略参数化下自然策略梯度的线性收敛。*arXiv*预印本 *arXiv:2209.15382*，2022年。

Yossi Arjevani、Yair Carmon、John C Duchi、Dylan J Foster、Ayush Sekhari 和 Karthik Sridharan。非凸随机优化中的二阶信息：优势与局限。发表于学习理论会议，第 242–299 页。PMLR，2020年。

S´ebastien Arnold、Pierre-Antoine Manzagol、Reza Ba-banezhad Harikandeh、Ioannis Mitliagkas 和 Nicolas Le Roux。通过传输过去的梯度降低在线优化中的方差。神经信息处理系统进展，32，2019年。

Hedy Attouch 和 J´erˆome Bolte。关于涉及解析特征的非光滑函数的近端算法收敛性。数学规划，116(1): 5–16，2009年。

Amrit Singh Bedi、Anjaly Parayil、Junyu Zhang、Mengdi Wang 和 Alec Koppel。关于连续控制中重尾策略搜索的样本复杂度与亚稳态。*arXiv* 预印本 *arXiv:2106.08414*，2021。

Amrit Singh Bedi、Souradip Chakraborty、Anjaly Parayil、Brian M Sadler、Pratap Tokekar 和 Alec Koppel。连续动作空间中策略镜像上升的隐含偏差。载于第*39*届国际机器学习会议论文集，第162卷，机器学习研究论文集，第1716–1731页。PMLR，2022年7月17–23日。

Jalaj Bhandari 和 Daniel Russo。策略梯度方法的全局最优性保证。*arXiv* 预印本 *arXiv:1906.01786*，2019年。

Jalaj Bhandari 和 Daniel Russo。有限马尔可夫决策过程策略梯度方法的线性收敛。载于人工智能与统计学国际会议，第 2386–2394 页。PMLR，2021。

Zaiwei Chen 和 Siva Theja Maguluri。离策略采样与线性函数逼近下基于策略方法的样本复杂度。载于人工智能与统计学国际会议，第 11195–11214 页。PMLR，2022年。

ZaiweiChen、Sajad Khodadadian 和 SivaThejaMaguluri。线性函数逼近下离策略自然演员–评论家的有限样本分析。*IEEE* 控制系统快报，6：2611–2616，2022年。

Ashok Cutkosky 和 Harsh Mehta。动量改进了归一化随机梯度下降。载于第 *37* 届国际机器学习会议论文集，第 119 卷

的 机器学习研究论文集，第 2260–2268 页。PMLR，2020年。

Ashok Cutkosky 和 Francesco Orabona。非凸随机梯度下降中的基于动量的方差缩减。神经信息处理系统进展，32，2019年。

Anirban DasGupta 和 Anirban DasGupta。指数族与统计应用。统计学与机器学习概率论：基础与进阶专题，第 583–612 页，2011。

Yuhao Ding、Junzi Zhang 和 Javad Lavaei。超越精确梯度：带熵正则化的随机 soft-max 策略梯度方法的收敛。*arXiv* 预印本 *arXiv:2110.10117*，2021。

Yuhao Ding、Junzi Zhang 和 Javad Lavaei。基于动量的策略梯度的全局最优收敛。发表于 国际人工智能与统计会议，第 1910–1934 页。PMLR，2022年。Cong Fang、Chris Junchi Li、Zhouchen Lin 和 Tong Zhang。SPIDER：通过随机路径积分微分估计器实现近最优非凸优化。发表于神经信息处理系统进展，第 31 卷，2018年。Ilyas Fatkhullin 和 Boris Polyak。优化静态线性反馈：梯度方法。*SIAM*控制与优化杂志，59(5)：3887–3911，2021。Ilyas Fatkhullin、Jalal Etesami、Niao He 和 Negar Kiyavash。全局 Kurdyka–Łojasiewicz 不等式下随机优化的 SHARP 分析。神经信息处理系统进展，2022年。Maryam Fazel、Rong Ge、Sham Kakade 和 Mehran Mesbahi。线性二次调节器策略梯度方法的全局收敛。发表于 国际机器学习会议，第 1467–1476 页。PMLR，2018年。Xavier Fontaine、Valentin De Bortoli 和 Alain Durmus。随机梯度下降及其连续时间对应形式的收敛速率与逼近结果。发表于 学习理论会议，第 1965–2058 页。PMLR，2021。S´ebastien Gadat、Fabien Panloup 和 Sofiane Saadane。随机重球法。电子统计学杂志，12(1)：461–529，2018年。Garage贡献者。Garage：用于可复现强化学习研究的工具包。https://github.com/rlworkgroup/garage，2019年。

Matilde Gargiani、Andrea Zanelli、Andrea Martinelli、Tyler Summers 和 John Lygeros。PAGE-PG：一种简单的采用概率梯度估计的无环方差缩减策略梯度方法。载于第*39*届国际机器学习会议论文集，机器学习研究论文集第162卷，页码7223–7240。PMLR，2022年。

Saeed Ghadimi 和 Guanghui Lan。用于非凸随机规划的随机一阶和零阶方法。*SIAM* 优化期刊，23(4):2341–2368，2013年。

Elad Hazan、Kfir Y. Levy 和 ShaiShalev-Shwartz。超越凸性：随机拟凸优化。*arXiv* 预印本 *arXiv:1507.02030*，2015。

Mingyi Hong、Hoi-To Wai、Zhaoran Wang 和 Zhuoran Yang。双层优化的双时间尺度框架：复杂度分析及其在演员-评论家中的应用。即将发表于 *SIAM* 优化期刊，*2022*年，*arXiv* 预印本 *arXiv:2007.05170*，2020年。Feihu Huang、Shangqian Gao、Jian Pei 和 Heng Huang。基于动量的策略梯度方法。发表于国际机器学习会议，第 4422–4433 页。PMLR，2020年。Sajad Khodadadian、Thinh T Doan、Justin Romberg 和 Siva Theja Maguluri。自然演员 - 评论家算法双时间尺度框架的有限样本分析。*IEEE* 自动控制汇刊，2022a。Sajad Khodadadian、Prakirt Raj Jhunjhunwala、Sushil Ma- havir Varma 和 Siva Theja Maguluri。自然策略梯度算法的线性和超线性收敛。系统与控制快报，164:105214，2022b。ISSN 0167-6911。Solomon Kullback。信息论与统计学。Courier Corporation，1997年。Krzysztof Kurdyka。关于 o-极小结构中可定义函数的梯度。发表于傅里叶研究所年刊，第 48 卷，第 769–783 页，1998年。Guanghui Lan。用于强化学习的策略镜像下降：线性收敛、新的采样复杂度及广义问题类别。数学规划，第 1–48 页，2022年。Zhize Li、Hongyan Bao、Xiangliang Zhang 和 Peter Richtarik。PAGE：一种用于非凸优化的简单且最优的概率梯度估计器。发表于第 *38* 届国际机器学习会议论文集，第 139 卷，机器学习研究论文集，第 6286–6295 页。PMLR，2021。

YanliLiu、Kaiqing Zhang、Tamer Basar 和 Wotao Yin。关于（方差缩减）策略梯度

与自然策略梯度方法的改进分析。神经信息处理系统进展，33:7624–7636，2020年。

Stanislaw Lojasiewicz。实解析子集的一个拓扑性质。《 ´偏微分方 ´程 ´》，117:87–89，1963年。

Saeed Masiha、Saber Salehkaleybar、Niao He、Negar Kiyavash 和 Patrick Thiran。随机二阶方法在梯度主导函数上可证明优于随机梯度下降。即将在《神经信息处理系统进展》上发表，*arXiv* 预印本 *arXiv:2205.12856v1*，2022年。

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。载于《第*37*届国际机器学习大会论文集》，第119卷《机器学习研究论文集》，第6820–6829页。PMLR，2020年。

Jincheng Mei、Bo Dai、Chenjun Xiao、Csaba Szepesvari 和 Dale Schuurmans。理解随机性对策略优化的影响。神经信息处理系统进展，34:19339–19351，2021a。

Jincheng Mei、YueGao、Bo Dai、Csaba Szepesvari 和 Dale Schuurmans。利用一阶非凸优化中的非均匀性。载于第 *38* 届国际机器学习会议论文集，第 139 卷，机器学习研究论文集，第 7555–7564 页。PMLR，2021b。

Francisco S Melo、Sean P Meyn 和 M Isabel Ribeiro。函数近似下强化学习的分析。载于第 *25* 届国际机器学习会议论文集，第 664–671 页，2008 年。

Lam M Nguyen、Jie Liu、Katya Scheinberg 和 Martin Tak´aˇc。Sarah：一种利用随机递归梯度解决机器学习问题的新方法。载于 国际机器学习会议，第 2613–2621 页。PMLR，2017 年。

Matteo Papini、Damiano Binaghi、Giuseppe Canonaco、Mat- teo Pirotta 和 Marcello Restelli。随机方差缩减策略梯度。载于国际机器学习会议，第 4026–4035 页。PMLR，2018年。

Nhan Pham、Lam Nguyen、Dzung Phan、Phuong Ha Nguyen、Marten Dijk 和 Quoc Tran-Dinh。用于强化学习的混合随机策略梯度算法。载于国际人工智能与统计会议，第 374–385 页。PMLR，2020年。

Boris Teodorovich Polyak。泛函最小化的梯度方法。计算数学与数学物理杂志，3(4):643–653，1963年。

Martin L Puterman。马尔可夫决策过程：离散随机动态规划。John Wiley & Sons，2014。Shuang Qiu、Zhuoran Yang、Jieping Ye 和 Zhaoran Wang。关于演员-评论家算法的有限时间收敛。*IEEE* 信息论选定领域期刊，2(2):652–664，2021。Saber Salehkaleybar、Sadegh Khorasani、Negar Kiyavash、Niao He 和 Patrick Thiran。结合二阶信息的自适应基于动量的策略梯度。*arXiv* 预印本 *arXiv:2205.08253*，2022年。

Kevin Scaman、Cedric Malherbe 和 Ludovic Dos Santos。在通用洛雅谢维奇条件与局部光滑性下非凸随机梯度下降的收敛速率。载于第*39*届国际机器学习会议论文集，第 19310–19327 页。PMLR，2022年6月。

John Schulman、Sergey Levine、Pieter Abbeel、Michael Jor- dan 和 Philipp Moritz。信赖域策略优化。载于国际机器学习会议，第 1889–1897 页。PMLR，2015。

John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Rad- ford 和 Oleg Klimov。近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017。

Zebang Shen、Alejandro Ribeiro、Hamed Hassani、Hui Qian 和 Chao Mi。海森矩阵辅助策略梯度。发表于国际机器学习会议，第 5729–5738 页。PMLR，2019。

David Silver、Guy Lever、Nicolas Heess、Thomas Degris、Daan Wierstra 和 Martin Riedmiller。确定性策略梯度算法。发表于国际机器学习会议，第 387–395 页。PMLR，2014。

Sebastian U. Stich。（随机）梯度方法的统一最优分析。*arXiv* 预印本 *arXiv:1907.04232v2*，2019年。

Richard S. Sutton、David McAllester、Satinder Singh 和 Yishay Mansour。使用函数逼近进行强化学习的策略梯度方法。载于神经信息处理系统进展，第 12 卷。MIT 出版社，1999 年。

Hoang Tran 和 Ashok Cutkosky。利用二阶动量改进随机梯度下降。*arXiv* 预印本 *arXiv:2103.03265*，2021 年。

J.N. Tsitsiklis 和 B. Van Roy。《使用函数逼近的时序差分学习分析》。*IEEE* 自动控制汇刊，42(5):674–690，1997年。数字对象唯一标识符：10.1109/9.580874。

Lingxiao Wang、Qi Cai、Zhuoran Yang 和 Zhaoran Wang。《神经策略梯度方法：全局最优性与收敛速率》。载于国际学习表征会议，2020年。

Ronald J Williams. 面向连接主义强化学习的简单统计梯度跟随算法. *Machine Learning*, 8:229–256, 1992年。

Yue Frank Wu、Weitong Zhang、Pan Xu 和 Quanquan Gu. 双时间尺度演员-评论家方法的有限时间分析. 神经信息处理系统进展, 33:17617–17628, 2020年。

Lin Xiao. 策略梯度方法的收敛速率. *Journal of Machine LearningResearch*, 23 (282):1–36, 2022年。

Pan Xu、Felicia Gao 和 Quanquan Gu. 采用递归方差缩减的样本高效策略梯度方法. 载于国际学习表征会议, 2020a。

Pan Xu、Felicia Gao 和 Quanquan Gu. 随机方差缩减策略梯度的改进收敛分析. 载于人工智能中的不确定性, 第541–551页. PMLR, 2020b。

Tengyu Xu、Zhe Wang 和 Yingbin Liang。改进（自然）演员-评论家算法的样本复杂度界。在神经信息处理系统进展，第 33 卷，第 4358–4369 页，2020c。

Long Yang、QianZheng 和 Gang Pan。策略梯度寻找二阶驻点的样本复杂度。人工智能会议论文集, 35(12):10630–10638，2021。

Huizhuo Yuan、Xiangru Lian、Ji Liu 和 Yuren Zhou。策略梯度方法的随机递归动量，2020。

Rui Yuan、Simon S Du、Robert M Gower、Alessandro Lazaric 和 Lin Xiao。对数线性策略下自然策略梯度方法的线性收敛。*arXiv* 预印本 *arXiv:2210.01400*，2022a。

Rui Yuan、Robert M. Gower 和 Alessandro Lazaric。普通策略梯度的一般样本复杂度分析。在第 *25* 届人工智能与统计学国际会议论文集，《机器学习研究论文集》第 151 卷，第 3332–3380 页。PMLR, 2022b。

Junyu Zhang、Alec Koppel、Amrit Singh Bedi、Csaba Szepesvari 和 Mengdi Wang。用于具有一般效用的强化学习的变分策略梯度方法。神经信息处理系统进展，33：4572–4583，2020a。

Junyu Zhang、Chengzhuo Ni、Csaba Szepesvari、Mengdi Wang 等人。论方差缩减策略梯度方法的收敛性与样本效率。神经信息处理系统进展，34：2228–2240，2021a。

Junzi Zhang、Jongho Kim、Brendan O'Donoghue 和 Stephen Boyd。使用 REINFORCE 实现样本高效的强化学习。*AAAI*人工智能会议论文集，35(12):10887–10895，2021b。

Kaiqing Zhang、Alec Koppel、Hao Zhu 和 Tamer Basar。策略梯度方法对（几乎）局部最优策略的全局收敛。*SIAM*控制与优化杂志，58(6):3586–3612，2020b。

Matthew S Zhang、Murat A Erdogdu 和 Animesh Garg。《弱光滑情形下策略梯度方法的收敛性与最优性》。载于*AAAI*人工智能会议论文集，第36卷，第9066–9073页，2022年。

目录

1 引言 **1**

1.1 贡献总结 .21.2 相关工作 .2

2 预备知识 **3**

2.1 问题表述 .32.2 一阶平稳性与全局最优性 .42.3 策略梯度与海森矩阵 .4

3 归一化动量策略梯度算法 **4**

3.1 带隐式梯度传输的归一化策略梯度 .53.2 海森矩阵辅助递归策略梯度 .5

4 全局收敛分析 **6**

4.1 假设.64.2 N-PG-IGT 的收敛分析.74.3 (N)-HARPG 的收敛分析.7

5 实验 **8**

5.1 与经过调优的初始步长比较 .85.2 对初始步长的稳健性 .8

6 结论 8

A 附加实验 与实现细节 **15**

A.1 实验1：与经过调优的初始步长比较 .15A.2 实验2：对初始步长的稳健性 .16A.3 实验3：比较原始策略梯度 和 HARPG 在使用较小且未经调优的步长时的表现 .16A.4 实验 4：与其他经过调优的步长方法进行额外比较 .16A.5 实验5：离散状态-动作空间下 soft-max 参数化的性能 17

B 关于第4.1节假设的进一步讨论 18

B.1 高斯策略 .18B.2 柯西策略 .19

C 主要结果的证明概要 **21**

D 进一步的相关工作 **22**

D.1 方差缩减策略梯度方法的一阶驻点性 . 22

D.2 精确策略梯度方法的全局最优性 .22D.3 关于随机策略优化的进一步讨论 .22D.4 与随机优化领域既有工作的比较 .23

E 符号说明与实用引理 24F 定理 1 的证明（N-PG-IGT）25F.1 全局收敛 .25F.2 收敛至一阶驻点 .29**G 定理 2 的证明（HARPG）31H 定理 3 的证明（N-HARPG）36I 归一化动量策略梯度方法的收敛分析（N-MPG）38J 技术引理 41**J.1 求解递推关系的引理 .41J.2 递减步长估计引理 .42

附录

A. 补充实验与实现细节

作为对第 5 节所呈现实证结果的补充，本节将介绍一组关于 MuJoCo 环境的额外实验，并提供更多实现细节。

实现细节。 参数 $\mu_\theta(s)$ 和 $\sigma_\theta(s)$ 高斯策略由一个包含两个隐藏层的前馈神经网络参数化，每层大小为 64，采用双曲正切 激活函数。各环境的参数汇总见表 3。

环境	Walker2d	Hopper	Halfcheetah	Reacher	Humanoid	Cartpole
时域	500	1000	500	500	500	200
时间步数	2×10^7	2×10^7	2×10^7	2×10^7	2×10^7	2×10^6
批量大小	20	10	10	200	10	5
最佳 γ_0 适用于原始策略梯度	0.02	0.02	0.02	0.02	0.05	0.1
最佳 γ_0 适用于N-PG-IGT	0.75	0.5	1.0	0.2	1.0	0.2
最佳 γ_0 适用于HARPG	0.002	0.002	0.001	0.002	0.002	0.02
最佳 γ_0 适用于NHARPG	0.1	0.05	0.2	0.1	2.0	0.1

表3：超参数和步长选择。初始步长从13个值组成的集合中选取，以最后一次迭代中的最佳表现为准。

A.1. 实验1：与调优后的初始步长进行比较

我们进一步测试了使用最佳调优步长的算法，测试环境包括`Walker2d` 和`Reacher`。与第5节中的实验类似，我们针对每个初始步长对每种算法分别运行五次 γ_0 从集合中

$$\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.75, 1, 2, 4\}. \tag{4}$$

然后，根据算法在上一次迭代中的表现来选择所报告的步长。在表4中，我们报告时变步长和动量参数的选择。关于表中所列参数选择的依据，请参阅第F、G 和H节。

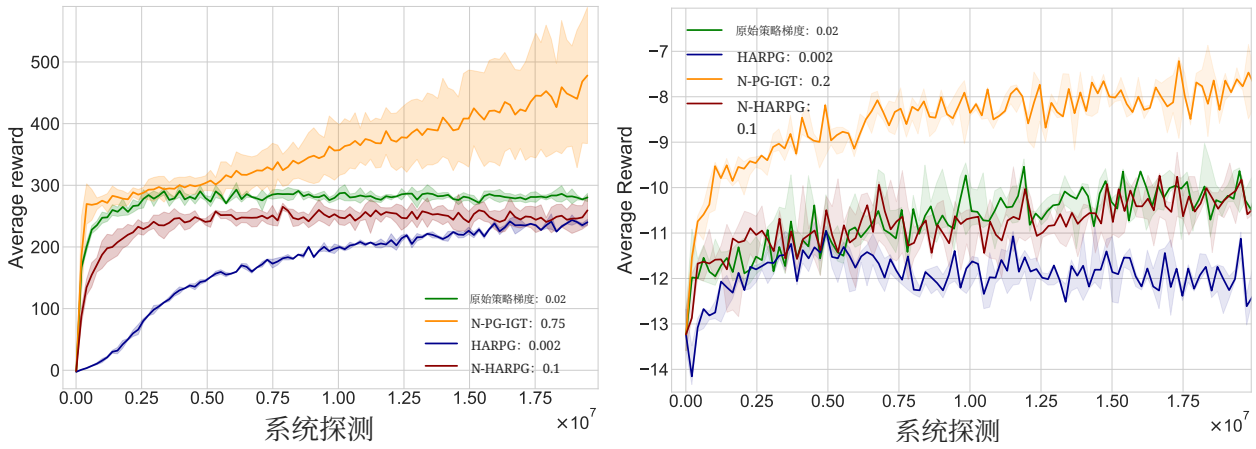

图2：原始策略梯度、N-PG-IGT、HARPG和N-HARPG 在 Walker2d（左）和 Reacher（右）环境中的表现。

与`Humanoid` 环境第5节中介绍的环境类似，我们观察到N-PG-IGT 优于其他方法，并迅速将奖励提高到450 以上（对于`Walker2d`），对于 -9 以上（对于`Reacher`），见图2。遗憾的是，

理论上更优的HARPG 和N-HARPG 方法在实践中并未表现出相较于原始策略梯度的改进，这可能是由于海森矩阵向量估计器的方差过大。

	原始策略梯度	N-PG-IGT	HARPG	N-HARPG
步长，γ_t	$\gamma_0 \left(\frac{2}{t+2}\right)^{2/3}$	$\frac{2\gamma_0}{t+2}$	$\gamma_0 \left(\frac{2}{t+2}\right)^{1/2}$	$\frac{2\gamma_0}{t+2}$
动量参数，η_t	N/A	$\left(\frac{2}{t+2}\right)^{4/5}$	$\frac{2}{t+2}$	$\frac{2}{t+2}$

表4：策略梯度方法的步长序列和动量参数汇总。初始步长参数 γ_0 在集合（4）内进行调优。各算法和环境对应的最佳值 γ_0 见表 3。

A.2. 实验2：对初始步长的鲁棒性

除了在 Reacher 环境上进行的实验（见图1）外，我们还报告了所研究方法在另外两个任务 Walker2d 和 Humanoid 上对初始步长的敏感性。对于这些任务，我们观察到类似的趋势：在小步长范围内，HARPG 可能比原始策略梯度更快，见图3。另一方面，归一化方法能够容忍更大的步长，尤其是N-PG-IGT，表现出对步长选择更强的鲁棒性。总体而言，我们观察到，归一化方法和非归一化方法的最佳步长范围自然有所不同，因此，为了比较这两类算法，有必要对步长进行广泛调优。另一个观察结果是，对于所测试的环境，方差缩减 HARPG 方法只有在较小且未经调优的步长下才优于原始策略梯度 。因此，在下一个实验中，我们将在小步长范围内比较这些方法。

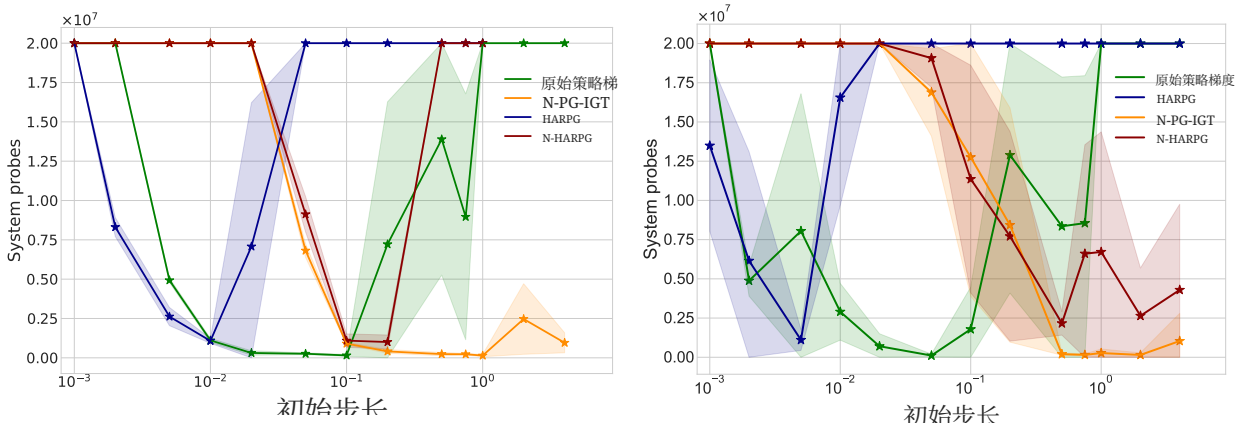

图 3：Walker2d（左）和 Humanoid（右）环境中对初始步长的鲁棒性。

A.3. 实验3：采用较小且未经调优的步长比较 原始策略梯度和HARPG

在本实验中，我们展示了方差缩减 HARPG方法在固定的较小初始步长下，能够胜过 原始策略梯度 γ_0，见图 4。[7] 我们在此观察到的结果与先前关于其他方差缩减策略梯度算法的研究报告一致，例如（Shen 等人，2019年；Huang 等人，2020年；Salehkaleybar 等人，2022年）。然而，当步长经过调优后，原始策略梯度 的表现始终优于方差缩减算法；此外，实验 4（图5）还比较了其他采用调优步长的方法。

A.4. 实验 4：与其他采用调优步长的方法进行补充比较

在本实验中，我们在`Hopper` 和 `Halfcheetah` 任务上测试采用调优步长的算法。我们将所介绍方法的表现与另外两种算法进行了比较。 VR-PG 方法也称为 STORM-PG（Yuan 等人，

[7]我们未纳入 norma 化算法未纳入此次比较，因为归一化会对步长产生额外的缩放。

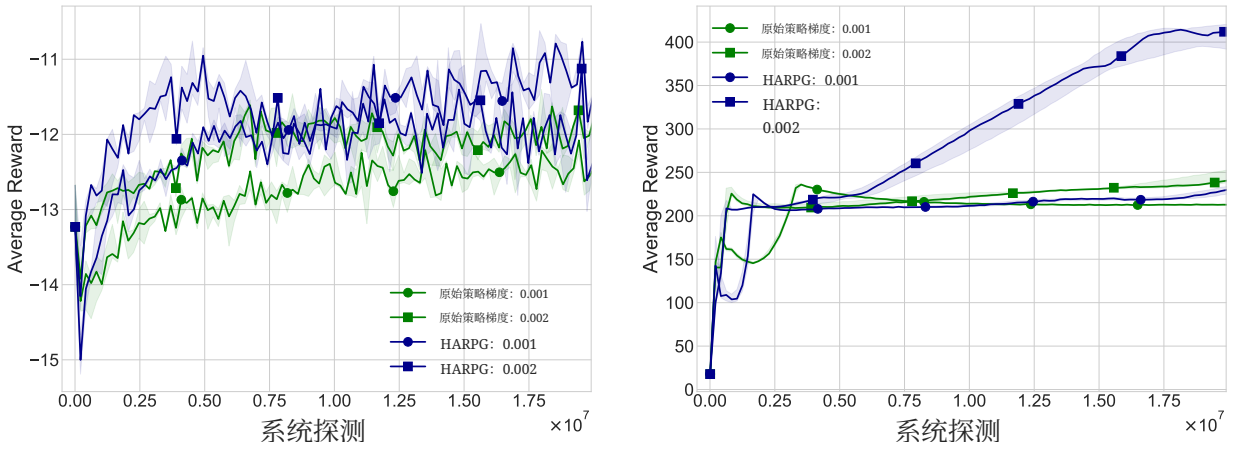

图 4：在初始步长未调优的情况下，原始策略梯度和HARPG 在 Reacher（左）和 Hopper（右）环境中的性能。

2020年）采用重要性采样（IS）权重实现方差缩减策略梯度方法，而非采用海森修正。为公平起见，我们还考虑了该方法的归一化版本，并将其与 N-HARPG进行比较。 [8] 我们从图5中观察到，尽管VR-PG 和N-VR-PG 的表现与HARPG 和N-HARPG相近或更好，但我们的方法N-PG-IGT仍然收敛得更快。

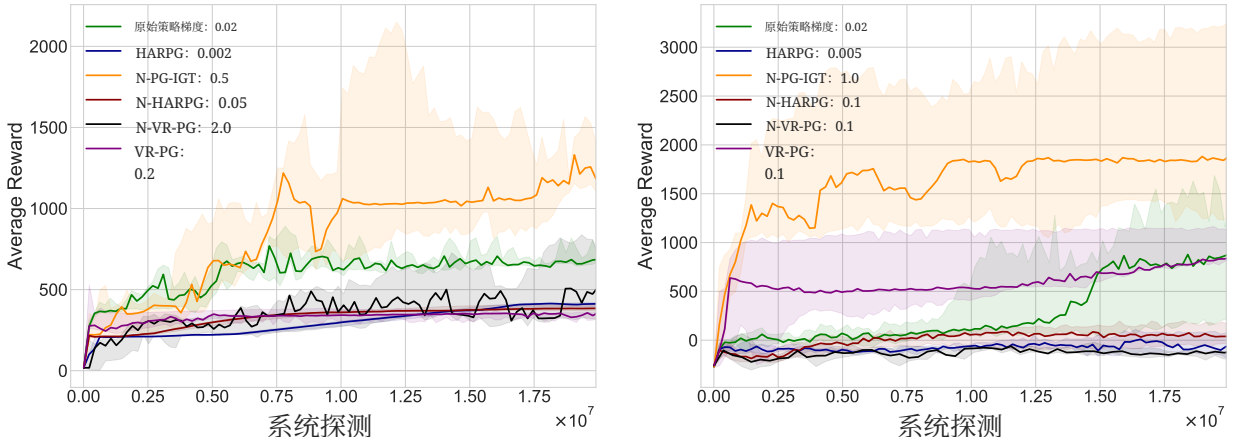

图 5：原始策略梯度、N-PG-IGT、HARPG、N-HARPG、VR-PG 和N-VR-PG 在 Hopper（左）和 Halfcheetah（右）环境中的性能。

A.5. 实验 5：离散状态-动作空间下 soft-max参数化的性能

尽管我们的主要理论结果（除定理 5中的 N-PG-IGT 的 FOSP 外）仅在费希尔非退化（FND）参数化假设 3下成立，但本文介绍的所有方法都无需任何修改即可应用于其他策略类别。在本实验中，我们在采用 Softmax策略参数化的热门 `Cartpole` 环境中测试所提出的算法。我们从图 6 中观察到，这些方法的表现与在其他环境中的表现相似。

[8] 我们注意到 这些方法的理论分析需要额外假设 IS 权重有界 ts.

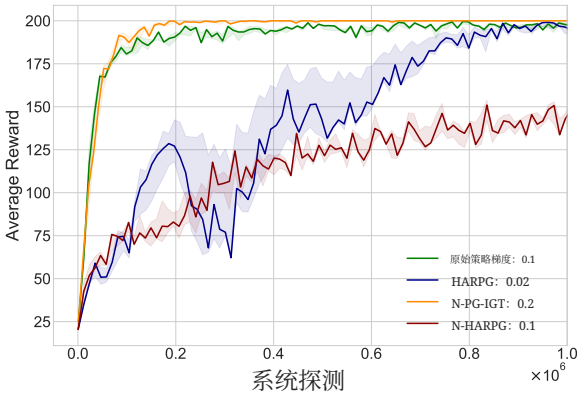

图6：在经过调优的初始步长下，原始策略梯度、N-PG-IGT、HARPG和N-HARPG 在 Cartpole 环境中的表现。

B. 关于第 4.1 节中假设的进一步讨论

本节将通过两个简单示例讨论我们的假设，同时也会说明这些假设在这两个示例之外的普适性。不过，我们希望强调的是，我们的假设集合具有一般性，并且是近期关于策略梯度方法收敛性研究中的标准假设（参见 Xu 等人 (2020a)；Zhang 等人 (2020b)；Liu 等人 (2020年)；Ding 等人 (2022年)；Gargiani 等人 (2022年)；Masiha 等人(2022年)；Salehkaleybar 等人 (2022年)等）。我们认为，放宽部分假设是一个有趣的问题，将留待未来研究。

B.1. 高斯策略

令 $\varphi: \mathcal{S} \to \mathbb{R}^d$ 为一个特征映射，并设[9] $\mathcal{A} = \mathbb{R}$。我们考虑一种常见的高斯策略，它采用线性均值参数化和常数方差参数 $\sigma^2 > 0$，该参数对每个 $\theta \in \mathbb{R}^d, (s,a) \in \mathcal{S} \times \mathcal{A}$ 定义如下：

$$\pi_\theta(a|s) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(a - \varphi(s)^\top \theta)^2}{2\sigma^2}\right). \tag{5}$$

更一般的高斯策略会使用神经网络进行均值和标准差参数化。请注意，还可以考虑更一般的参数化形式，例如指数族或对称 α-稳定策略，其中高斯策略是一个特例。感兴趣的读者可参阅（Bedi 等人，2021）中关于此类重尾策略参数化的精彩介绍。对于（5），得分函数对每个 $\theta \in \mathbb{R}^d, (s,a) \in \mathcal{S} \times \mathcal{A}$ 可写为：

$$\nabla \log \pi_\theta(s,a) = \frac{a - \varphi(s)^\top \theta}{\sigma^2} \varphi(s).$$

假设 1 -（i）（有界得分函数）：关于这一假设，有几点需要说明 ：

1. 对于高斯策略这一特定情形，我们注意到，严格来说，即使特征映射有界（正如文献中先前指出的那样（Yuan 等人，2022b；Bedi 等人，2022年）），对于所有 $a \in \mathbb{R}$和 $\theta \in \mathbb{R}^d$，这一假设也并不成立。然而，近期有一些研究（Papini 等人，2018年，C 节）、（Xu 等人，2020b，第 7 节）、（Liu 等人，2020年，第 4 节）、（Xu 等人，2020a，推论 4.8）假设得分函数有界。显然，如果进一步假设采样动作和均值参数化 $\varphi(s)^\top\theta$有界，就能保证这一假设成立。实践中，通常通过对策略选出的动作进行裁剪来满足这一条件 π_θ，并将梯度方法的更新投影到固定半径球内。

2. 此前已有研究表明，这一标准假设可以放宽为期望意义上的假设（见假设 4.1（Yuan 等人，2022b）），而不必对所有状态-动作对都成立。这种放宽后的版本允许在无需额外假设的情况下采用高斯

[9]更一般的情形 $\mathcal{A} = \mathbb{R}^q (q > 1)$ 也可以用类似的方法处理，只需对高斯策略的定义稍作调整，例如采用维度适当的矩阵特征映射（而非向量特征映射）。在下面讨论假设 3 时，我们会简要评论这种情形。

参数化，同时只需验证 J 的光滑性以及随机梯度的有界方差，见（Yuan 等人，2022b，引理 4.2 和 4.4）。然而，本文不关注这类技术性放宽，而是采用有界得分函数这一标准假设。

3. 对于得分函数无界的策略，一种替代方法是根据得分函数有界且可积的情况（例如，利用探索参数）对动作空间进行划分。请参阅Bedi 等人 (2021) 的第 5.1 节。采用这种方法需要调整分析过程。

假设 1-(ii)(得分函数的利普希茨连续性)：我们计算 ∇^2 对数 $\pi_\theta(s,a) = -\frac{1}{\sigma^2}\varphi(s)\varphi(s)^\top$，当特征映射有界时，该式也有界。

假设 2(得分函数的二阶光滑性)：由于 ∇^2 对数 $\pi_\theta(s,a)$ 与 θ 无关，因此该假设显然成立。关于这一假设，我们还要强调，(N)-HARPG并不需要该假设（见定理 2和3）。只有 N-PG-IGT需要该假设。尽管该假设对于 N-PG-IGT的改进至关重要，但我们指出：(a) 正如刚才所述，这是一个较弱的假设，高斯策略满足该假设；(b) 表 2 (“无二阶光滑性”列) 将其列为相较于现有结果的一项额外假设，以便公平比较。具体而言，我们将结果与作出同一额外假设的（Masiha 等人，2022年）进行比较。

假设3（费舍尔非退化性）：高斯策略的费舍尔信息矩阵可写为 $F_\rho(\theta) = \frac{1}{\sigma^2}\varphi(s)\varphi(s)^\top$，对于满行秩特征映射，该矩阵为正定矩阵。更一般地[10]，

1. 线性均值参数化：当高斯策略为多元策略（$\mathcal{A}=\mathbb{R}^q$），且参数化形式为 $\mu_\theta(s) = \Phi(s)^T\theta$，其中 $\Phi: \mathcal{S}\to\mathbb{R}^{d\times q}$ (d 为参数 θ的维度），且 Σ 为固定协方差矩阵时，费舍尔信息矩阵与 θ 无关，因为它由 $\Phi(s)\Sigma^{-1}\Phi(s)^T$. 给出。因此，若 $\Phi(s)$ 为满行秩，假设3便成立。对于通常的情形，即参数 θ 的维度小于动作空间的维度（即 $d<q$），令 $\Phi(s)$的各行线性无关，即可满足这一满秩要求。在强化学习文献中，考虑线性函数逼近时通常会作出这一假设（例如参见 Tsitsiklis 和 Van Roy（1997）；Melo 等人（2008））。

2. **神经网络均值参数化**：当 $\mu_\theta(s)$ 关于 θ 是非线性的（如著名的神经网络情形），在对 3 的雅可比矩阵 $\mu_\theta(s)$ 满足适当（关于 θ 一致）的假设时，便可满足假设 3；这些条件与线性均值参数化的条件类似。

3. 满秩指数族分布（DasGupta 和 DasGupta，2011年）若具有参数化均值 $\mu_\theta(s)$，也满足假设 3。值得一提的是，许多正则统计模型的费舍尔信息矩阵都是正定的（Kullback，1997年）。

B.2. 柯西策略

令 $\varphi: \mathcal{S}\to\mathbb{R}^d$ 为特征映射，并设定 $\mathcal{A}=\mathbb{R}$。我们假设该特征映射有界（即，对于每个 $s\in\mathcal{S}$，都有 $\|\phi(s)\|\le D$，其中某个正常数为 D），且具有满行秩。采用线性均值参数化$\varphi(s)^\top\theta$、尺度参数 $\sigma>0$ 的柯西分布定义如下：

$$\pi_\theta(a|s) = \frac{1}{\pi\sigma\left(1+\left(\frac{a-\varphi(s)^\top\theta}{\sigma}\right)^2\right)},$$

对每个 $\theta\in\mathbb{R}^d$，$s\in\mathcal{S}$，$a\in\mathcal{A}$.，在这种情况下，得分函数可以写为

$$\nabla\log\pi_\theta(a|s) = \frac{2x_\theta}{1+x_\theta^2}\frac{\varphi(s)}{\sigma},$$

其中 $x_\theta = \left(a-\varphi(s)^\top\theta\right)/\sigma$。

根据（Bedi 等人，2022年，引理4.7），我们在这种策略参数化下验证这些假设。

[10]为完整起见，也为方便读者，我们在此转述 Liu 等人（2020年）附录 B.2 讨论中的一些评论。

假设 1-(i)(有界得分函数)： 由于 $\frac{2x}{1+x^2} \leq 1$ 对所有 ~~$x \in$~~ $\mathbb{R}$，我们有 $\|\nabla$对数 $\pi_\theta(a|s)\| \leq M_g$，且$M_g = \frac{D}{\sigma}$ 对所有 $\theta \in \mathbb{R}^d$ 以及 $s \in \mathcal{S}$， $a \in \mathcal{A}$。

假设 1-(ii)(得分函数的利普希茨连续性)： 我们计算 $\nabla^2 \log \pi_\theta(a|s) = \frac{2}{(1-x_\theta^2)(1+x_\theta^2)^2} \begin{pmatrix} \varphi(s)\varphi(s)^\top \\ \sigma^2 \end{pmatrix} \cdot$因此，由于$\frac{2(1-x^2)}{(1+x^2)^2} \leq \frac{3}{2}$对所有 $x \in \mathbb{R}$ 均成立，我们可以用 $M_h = \frac{3D^2}{2\sigma^2}$ 对 $\left\|\nabla^2 \log \pi_\theta(a|s)\right\| \leq M_h$ 进行界定。

假设2(得分函数的二阶光滑性)： 我们任取一个非零向量 $u \in \mathbb{R}^d$，并计算$\nabla(\nabla^2 \log \pi_\theta(a|s)\, u) = \frac{4x_\theta}{(x_\theta^2-3)(1+x_\theta^2)^3} \frac{1}{\sigma^3} \varphi(s)\varphi(s)^\top \varphi(s)^\top u$。~~因此，~~$\|\nabla^2 \log \pi_\theta(a|s) - \nabla^2 \log \pi_{\theta'}(a|s)\| \leq l_2 \|\theta - \theta'\|$可用 $l_2 = \frac{4D^3}{\sigma^3}$ 界定。

假设3（费舍尔非退化性）： 利用柯西分布费舍尔信息的重参数化定理，$\pi_\theta(a|s)$ 的费舍尔信息矩阵可写为 $F_\rho(\theta) = \frac{1}{2\sigma^2} \varphi(s)\varphi(s)^\top$，对于满行秩特征映射，该矩阵为正定矩阵。

关于假设4 （有界相容函数逼近迁移误差）： 如前所述，这一假设已在多项研究中使用（Agarwal 等人，2021；Liu 等人，2020年；Ding 等人，2022年；Yuan 等人，2022b）。

C. 主要结果证明概要

本节将简要概述我们的结果证明，以突出说明相较于先前研究实现样本复杂度改进所采用的证明技术。样本复杂度的改进源于两方面的共同作用：（a）利用宽松弱梯度支配；（b）推导出更优的方差误差控制。完整的形式化证明将在后续章节中给出。

步骤 1. 类上升引理。定义$\hat{e}_t \overset{\text{def}}{=} d_t - \nabla J_H(\theta_t)$。首先，在引理5 中，我们证明：对于任意序列 θ_t，若其通过 $\theta_{t+1} = \theta_t + \gamma_t d_t/\|d_t\|$ 更新，且更新方向 d_t 非零（将由N-PG-IGT 或N-HARPG给出），则以下不等式成立：

$$J(\theta_{t+1}) - J^* \geq J(\theta_t) - J^* + \frac{\gamma_t}{3}\|\nabla J(\theta_t)\| - \frac{8\gamma_t}{3}\|\hat{e}_t\| - \mathcal{O}(\gamma_t^2 + \gamma^H \gamma_t)\,. \tag{6}$$

对于变体HARPG，不进行归一化时也有类似的引理，见引理8。

步骤 2. 宽松弱梯度支配不等式。结合假设3 和4，并沿用Ding 等人(2022年）的推导，我们得到期望回报函数 $J(\theta)$所满足的宽松弱梯度支配不等式（引理2）：

$$\forall \theta \in \mathbb{R}^d, \quad \varepsilon' + \|\nabla J(\theta)\| \geq \sqrt{2\mu}\,(J^* - J(\theta))\,, \tag{7}$$

其中 $\varepsilon' = \frac{\mu_F \sqrt{\varepsilon_{\text{bias}}}}{M_g\sqrt{1-\gamma}}()$ 且 $\mu = \frac{\mu_F^2}{2M_g^2}$. 目标函数的这一结构性质是我们分析中的关键工具之一。注意，我们还有 $\varepsilon' = \frac{\mu_F \sqrt{\varepsilon_{\text{bias}}}}{M_g\sqrt{1-\gamma}}()$ 且 $\mu = \frac{\mu_F^2}{2M_g^2}$. 特别是，误差 ε' 与策略参数化的表达能力密切相关，该表达能力由 $\varepsilon_{\text{bias}}$ 量化，如假设4中所述。注意，与Ding 等人(2022年）相比，我们没有推导期望梯度范数的界并将其代入（7）以获得最终的全局收敛速率。相反，我们将（7）仔细地纳入为（归一化）梯度方法推导的类上升引理中，从而得到关于 $\delta_t \overset{\text{def}}{=} J^* - \mathbb{E}\,[J(\theta_t)]$, 的递推关系，进而获得更优的收敛速率。

注释 2（关于梯度支配）。请注意，我们放宽的弱梯度支配条件采用梯度范数，而非梯度范数的平方。对于我们在此考虑的标准强化学习设定，若干近期研究已揭示并利用了这种结构性质（*Agarwal* 等人，*2021*；*Xiao*，*2022*年）。请注意，该条件不同于式（7）（其中 $\epsilon' = 0$）中的 *Polyak-Łojasiewicz*（*PŁ*）类条件，后者采用梯度范数的平方。连续控制中线性二次调节问题的状态反馈版本也有类似的 *PŁ* 条件（*Fazel* 等人，*2018*年；*Fatkhullin* 和 *Polyak*，*2021*），但这与我们当前的设定不同。

步骤 3：方差缩减控制。这一步至关重要，具体做法取决于所用算法，包括仔细推导关于 $\mathbb{E}\,[\|e_t\|]$ 或 $\mathbb{E}\left[\|e_t\|^2\right]$. 的递推关系。例如，在引理 7 中，对于 N-PG-IGT 方法，我们以向量形式推导出以下递推关系

$$\hat{e}_t = (1-\eta_t)\hat{e}_{t-1} + \eta_t e_t + (1-\eta_t)S_t + \eta_t Z_t, \tag{8}$$

其中 $\mathbb{E}\,[\|e_t\|] \leq \sigma_g$，而各项 S_t 和 Z_t 分别对应于 J 在不同点处的二阶泰勒近似。这使我们能够结合假设 2 和归一化更新规则（该规则保证 $\|\theta_{t+1} - \theta_t\| = \gamma_t$），并推导出对 $\|S_t\| \leq \mathcal{O}\left(\gamma_t^2 + \gamma^H\right)$的控制，$\|Z_t\| \leq \mathcal{O}\left(\gamma_t^2\eta_t^{-2} + \gamma^H\right)$。

通过选择 $\gamma_t = \mathcal{O}\left(t^{-1}\right)$，$\eta_t = \mathcal{O}\left(t^{-4/5}\right)$, $H = (1-\gamma)^{-1}\log(T+1)$，我们可以保证

$$\mathbb{E}\,[\|\hat{e}_t\|] \leq \mathcal{O}\left(\eta_t^{1/2} + \gamma_t^2\eta_t^{-2} + \gamma^H\eta_t^{-1} + \gamma^H\gamma_t\eta_t^{-1}\right) = \mathcal{O}\left(t^{-2/5}\right)\,.$$

我们强调，在分析中使用随时间变化的步长和动量参数十分重要，而不是使用依赖于迭代总次数 T 的恒定步长。具体而言，选择随时间变化的 η_t 可以使我们对所有8求解（$t \geq 1$，并将其代入（6）。

步骤 4：结合方差缩减估计和类上升引理。最后，我们将步骤 2–3 的结果代入步骤 1 中推导出的类上升引理，从而得到序列 (δ_t) 的递推关系。例如，对于N-PG-IGT方法，我们得到

$$\delta_{t+1} \leq (1 - \Omega(\gamma_t))\delta_t + \mathcal{O}(\gamma_t t^{-2/5} + \gamma_t^2 + \varepsilon'\gamma_t).$$

接下来需要利用第 (δ_t) 节中的技术引理，求解由此得到的序列上的递推关系，从而获得最终收敛速率。J

D. 补充相关工作

本节将补充简要讨论，包括方差缩减策略梯度方法的一阶驻点性、精确策略梯度方法的全局最优性，以及一些补充相关工作，以完善论文第 1 节的主要讨论。

D.1. 方差缩减策略梯度方法的一阶驻点性

过去几年，文献中提出了多种方差缩减策略梯度方法，以提高相较于原始策略梯度方法的样本效率；原始策略梯度方法的样本复杂度为 $\tilde{\mathcal{O}}(\varepsilon^{-4})$。Papini 等人（2018年）提出了一种随机方差缩减策略梯度（SVRPG）方法，其样本复杂度为 $\tilde{\mathcal{O}}(\varepsilon^{-4})$。Xu 等人（2020b）重新审视了该算法，并证明其样本复杂度为 $\tilde{\mathcal{O}}(\varepsilon^{-10/3})$。Shen 等人（2019年）利用二阶信息，或 Xu 等人（2020a）、Pham 等人（2020年）和 Gargiani 等人（2022年）采用 SARAH（Nguyen 等人，2017）、SPIDER（Fang 等人，2018）或 PAGE（Li 等人，2021）等方差缩减方法的变体，进一步将这一结果改进至 $\tilde{\mathcal{O}}(\varepsilon^{-3})$。Yuan 等人和 Huang 等人（Yuan 等人，2020年；Huang 等人，2020年）提出了基于 STORM（Cutkosky 和 Orabona，2019）的基于动量的策略梯度方法。具体而言，Huang 等人（2020年）消除了方差缩减对大批量轨迹的需求。本节上述所有提出方差缩减策略梯度方法的工作（使用二阶信息的 Shen 等人（2019年）除外），均采用重要性采样来处理采样分布的非平稳性。由于使用重要性采样机制来解决非平稳性问题，以往大多数工作都要求一个无法验证的假设，即重要性采样权重的方差有界。为解决这一问题，Zhang 等人（2021a）针对 softmax 参数化提出了一种与重要性采样互补的梯度截断机制，而 Shen 等人（2019年）和 Salehkaleybar 等人（2022年）则引入了二阶信息。这两种方法都实现了相同的 $\tilde{\mathcal{O}}(\varepsilon^{-3})$ FOSP 样本复杂度。

D.2. 精确策略梯度方法的全局最优性

近期，一系列研究转而关注建立全局最优收敛率，即推导达到期望回报函数的 ε 近似全局最优解所需的样本复杂度（Agarwal 等人，2021；Fazel 等人，2018；Bhandari 和 Russo，2019；Mei 等人，2020；Zhang 等人，2020b；2021b），而非仅仅达到 ε 近似一阶驻点。这一系列研究利用了策略优化问题的额外结构，其形式包括梯度支配（Agarwal 等人，2021；Fazel 等人，2018）、Łojasiewicz 型不等式（Mei 等人，2020；2021b）或隐式凸性（Zhang 等人，2021a）。这些结构性质保证非凹策略优化目标不存在次优驻点。大多数此类工作假设可以访问精确策略梯度，以证明快速收敛速率（Mei 等人，2020；2021b；Xiao（2022））。另一些工作也在假设可以访问精确梯度的情况下，为自然策略梯度方法建立了全局最优性保证（Khodadadian 等人，2022b；Bhandari 和 Russo，2021）。具体而言，Khodadadian 等人（2022b）和 Bhandari 和 Russo（2021）通过揭示策略梯度与策略迭代之间的联系，证明采用自适应步长的自然策略梯度方法具有线性收敛速率。Lan（2022）和 Xiao（2022）进一步改进了这些结果，证明在表格型设定下，包括自然策略梯度方法和投影 Q 下降方法在内的一般策略镜像下降方法，在采用几何递增步长时具有线性收敛速率，优于此前已知的较简单策略梯度方法的次线性收敛率（Agarwal 等人，2021；Mei 等人，2020）。Zhang 等人（2020a）考虑了一个更一般的问题，即最大化状态-动作占用度量的一般凹效用函数。利用该问题的隐式凸性结构，他们基于一个针对一般效用强化学习的新型变分策略梯度定理，提出了一种全局收敛算法。

D.3. 关于随机策略优化的进一步讨论

对于随机同策略梯度，在真实梯度设定下利用几何信息带来的加速效果，显然无法直接应用于随机梯度情形；Mei 等人(2021a)针对表格型 softmax 参数化准确地报告了这一点。具体而言，Mei 等人 (2021a)揭示了利用几何信息加速收敛与克服随机梯度噪声（其方差可能为无穷大）之间的权衡。我们还要指出，有若干研究（例如，Chen 等人 (2022)；Chen 和 Maguluri (2022)；Khodadadian 等人 (2022a)；Hong 等人 (2020)；Qiu 等人 (2021)；Wu 等人 (2020)；Xu 等人(2020c)）提出了有限时间分析，在底层马尔可夫链满足遍历性假设的条件下，使用（线性）函数逼近来逼近 Q 函数或优势函数，以求解策略优化问题的演员-评论家方法。我们的工作聚焦于基于蒙特卡洛的策略梯度方法（类似 REINFORCE），这类方法

无需评论家（因此不会产生相对于该评论家过程的函数近似误差），也无需遍历性假设。相较之下，我们依赖兼容函数逼近框架。除这一系列研究外，Zhang 等人(2021b)针对回合式强化学习中的 REINFORCE 算法，在采用固定小批量而非任意大的批次大小时，给出了全局收敛速率。这些结果包括任意时刻均成立的次线性高概率遗憾界，以及平均遗憾几乎必然全局收敛且收敛速率渐近次线性。具体而言，在遗憾分析中，他们考虑了带有渐消对数障碍正则化的 softmax 参数化，所得结果对应于一个 $\tilde{\mathcal{O}}(\varepsilon^{-6})$ 全局样本复杂度，这一点也见于 (Yuan 等人, 2022b)。最近，若干研究者考虑了重尾策略参数化，这种参数化能在连续控制问题中提供更好的探索能力（Bedi 等人, 2021; 2022）。

D.4. 随机优化领域与先前工作的比较

在随机优化设定中，近期有若干研究考虑优化满足弱梯度支配不等式的函数（见引理 2 (7)，其中 $\varepsilon' = 0$）。具体而言，(Fontaine 等人，2021；Fatkhullin 等人，2022；Scaman 等人，2022) 得到了 $\mathcal{O}(\varepsilon^{-3})$ 的样本复杂度，可在期望意义下找到全局最优解；其随机梯度下降[11], 的复杂度与强化学习设定中的（随机）原始策略梯度算法相同（Yuan 等人，2022b）。Fatkhullin 等人（2022）进一步在额外的平均光滑性假设下，通过提出一种名为 PAGER 的方差缩减算法，将这一复杂度改进为 $\mathcal{O}(\varepsilon^{-2})$。该算法采用双循环结构，并在两种不同的批量大小序列（随迭代递增）之间进行概率切换，因此实现和分析都更加困难。Masiha 等人（2022）的另一项研究则在类似假设下考察随机二阶方法。他们提出的随机三次正则化牛顿算法（SCRN）在一组类似假设下达到了 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ 的样本复杂度。然而，SCRN 使用一个子程序近似求解三次正则化子问题（导致计算效率低下），需要使用随机海森矩阵样本，并且在每次迭代中对随机梯度和海森矩阵都要求较大的批量。据我们所知，(Fatkhullin 等人，2022) 和 (Masiha 等人，2022) 是仅有的针对满足弱梯度支配条件的函数进行随机优化、并给出优于 $\mathcal{O}(\varepsilon^{-3})$ 样本复杂度的研究。然而，这些优化研究采用了复杂的算法，每次迭代涉及参数重启、三次正则化子问题和/或较大的小批量。相比之下，我们提出的算法（N-PG-IGT、HARPG、N-HARPG）以简单的算法将样本复杂度改进至 $\mathcal{O}(\varepsilon^{-3})$ 以下；这些算法计算高效且无需批量（每次迭代只需采样一条轨迹）。此外，我们的分析比 (Fatkhullin 等人，2022) 和 (Masiha 等人，2022) 中算法的分析简单得多。我们的分析对于研究弱梯度支配假设下的随机优化也具有独立价值。

[11]这些工作考虑了更一般的假设，例如 α-PŁ，即存在 $\alpha \in [1,2]$，使得对所有 $\theta \in \mathbb{R}^d$，均有 $\|\nabla J(\theta)\|^\alpha \geq (2\mu)^{\alpha/2}(J^* - J(\theta))$，并推导出取决于 α 值的样本复杂度。结果表明，当 $\alpha = 1$（如我们的设定）时，与其他 $\alpha > 1$ 相比，样本复杂度最差（关于 ε）。

E. 符号说明与实用引理

符号说明。对于任意整数 p，欧几里得空间 $\mathbb{R}^p$ 配备其通常的内积 $\langle\cdot,\ \cdot\rangle$ 以及相应的二范数 $\|\cdot\|$ 。向量 $x \in \mathbb{R}^p$ 的转置（其中 p 为任意整数）记为 $x^\top$，维度为 p的单位矩阵记为 I_p。对于两个矩阵 $A,\ B \in \mathbb{R}^{p\times q}$，若 p和 $q \geq 1$为两个整数，记号 $A \succcurlyeq B$ 表示矩阵 $A-B$ 为半正定。对于两个非负实数序列 (a_n)，(b_n)，如果存在 $c>0$ 使得对每个整数 n 都有 $a_n \leq c\, b_n$，我们就使用记号 $a_n = \mathcal{O}(b_n)$。因此，除非另有说明，我们使用 $\mathcal{O}(\cdot)$ 隐去 $\tilde{}$，其中只包含与任何问题参数无关的绝对常数；此外，我们使用记号 $\mathcal{O}(\cdot)$ 隐去绝对常数和对数因子。

优势函数。为完整起见，我们定义了优势函数，它出现在本文主体部分假设 4 的陈述中所提到的相容函数逼近误差中。为此，我们首先为每个策略定义 π 状态-动作价值函数 $Q^\pi:\ \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ 对于每个 $s\in\mathcal{S},\ a\in\mathcal{A}$ 定义如下：

$$Q^\pi(s,a) \stackrel{\text{def}}{=} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=a, a_0=a\right].$$

对于相同的策略 π，状态价值函数 $V^\pi:\ \mathcal{S}\to\mathbb{R}$ 和优势函数 $A^\pi:\ \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ 对每个 $s\in\mathcal{S},\ a\in\mathcal{A}$ 定义如下：

$$V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)],$$
$$A^\pi(s,a) \stackrel{\text{def}}{=} Q^\pi(s,a) - V^\pi(s).$$

现在我们陈述两个在分析中常用的实用引理。第一个引理表明，在策略参数化满足二阶正则性条件时，期望回报函数的海森矩阵是利普希茨连续的。

引理 3（（Zhang 等人，2020b）中的（A.68））。设假设 *1* 和 *2* 成立，则有

$$\left\|\nabla^2 J(\theta) - \nabla^2 J(\theta')\right\| \leq L_h \|\theta-\theta'\|_2,$$

其中 $L_h = \frac{r_{\max}M_g M_h(1-\gamma)^2}{M_g(1+\gamma)+M_h\gamma\ 1-\gamma} + \frac{r_{\max}M_g^3(1+\gamma)(1-\gamma)^3}{1-\gamma} + \frac{r_{\max}M_g}{1-\gamma}$ 最大 $\left\{M_h, \frac{\gamma M_g^2}{1-\gamma}, \frac{l_2}{M_g}, \frac{M_h\gamma}{1-\gamma}\right\}$。

第二个引理控制了在无限时域设定下，将模拟轨迹截断至时域 H 所产生的截断误差。值得注意的是，该误差随时域 H的增长呈几何速度迅速趋于零。

引理 4 (（Masiha 等人，2022年）中的引理 2)。设假设 *1* 成立，则对所有 $\theta\in\mathbb{R}^d$ 以及每个 $H\geq 1$，有

$$\|\nabla J_H(\theta) - \nabla J(\theta)\| \leq D_g\gamma^H, \qquad \left\|\nabla^2 J_H(\theta) - \nabla^2 J(\theta)\right\| \leq D_h\gamma^H,$$

其中 D_g 定义= $\frac{M_g r_{\max}}{1-\gamma}\sqrt{\frac{1}{1-\gamma}+H}$, D_h 定义= $\frac{r_{\max}(}{\frac{M_h+M_g^2}{1-\gamma}\left(H+\frac{1}{1-\gamma}\right)}$.

定理证明 F1 (N-PG-IGT)

F.1. 全局收敛

我们首先给出定理1的更详细版本。

定理 4. 设假设 *1, 2, 3* 和 *4* 成立。令 $\gamma_t = \frac{6}{\sqrt{2}\mu(t+2)} = \frac{6M_g}{\mu_F(t+2)}$, $\eta_t = \left(\frac{2}{t+2}\right)^{4/5}$, $H = 9\,5\,(1-\gamma)^{-1}\log(T+1)$。则对于每个整数 $T \geq 1$，*N-PG-IGT* 的输出 N-PG-IGT （算法 *1*）满足

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{J^* - J(\theta_0)}{(T+1)^2} + \frac{D_h}{\mu(T+1)^2} + \frac{L_g + \sqrt{\mu}D_g}{\mu(T+1)} + \frac{\sigma_g\mu + L_h}{\mu^{3/2}(T+1)^{2/5}} + \frac{\varepsilon'}{\sqrt{\mu}}\right). \tag{9}$$

其中 σ_g, L_g, L_h 分别定义于命题 *1* 和引理 *1* 中， D_g 定义于引理 *4*中， ε' 和 μ 定义于引理 *2*中。这里 $\mathcal{O}(\cdot)$ 仅省略绝对数值常数。

我们首先针对函数 J （我们要最大化 J）证明一个类上升引理。对于任意一列更新方向 (d_t)，以下结果适用于一般的归一化更新规则。

引理 5. 设假设 *1* 成立。令(d_t) 为 $\mathbb{R}^d$ 中的任意向量序列，并考虑由下式定义的序列 (θ_t):

$$\theta_{t+1} = \theta_t + \gamma_t \frac{d_t}{\|d_t\|},$$

其中 $\theta_0 \in \mathbb{R}^d$ （若 $d_t = 0$，则 $\theta_{t+1} = \theta_t$ ）。那么，对于每个整数 $t \geq 1$ 以及任意正步长 γ_t，有

$$-J(\theta_{t+1}) \leq -J(\theta_t) - \frac{\gamma_t}{3}\|\nabla J(\theta_t)\| + \frac{8\gamma_t}{3}\|d_t - \nabla J_H(\theta_t)\| + \frac{L_g\gamma_t^2}{2} + \frac{4\gamma_t}{3}\|\nabla J_H(\theta_t) - \nabla J(\theta_t)\|. \tag{10}$$

证明。 根据期望回报函数的光滑性 J （命题 1），并利用 θ_t 的更新规则，可得

$$\begin{aligned}
-J(\theta_{t+1}) &\leq -J(\theta_t) - \langle \nabla J(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L_g}{2}\|\theta_{t+1} - \theta_t\|^2 \\
&= -J(\theta_t) - \gamma_t \frac{\langle \nabla J(\theta_t), d_t\rangle}{\|d_t\|} + \frac{L_g\gamma_t^2}{2} \\
&\leq -J(\theta_t) - \gamma_t \frac{\langle \nabla J_H(\theta_t), d_t\rangle}{\|d_t\|} + \frac{L_g\gamma_t^2}{2} + \gamma_t\|\nabla J_H(\theta_t) - \nabla J(\theta_t)\|.
\end{aligned}$$

现在来界定上述不等式中的第二项。定义 $\hat{e}_t = d_t - \nabla J_H(\theta_t)$。我们分两种情况讨论。首先，若 $\|\hat{e}_t\| \leq \frac{1}{2}\|\nabla J_H(\theta_t)\|$，则

$$\begin{aligned}
-\frac{\langle \nabla J_H(\theta_t), d_t\rangle}{\|d_t\|} &= \frac{-\|\nabla J_H(\theta_t)\|^2 - \langle \nabla J_H(\theta_t), \hat{e}_t\rangle}{\|d_t\|} \\
&\leq \frac{-\|\nabla J_H(\theta_t)\|^2 + \|\nabla J_H(\theta_t)\|\,\|\hat{e}_t\|}{\|d_t\|} \\
&\leq \frac{-\|\nabla J_H(\theta_t)\|^2 + \frac{1}{2}\|\nabla J_H(\theta_t)\|^2}{\|d_t\|} \\
&\leq -\frac{\|\nabla J_H(\theta_t)\|^2}{2\left(\|\nabla J_H(\theta_t)\| + \|\hat{e}_t\|\right)} \\
&\leq -\frac{1}{3}\|\nabla J_H(\theta_t)\|.
\end{aligned}$$

否则，如果 $\|\hat{e}_t\| \geq \frac{1}{2}\|\nabla J_H(\theta_t)\|$，则有

$$\begin{aligned}
-\frac{\langle \nabla J_H(\theta_t), d_t\rangle}{\|d_t\|} &\leq \|\nabla J_H(\theta_t)\| \\
&= -\frac{1}{3}\|\nabla J_H(\theta_t)\| + \frac{4}{3}\|\nabla J_H(\theta_t)\| \\
&\leq -\frac{1}{3}\|\nabla J_H(\theta_t)\| + \frac{8}{3}\|\hat{e}_t\|,
\end{aligned}$$

综合这两种情况，即完成了该引理的证明：

$$\begin{aligned} -J(\theta_{t+1}) &\leq -J(\theta_t) - \frac{\gamma_t}{3}\left\|\nabla J_H(\theta_t)\right\| + \frac{8\gamma_t}{3}\left\|\hat{e}_t\right\| + \frac{L_g\gamma_t^2}{2} + \gamma_t\left\|\nabla J_H(\theta_t) - \nabla J(\theta_t)\right\| \\ &\leq -J(\theta_t) - \frac{\gamma_t}{3}\left\|\nabla J(\theta_t)\right\| + \frac{8\gamma_t}{3}\left\|\hat{e}_t\right\| + \frac{L_g\gamma_t^2}{2} + \frac{4\gamma_t}{3}\left\|\nabla J_H(\theta_t) - \nabla J(\theta_t)\right\| . \end{aligned} \tag{11}$$

□

我们为本文后续内容引入一种便捷的简写符号：

$$\delta_t \overset{\text{def}}{=} J^\star - \mathbb{E}\left[J(\theta_t)\right] , \tag{12}$$

其中， $J^\star$ 是最优期望回报，而 (d_t),(θ_t) 是算法计算得到的序列（本节稍后介绍N-PG-IGT，后续各节介绍(N)-HARPG或其他算法）。

在下一个引理中，我们将宽松弱梯度支配不等式代入前面的类上升引理，以界定梯度范数。由此得到的不等式是关于期望回报差的递归界，这对于推导我们的收敛速率至关重要。

引理 6. 设假设 *1*、*3* 和 *4* 均成立。在引理 *5*的设定下，对每个整数 $t \geq 1$以及任意正步长 γ_t,

$$\delta_{t+1} - \delta_t \leq -\frac{\sqrt{2\mu}\gamma_t}{3}\delta_t + \frac{8\gamma_t}{3}\mathbb{E}[\|d_t - \nabla J_H(\theta_t)\|] + \frac{L_g\gamma_t^2}{2} + \frac{\varepsilon'\gamma_t}{3} + \frac{4}{3}\gamma_t D_g\gamma^H , \tag{13}$$

这里回顾一下，(δ_t) 定义于 (12)。

证明。 根据假设 3 和 4，我们可以使用引理 2并结合引理 5来界定梯度范数，得到

$$-J(\theta_{t+1}) \leq -J(\theta_t) - \frac{\sqrt{2\mu}\gamma_t}{3}\left(J^\star - J(\theta_t)\right) + \frac{8\gamma_t}{3}\left\|d_t - \nabla J_H(\theta_t)\right\| + \frac{L_g\gamma_t^2}{2} + \frac{\varepsilon'\gamma_t}{3} + \frac{4}{3}\gamma_t D_g\gamma^H,$$

其中，我们使用引理 4来控制（10）最后一项中的截断误差。将 $J^\star$ 加到两边并取期望，得到（13）。

□

根据引理 6，现在我们来控制剩余的误差项 $\mathbb{E}[\|d_t - \nabla J_H(\theta_t)\|]$.

引理 7。 设假设 *1* 和 *2* 成立。设 (θ_t)、(d_t) 和 (θ_t) 为通过 N-PG-IGT算法（见算法 *1*）生成的序列，其中 $\eta_t = \left(\frac{2}{t+2}\right)^{4/5}$ 和 $\gamma_t = \frac{6}{\sqrt{2\mu}\,(t+2)}$。那么，对于任意整数 $t \geq 0$，均有

$$\begin{aligned} \mathbb{E}\left[\|d_t - \nabla J_H(\theta_t)\|\right] \leq\ & 2\cdot\sqrt{C\left(\frac{8}{5},\frac{4}{5}\right)\sigma_g\eta_t^{1/2} + C\left(\frac{6}{5},\frac{4}{5}\right)L_h\gamma_t^2\eta_t^{-2}} \\ & + 2D_g\gamma^H C\left(0,\frac{4}{5}\right)\eta_t^{-1} + D_h\gamma^H C\left(1,\frac{4}{5}\right)\gamma_t\eta_t^{-1}, \end{aligned} \tag{14}$$

其中 $C(p,\ q)$ 对于 $q \in [0,1)$， $p > 0$ 是引理 *15*中定义的一个数值常数。

证明。 定义$\hat{e}_t \overset{\text{def}}{=} d_t - \nabla J_H(\theta_t)$, $e_t \overset{\text{def}}{=} g(\tilde{\tau}_t,\tilde{\theta}_t) - \nabla J_H(\tilde{\theta}_t)$,

$$S_t \overset{\text{def}}{=} \nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + \nabla^2 J_H(\theta_t)\left(\theta_{t-1} - \theta_t\right),$$

$$\bar{S}_t \overset{\text{def}}{=} \nabla J(\theta_{t-1}) - \nabla J(\theta_t) + \nabla^2 J(\theta_t)\left(\theta_{t-1} - \theta_t\right)$$

$$Z_t \overset{\text{def}}{=} \nabla J_H(\tilde{\theta}_t) - \nabla J_H(\theta_t) + \nabla^2 J_H(\theta_t)\left(\tilde{\theta}_t - \theta_t\right).$$

$$\bar{Z}_t \overset{\text{def}}{=} \nabla J(\tilde{\theta}_t) - \nabla J(\theta_t) + \nabla^2 J(\theta_t)\left(\tilde{\theta}_t - \theta_t\right).$$

注意，根据引理 1 （二阶光滑性），引理4 以及三角不等式，可得

$$\begin{aligned} \|S_t\| &\leq L_h \|\theta_t - \theta_{t-1}\|^2 + \left\|S_t - \bar{S}_t\right\| = L_h \gamma_{t-1}^2 + \left\|S_t - \bar{S}_t\right\|, &(15)\\ \left\|S_t - \bar{S}_t\right\| &\leq 2D_g\gamma^H + D_h\gamma^H\gamma_{t-1} &(16)\end{aligned}$$

$$\begin{aligned} \|Z_t\| &\leq L_h \left\|\tilde{\theta}_t - \theta_t\right\|^2 + \left\|Z_t - \bar{Z}_t\right\| = L_h \frac{(1-\eta_t)^2}{\eta_t^2}\|\theta_t - \theta_{t-1}\|^2 + \left\|Z_t - \bar{Z}_t\right\|, &(17)\\ \left\|Z_t - \bar{Z}_t\right\| &\leq 2D_g\gamma^H + D_h\gamma^H\frac{\gamma_{t-1}}{\eta_t}. &(18)\end{aligned}$$

接下来，通过写出序列 $(\hat{e}_t)$ 的递推关系，我们可以推导出

$$\begin{aligned} \hat{e}_t &= d_t - \nabla J_H(\theta_t)\\ &\overset{(i)}{=} (1-\eta_t)d_{t-1} + \eta_t g(\tilde{\tau}_t, \tilde{\theta}_t) - \nabla J_H(\theta_t)\\ &= (1-\eta_t)\left(\hat{e}_{t-1} + \nabla J_H(\theta_{t-1})\right) + \eta_t g(\tilde{\tau}_t, \tilde{\theta}_t) - \nabla J_H(\theta_t)\\ &= (1-\eta_t)\hat{e}_{t-1} + \eta_t e_t + (1-\eta_t)\left(\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t)\right) + \eta_t\left(\nabla J_H(\tilde{\theta}_t) - \nabla J_H(\theta_t)\right)\\ &\overset{(ii)}{=} (1-\eta_t)\hat{e}_{t-1} + \eta_t e_t + (1-\eta_t)\left(\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + \nabla^2 J_H(\theta_t)\left(\theta_{t-1} - \theta_t\right)\right)\\ &\quad + \eta_t\left(\nabla J_H(\tilde{\theta}_t) - \nabla J_H(\theta_t) + \nabla^2 J_H(\theta_t)\left(\tilde{\theta}_t - \theta_t\right)\right)\\ &\quad - (1-\eta_t)\nabla^2 J_H(\theta_t)(\theta_{t-1} - \theta_t) - \eta_t\nabla^2 J_H(\theta_t)(\tilde{\theta}_t - \theta_t) \qquad (19)\\ &\overset{(iii)}{=} (1-\eta_t)\hat{e}_{t-1} + \eta_t e_t + (1-\eta_t)S_t + \eta_t Z_t, \end{aligned}$$

其中，(i)由序列(d_t)的更新规则得到，而(iii)则源于注意到 S_t, Z_t 并观察到，由于N-PG-IGT 中序列(θ_t)的更新规则，(ii)中的˜最后一项为零（最后˜项中的海森校正项消失了）。请注意，这一简化正是解释序列 θ_t的更新规则的主要原因。

考虑一个整数 $T \geq 1$.。展开递推关系并使用符号$\zeta_{t+1,T}$定义= $\prod_{\tau=t+1}^{T-1}(1-\eta_{\tau+1})$（其中$\zeta_{T,T} = 1$且$\zeta_{0,T} = \prod_{t=0}^{T-1}($ $1-\eta_{t+1})$），可得

$$\hat{e}_T = \zeta_{0,T}\hat{e}_0 + \sum_{t=0}^{T-1}\eta_{t+1}\zeta_{t+1,T}e_{t+1} + \sum_{t=0}^{T-1}(1-\eta_{t+1})\zeta_{t+1,T}S_{t+1} + \sum_{t=0}^{T-1}\eta_{t+1}\zeta_{t+1,T}Z_{t+1}\,.$$

对每个整数 $t \geq 1$，定义 σ-域 $\mathcal{F}_t$ 定义= $\sigma(\left\{\tilde{\theta}_0, \tilde{\tau}_0, \ldots, \tilde{\tau}_{t-1}\right\})$，其中对每个 $0 \leq s \leq t-1$， $\tilde{\tau}_s \sim p(\cdot|\pi_{\tilde{\theta}_s})$。注意，对于任意整数 $t_2 > t_1 \geq 1$，我们有

$$\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle|\mathcal{F}_{t_2}\right]\right] = \mathbb{E}\left[\langle e_{t_1}, \mathbb{E}\left[e_{t_2}|\mathcal{F}_{t_2}\right]\rangle\right] = 0.$$

然后利用三角不等式，取期望并应用詹森不等式，可得

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{e}_T\|\right] &\leq \zeta_{0,T}\mathbb{E}\left[\|\hat{e}_0\|\right] + \mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\eta_{t+1}\zeta_{t+1,T}e_{t+1}\right\|\right] \\
&\quad + \mathbb{E}\left[\left\|\sum_{t=0}^{T-1}(1-\eta_{t+1})\zeta_{t+1,T}S_{t+1}\right\|\right] + \mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\eta_{t+1}\zeta_{t+1,T}Z_{t+1}\right\|\right] \\
&\leq \zeta_{0,T}\sigma_g + \left(\mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\eta_{t+1}\zeta_{t+1,T}e_{t+1}\right\|^2\right]\right)^{1/2} \\
&\quad + \sum_{t=0}^{T-1}(1-\eta_{t+1})\zeta_{t+1,T}\mathbb{E}\left[\|S_{t+1}\|\right] + \sum_{t=0}^{T-1}\eta_{t+1}\zeta_{t+1,T}\mathbb{E}\left[\|Z_{t+1}\|\right] \\
&\overset{(i)}{\leq} \zeta_{0,T}\sigma_g + \left(\sum_{t=0}^{T-1}\eta_{t+1}^2\zeta_{t+1,T}^2\mathbb{E}\left[\|e_{t+1}\|^2\right]\right)^{1/2} \\
&\quad + L_h\sum_{t=0}^{T-1}(1-\eta_{t+1})\zeta_{t+1,T}\gamma_t^2 + L_h\sum_{t=0}^{T-1}\frac{\gamma_t^2}{\eta_{t+1}}\zeta_{t+1,T} \\
&\quad + \sum_{t=0}^{T-1}(1-\eta_{t+1})\zeta_{t+1,T}\left\|S_{t+1}-\bar{S}_{t+1}\right\| + \sum_{t=0}^{T-1}\eta_{t+1}\zeta_{t+1,T}\left\|Z_{t+1}-\bar{Z}_{t+1}\right\| \\
&\leq \zeta_{0,T}\sigma_g + \left(\sum_{t=0}^{T-1}\eta_{t+1}^2\zeta_{t+1,T}^2\right)^{1/2}\sigma_g + 2L_h\sum_{t=0}^{T-1}\frac{\gamma_t^2}{\eta_{t+1}}\zeta_{t+1,T} \\
&\quad + 2D_g\gamma^H\sum_{t=0}^{T-1}\zeta_{t+1,T} + D_h\gamma^H\sum_{t=0}^{T-1}\gamma_t\zeta_{t+1,T},
\end{aligned}
\tag{20}
$$

其中在 (i) 中，我们应用 (15) 和 (17)，并利用独立性 $\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle\right] = 0$，对任意整数 $t_1 \neq t_2$（大于 1）。在最后一个不等式中，我们对任意 $\mathbb{E}\left[\|e_{t+1}\|^2\right] \leq \sigma_g^2$ 使用 $t \in \{0, \cdots, T-1\}$。

此外，利用引理14和15，我们有

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{e}_T\|\right] &\leq \eta_T\sigma_g + \sqrt{C\left(\frac{8}{5},\frac{4}{5}\right)}\eta_T^{1/2}\sigma_g + C\left(\frac{6}{5},\frac{4}{5}\right)\gamma_T^2\eta_T^{-2}L_h \\
&\quad + 2D_g\gamma^H C\left(0,\frac{4}{5}\right)\eta_T^{-1} + D_h\gamma^H C\left(1,\frac{4}{5}\right)\gamma_T\eta_T^{-1} \\
&\leq 2\cdot\sqrt{C\left(\frac{8}{5},\frac{4}{5}\right)}\eta_T^{1/2}\sigma_g + C\left(\frac{6}{5},\frac{4}{5}\right)\gamma_T^2\eta_T^{-2}L_h \\
&\quad + 2D_g\gamma^H C\left(0,\frac{4}{5}\right)\eta_T^{-1} + D_h\gamma^H C\left(1,\frac{4}{5}\right)\gamma_T\eta_T^{-1},
\end{aligned}
$$

其中 $C(p,\ q)$ 对于 $q \in [0,1)$，$p > 0$ 在引理 15中定义。 □

定理4 （从而定理1）的证明结束。 现在，我们将结合纳入宽松弱梯度支配的类上升引理，以及前一个引理中得到的消失误差估计。

证明。 结合引理6 与引理7 的结果，我们得到

$$\begin{aligned}\delta_{t+1} &\leq \left(1-\frac{\sqrt{2\mu}\gamma_t}{3}\right)\delta_t+\frac{8\gamma_t}{3}\mathbb{E}\left[\|d_t-\nabla J_H(\theta_t)\|\right]+\frac{L_g\gamma_t^2}{2}+\frac{\gamma_t\varepsilon'}{3}+\frac{4\gamma_t D_g}{3}\gamma^H\\ &\leq \left(1-\frac{\sqrt{2\mu}\gamma_t}{3}\right)\delta_t+\frac{16}{3}\sqrt{C\left(\frac{8}{5},\frac{4}{5}\right)}\sigma_g\gamma_t\eta_t^{1/2}+\frac{8}{3}C\left(\frac{6}{5},\frac{4}{5}\right)L_h\gamma_t^3\eta_t^{-2}+\frac{L_g\gamma_t^2}{2}+\frac{\gamma_t\varepsilon'}{3}+\frac{4\gamma_t D_g}{3}\gamma^H\\ &\quad+\frac{16}{3}D_g\gamma^H C\left(0,\frac{4}{5}\right)\gamma_t\eta_t^{-1}+\frac{8}{3}D_h\gamma^H C\left(1,\frac{4}{5}\right)\gamma_t^2\eta_t^{-1}. \end{aligned}\tag{21}$$

使用引理12结合 $\alpha_t=\gamma_t,\quad a=\frac{3}{\sqrt{2\mu}},\quad t_0=0,\quad \tau=2,\quad \beta_t=\frac{16}{3}\sqrt{C\left(\frac{8}{5},\frac{4}{5}\right)}\sigma_g\gamma_t\eta_t^{1/2}+\frac{8}{3}C\left(\frac{6}{5},\frac{4}{5}\right)L_h\gamma_t^3\eta_t^{-2}+\frac{L_g\gamma_t^2}{2}+\frac{\gamma_t\varepsilon'}{3}+\frac{4\gamma_t D_g}{3}\gamma^H+\frac{16}{3}D_g\gamma^H C\left(0,\frac{4}{5}\right)\gamma_t\eta_t^{-1}+\frac{8}{3}D_h\gamma^H C\left(1,\frac{4}{5}\right)\gamma_t^2\eta_t^{-1}$

$$\begin{aligned}\delta_T &\leq \frac{\delta_0}{(T+1)^2}+\frac{\sum_{t=0}^{T-1}\beta_t(t+2)^2}{(T+1)^2}\\ &\leq \frac{\delta_0}{(T+1)^2}+\frac{16}{3}\sqrt{C\left(\frac{8}{5},\frac{4}{5}\right)}\frac{\sigma_g\gamma_0}{(T+1)^{2/5}}+\frac{16}{3}C\left(\frac{6}{5},\frac{4}{5}\right)\frac{L_h\gamma_0^3}{(T+1)^{2/5}}\\ &\quad+\frac{L_g\gamma_0^2}{2}\frac{1}{T+1}+\frac{\gamma_0\varepsilon'}{3}+\frac{4\gamma_0 D_g}{3}\gamma^H+\frac{16\gamma_0}{3}D_g\gamma^H C\left(0,\frac{4}{5}\right)(T+1)^{4/5}+\frac{8\gamma_0^2}{3}D_h\gamma^H C\left(1,\frac{4}{5}\right)\frac{1}{(T+1)^{1/5}}.\end{aligned}$$

□

F.2. 收敛到一阶驻点

定理*5.* 设假设 *1* 和 *2* 成立。令 $\gamma_t=\left(\frac{2}{t+2}\right)^{5/7}$, $\eta_t=\left(\frac{2}{t+2}\right)^{4/7}$, $H=(1-\gamma)^{-1}\log(T+1)$。令 $\bar{\theta}_T$ 从 *N-PG-IGT*（算法 *1*）的迭代结果中以概率分布 $\mathbb{P}(\theta_T=\theta_t)=\frac{\gamma_t}{\sum_{t=0}^{T-1}\gamma_t}$ 进行采样。则对于每个整数 $T\geq 1$，θ_T 满足

$$\mathbb{E}\left[\left\|\nabla J(\bar{\theta}_T)\right\|\right]\leq\mathcal{O}\left(\frac{J^*-J(\theta_0)}{(T+1)^{2/7}}+\frac{L_g}{(T+1)^{5/7}}+\frac{D_g}{(T+1)}+\frac{D_h}{(T+1)^{13/7}}+\frac{(L_h+\sigma_g)\log(T+1)}{(T+1)^{2/7}}\right).$$

其中 σ_g, L_g, L_h 分别定义于命题 *1* 和引理 *1* 中，D_g 定义于引理 *4*中。此处 $\mathcal{O}(\cdot)$ 仅代表绝对数值常数。

证明。 由引理 5 结合引理4 可得

$$-J(\theta_{t+1})\leq -J(\theta_t)-\frac{\gamma_t}{3}\|\nabla J(\theta_t)\|+\frac{8\gamma_t}{3}\|d_t-\nabla J_H(\theta_t)\|+\frac{L_g\gamma_t^2}{2}+\frac{4\gamma_t D_g}{3}\gamma^H.\tag{22}$$

我们现在对误差项 $\|\hat{e}_t\|=\|d_t-\nabla J_H(\theta_t)\|$ 进行如下控制。根据引理 7的证明中的 (20)，我们有

$$\begin{aligned}\mathbb{E}\left[\|\hat{e}_T\|\right] &\leq \zeta_{0,T}\sigma_g+\left(\sum_{t=0}^{T-1}\eta_{t+1}^2\zeta_{t+1,T}^2\right)^{1/2}\sigma_g+2L_h\sum_{t=0}^{T-1}\frac{\gamma_t^2}{\eta_{t+1}}\zeta_{t+1,T}+2D_g\gamma^H\sum_{t=0}^{T-1}\zeta_{t+1,T}+D_h\gamma^H\sum_{t=0}^{T-1}\gamma_t\zeta_{t+1,T}\\ &\overset{(i)}{\leq} 2\cdot\sqrt{C\left(\frac{8}{7},\frac{4}{7}\right)}\eta_T^{1/2}\sigma_g+C\left(\frac{6}{7},\frac{4}{7}\right)\gamma_T^2\eta_T^{-2}L_h+2D_g\gamma^H C\left(0,\frac{4}{7}\right)\eta_T^{-1}+D_h\gamma^H C\left(\frac{5}{7},\frac{4}{7}\right)\gamma_T\eta_T^{-1},\end{aligned}$$

其中在 (i) 中我们应用了引理 14 和 15。对于 $C(p,q)$ 在 $q\in[0,1)$，$p>0$ 定义于引理 15，$\hat{e}_t=d_t-\nabla J_H(\theta_t)$

将上述不等式与（22）结合，可得

$$\begin{aligned}\mathbb{E}\left[J(\theta_t)-J(\theta_{t+1})\right] &\leq -\frac{\gamma_t}{3}\mathbb{E}\left[\|\nabla J(\theta_t)\|\right]+8\gamma_t\sqrt{C\left(\frac{8}{7},\frac{4}{7}\right)}\eta_T^{1/2}\sigma_g+8\gamma_t C\left(\frac{6}{7},\frac{4}{7}\right)\gamma_T^2\eta_T^{-2}L_h+\frac{L_g\gamma_t^2}{2}+\frac{\varepsilon'\gamma_t}{3}\\ &\quad+\frac{4\gamma_t D_g}{3}\gamma^H+\frac{16}{3}D_g\gamma^H C\left(0,\frac{4}{7}\right)\gamma_t\eta_t^{-1}+\frac{8}{3}D_h\gamma^H C\left(\frac{5}{7},\frac{4}{7}\right)\gamma_t^2\eta_T^{-1}.\end{aligned}\tag{23}$$

对 $t=0$、.....、$T-1$ 的上述不等式求和并重新整理各项，可得 t

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla J(\bar{\theta}_T)\right\|\right] &= \frac{\sum_{t=0}^{T-1}\gamma_t\mathbb{E}\left[\|\nabla J(\theta_t)\|\right]}{\sum_{t=0}^{T-1}\gamma_t} \\ &\leq \mathcal{O}\left(\frac{J^*-J(\theta_0)}{(T+1)^{2/7}}+\frac{L_g}{(T+1)^{5/7}}+\frac{D_g}{(T+1)}+\frac{D_h}{(T+1)^{13/7}}+\frac{(L_h+\sigma_g)\log(T+1)}{(T+1)^{2/7}}\right). \quad (24)\end{aligned}$$

□

G. 定理证明 2 (HARPG)

本节中，为证明定理 2 （关于HARPG），我们引入以下有用的补充简写符号：

$$\delta_t \overset{\text{def}}{=} J^\star - \mathbb{E}\left[J(\theta_t)\right], \tag{25}$$

$$V_t \overset{\text{def}}{=} \mathbb{E}\left[\left\|d_t - \nabla J(\theta_t)\right\|^2\right], \tag{26}$$

$$\Lambda_t \overset{\text{def}}{=} \delta_t + \lambda_{t-1} V_t, \tag{27}$$

$$R_t \overset{\text{def}}{=} \mathbb{E}\left[\left\|\theta_{t+1} - \theta_t\right\|^2\right], \tag{28}$$

其中 (θ_t)，(d_t) 是由 HARPG 生成的序列，而 (λ_t) 是一个待定的正标量序列。首先，我们给出定理2的详细表述，然后再提供完整证明。

定理 6. 设假设 *1*、*3*和 *4*成立。令 $\gamma_t = \gamma_0 \eta_t^{1/2}$， $\eta_t = \frac{5}{t+5}$其中 $\gamma_0 = \min\left\{\frac{1}{8\sqrt{3}}, \frac{\eta_0^{1/2}}{(L_g+\sigma_g+D_h\gamma^{2H})}, \frac{1}{\sigma_g\sqrt{3\mu}}\right\}$且 $H = 2(1-\gamma)^{-1}\log(T+4)$。令 $\lambda_t = 4\gamma_0\eta_t^{-1/2}$。则对于任意整数 $T \geq 1$，*HARPG* θ_T的输出 （参见算法 *2*）满足

$$J^\star - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{\Lambda_0}{(T+4)^2} + \frac{\sigma_g\sqrt{\mu} + L_g + \sigma_h}{(T+4)^{1/2}} + \frac{\lambda_0 D_g^2}{(T+4)^{5/2}}\right) + \frac{\sqrt{2}\varepsilon'}{\sqrt{\mu}},$$

其中 σ_g， σ_h， L_g 分别由命题 *1* 和引理 *1* 定义， D_g 和 D_h 由引理 *4*定义， ε' 和 μ 由引理 *2*定义。

证明的第一步是一个类上升引理，它不同于此前使用的引理 5和引理 6 中的类似结果，因为 HARPG算法不会使用归一化方向来更新策略参数。

引理 8. 假设 *1*、*3* 和 *4* 成立。设 (d_t) 是 $\mathbb{R}^d$ 中的任意向量序列，并考虑如下定义的序列 (θ_t)：

$$\theta_{t+1} = \theta_t + \gamma_t d_t,$$

其中 $\theta_0 \in \mathbb{R}^d$。那么，对于每个整数 $t \geq 1$，对于任意满足 $\gamma_t \leq \frac{1}{2L_g}$ 的正步长 γ_t，我们有

$$-J(\theta_{t+1}) \leq -J(\theta_t) - \frac{\mu\gamma_t}{2}\left(J^\star - J(\theta_t)\right)^2 + \gamma_t\left\|d_t - \nabla J_H(\theta_t)\right\|^2 - \frac{1}{4\gamma_t}\left\|\theta_{t+1} - \theta_t\right\|^2 + \frac{\left(\varepsilon'\right)^2\gamma_t}{2} + \gamma_t D_g^2\gamma^{2H}. \tag{29}$$

此外，我们还有

$$\delta_{t+1} - \delta_t \leq -\frac{\gamma_t\mu}{2}\delta_t^2 + \gamma_t V_t - \frac{1}{4\gamma_t}R_t + \frac{\left(\varepsilon'\right)^2\gamma_t}{2} + \gamma_t D_g^2\gamma^{2H}, \tag{30}$$

其中我们使用与 (25) 至 (28)中相同的符号。

Proof. 利用 J （命题 1）的光滑性和序列的更新规则 (θ_t)， 可得

$$\begin{aligned}
-J(\theta_{t+1}) &\leq -J(\theta_t) - \langle\nabla J(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L_g}{2}\left\|\theta_{t+1} - \theta_t\right\|^2 \\
&= -J(\theta_t) - \gamma_t\left(\frac{1}{2}\left\|\nabla J(\theta_t)\right\|^2 + \frac{1}{2}\left\|d_t\right\|^2 - \frac{1}{2}\left\|d_t - \nabla J(\theta_t)\right\|^2\right) + \frac{L_g}{2}\left\|\theta_{t+1} - \theta_t\right\|^2 \\
&= -J(\theta_t) - \frac{\gamma_t}{2}\left\|\nabla J(\theta_t)\right\|^2 + \frac{\gamma_t}{2}\left\|d_t - \nabla J(\theta_t)\right\|^2 - \left(\frac{1}{2\gamma_t} - \frac{L_g}{2}\right)\left\|\theta_{t+1} - \theta_t\right\|^2 \\
&\leq -J(\theta_t) - \frac{\gamma_t}{2}\left\|\nabla J(\theta_t)\right\|^2 + \gamma_t\left\|d_t - \nabla J_H(\theta_t)\right\|^2 - \left(\frac{1}{2\gamma_t} - \frac{L_g}{2}\right)\left\|\theta_{t+1} - \theta_t\right\|^2 + \gamma_t\left\|\nabla J(\theta_t) - \nabla J_H(\theta_t)\right\|^2 \\
&\leq -J(\theta_t) - \frac{\gamma_t}{2}\left\|\nabla J(\theta_t)\right\|^2 + \gamma_t\left\|d_t - \nabla J_H(\theta_t)\right\|^2 - \left(\frac{1}{2\gamma_t} - \frac{L_g}{2}\right)\left\|\theta_{t+1} - \theta_t\right\|^2 + \gamma_t D_g^2\gamma^{2H},
\end{aligned}$$

其中，最后一个不等式由引理 4得出。由引理 2 可知，对于任意 $\theta \in \mathbb{R}^d$

$$(\varepsilon')^2 + \|\nabla J(\theta)\|^2 \geq \frac{(\varepsilon' + \|\nabla J(\theta)\|)^2}{2} \geq \mu(J(\theta) - J^\star)^2.$$

利用上述不等式和关于步长的假设，我们得到（29）。在两边加上 $J^\star$，取期望，并对 $x \mapsto x^2$应用詹森不等式，可得（30）。 □

根据引理 8中的（30），现在我们为序列(V_t)（定义见（26））给出一个递推界。**引理 9**。 假设假设条件 *1* 成立，且设 $\eta_t \leq 1$。设序列 (d_t) 对每个整数 $t \geq$ 1 *via*

$$d_t = (1-\eta_t)\left(d_{t-1} + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})\right) + \eta_t g(\tau_t, \theta_t)\,,$$

按照算法 *2*更新。那么，对于每个整数 t，序列(V_t) 满足以下递推关系

$$V_{t+1} - V_t \leq -\eta_{t+1} V_t + 2\sigma_g^2\eta_{t+1}^2 + 12(2L_g^2 + \sigma_h^2 + D_h^2\gamma^{2H})\mathbb{E}\left[\|\theta_{t+1} - \theta_t\|^2\right] + 24D_g^2\gamma^{2H}.$$

证明。 定义 $\mathcal{V}_t \overset{\text{def}}{=} g(\tau_t, \theta_t) - \nabla J_H(\theta_t)$, $\mathcal{W}_t \overset{\text{def}}{=} \nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})$。注意到 $\mathbb{E}[\mathcal{V}_t] = 0$和 $\mathbb{E}\left[\|\mathcal{V}_t\|^2\right] \leq \sigma_g^2$。此外，由于 $\mathbb{E}[\mathcal{W}_t] = 0$

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_t] &= \mathbb{E}\left[\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1}) \mid \theta_{t-1}, \theta_t, \hat{\theta}_t\right]\right] \\
&= \mathbb{E}\left[\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + \nabla^2 J_H(\hat{\theta}_t)(\theta_t - \theta_{t-1})\right] \\
&= \mathbb{E}\left[\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t)\right] + \mathbb{E}\left[\int_0^1 \nabla^2 J_H(q\theta_t + (1-q)\theta_{t-1})(\theta_t - \theta_{t-1})dq\right] = 0.
\end{aligned}$$

此外，我们可以如下控制序列 $(\mathcal{W}_t)$ 的方差:

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{W}_t\|^2\right] &= \mathbb{E}\left[\left\|\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})\right\|^2\right] \\
&\leq 6\,\mathbb{E}\left[\|\nabla J_H(\theta_{t-1}) - \nabla J(\theta_{t-1})\|^2\right] + 6\,\mathbb{E}\left[\|\nabla J(\theta_{t-1}) - \nabla J(\theta_t)\|^2\right] \\
&\quad + 6\,\mathbb{E}\left[\|\nabla J(\theta_t) - \nabla J_H(\theta_t)\|^2\right] + 6\,\mathbb{E}\left[\left\|\left(B(\hat{\tau}_t, \hat{\theta}_t) - \nabla^2 J_H(\hat{\theta}_t)\right)(\theta_t - \theta_{t-1})\right\|^2\right] \\
&\quad + 6\,\mathbb{E}\left[\left\|\left(\nabla^2 J_H(\hat{\theta}_t) - \nabla^2 J(\hat{\theta}_t)\right)(\theta_t - \theta_{t-1})\right\|^2\right] + 6\,\mathbb{E}\left[\left\|\nabla^2 J(\hat{\theta}_t)(\theta_t - \theta_{t-1})\right\|^2\right] \\
&\leq 2(6L_g^2 + 3\sigma_h^2)\mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^2\right] + 12D_g^2\gamma^{2H} + 6D_h^2\gamma^{2H}\mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^2\right], \qquad (31)
\end{aligned}$$

其中，最后一个不等式源于利用函数的光滑性 J、海森矩阵估计方差的有界性（两项结果均见命题1 ）以及引理4（用于控制截断误差）。

现在，我们利用序列 (d_t) 的更新规则，以及为证明此结论而引入的序列 $(\mathcal{V}_t)$ 和 $(\mathcal{W}_t)$ 的定义，重新改写我们关注的量 $d_t - \nabla J_H(\theta_t)$。于是有

$$\begin{aligned}
d_t - \nabla J_H(\theta_t) &= (1-\eta_t)\left(d_{t-1} + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})\right) + \eta_t g(\tau_t, \theta_t) - \nabla J_H(\theta_t) \\
&= (1-\eta_t)d_{t-1} + \eta_t g(\tau_t, \theta_t) + (1-\eta_t)B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1}) - \nabla J_H(\theta_t) \\
&= (1-\eta_t)\left(d_{t-1} - \nabla J_H(\theta_{t-1})\right) + \eta_t g(\tau_t, \theta_t) \\
&\qquad + (1-\eta_t)B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1}) - \nabla J_H(\theta_t) + (1-\eta_t)\nabla J_H(\theta_{t-1}) \\
&= (1-\eta_t)\left(d_{t-1} - \nabla J_H(\theta_{t-1})\right) + \eta_t\left(g(\tau_t, \theta_t) - \nabla J_H(\theta_t)\right) \\
&\qquad + (1-\eta_t)\left(\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})\right) \\
&= (1-\eta_t)\left(d_{t-1} - \nabla J_H(\theta_{t-1})\right) + \eta_t\mathcal{V}_t + (1-\eta_t)\mathcal{W}_t.
\end{aligned}$$

使用 t 根据上述分解，我们可以推导出序列 (V_t) 所需的递归上界 :

$$\begin{aligned} V_t &= \mathbb{E}\left[\|d_t - \nabla J_H(\theta_t)\|^2\right] \\ &= \mathbb{E}\left[\|(1-\eta_t)\left(d_{t-1} - \nabla J_H(\theta_{t-1})\right) + \eta_t \mathcal{V}_t + (1-\eta_t)\mathcal{W}_t\|^2\right] \\ &= (1-\eta_t)^2 \mathbb{E}\left[\|d_{t-1} - \nabla J_H(\theta_{t-1})\|^2\right] + \mathbb{E}\left[\|\eta_t \mathcal{V}_t + (1-\eta_t)\mathcal{W}_t\|^2\right] \\ &\leq (1-\eta_t) V_{t-1} + 2\eta_t^2 \mathbb{E}\left[\|\mathcal{V}_t\|^2\right] + 2\mathbb{E}\left[\|\mathcal{W}_t\|^2\right] \\ &\leq (1-\eta_t) V_{t-1} + 2\sigma_g^2 \eta_t^2 + 12(2L_g^2 + \sigma_h^2 + D_h^2 \gamma^{2H}) \mathbb{E}\left[\|\theta_t - \theta_{t-1}\|^2\right] + 24 D_g^2 \gamma^{2H}, \end{aligned}$$

其中，第三个等式是通过注意到随机变量 $\eta_t \mathcal{V}_t + (1-\eta_t)\mathcal{W}_t$ 关于 θ_{t-1} 的条件期望等于零，并利用条件期望的塔式性质得出的；第一个不等式使用了假设 $\eta_t \leq 1$，最后一个不等式则利用了本引理证明开头建立的方差界。

□

接下来，我们将引理 8 和引理 9的结果结合起来。为此，我们将式（(Λ_t) 中定义的序列 27 作为李雅普诺夫函数。注意，该函数将量 δ_t 和 V_t 与适当的权重序列 (λ_t) 结合起来（将在下一个结果中定义）。为方便起见，我们在此回顾序列的定义(Λ_t):

$$\Lambda_t \stackrel{\text{def}}{=} \delta_t + \lambda_{t-1} V_t \,.$$

引理 10. 设假设 *1*、*3* 和 *4* 成立。令 $\eta_t = \frac{5}{t+5}$ 和 $\gamma_t \leq \gamma_0 \eta_t^{1/2}$ ~~σ~~ 令 $\lambda_t = \lambda_0 \eta_t^{-1/2}$，其中 $\lambda_0 = 4\gamma_0$且 $\gamma_0 \leq \frac{1}{8\sqrt{3}\left(L_g + \sigma_h + D_h \gamma^{2H}\right)}$。那么，对于任意整数 $t \geq 1$，要么 ~~δ_t~~ $\leq \frac{\sqrt{2}\varepsilon'}{\sqrt{\mu}}$，要么

$$\Lambda_{t+1} \leq \left(1 - \frac{\eta_t}{5}\right) \Lambda_t + \left(\frac{1}{\mu \gamma_0} + 2\lambda_0 \sigma_g^2\right) \eta_t^{3/2} + 24 \lambda_0 \eta_t^{-1/2} D_g^2 \gamma^{2H}.$$

证明。 根据引理 8，可得

$$\delta_{t+1} - \delta_t \leq -\frac{\gamma_t \mu}{2} \delta_t^2 + \gamma_t V_t - \frac{1}{4\gamma_t} R_t + \frac{\left(\varepsilon'\right)^2 \gamma_t}{2} + \gamma_t D_g^2 \gamma^{2H}. \tag{32}$$

引理 9 给出以下不等式:

$$V_{t+1} - V_t \leq -\eta_{t+1} V_t + 2\sigma_g^2 \eta_{t+1}^2 + 12(2L_g^2 + \sigma_h^2 + D_h^2 \gamma^{2H}) R_t + 24 D_g^2 \gamma^{2H}. \tag{33}$$

将（32）与 $\lambda_t = \lambda_0 \eta_t^{-1/2}$ 倍的（33）相加，可得

$$\begin{aligned} \Lambda_{t+1} - \Lambda_t &= \delta_{t+1} - \delta_t + \lambda_t (V_{t+1} - V_t) + (\lambda_t - \lambda_{t-1}) V_t \\ &\leq -\frac{\gamma_t \mu}{2} \delta_t^2 + \left(\frac{\gamma_t}{2} + (\lambda_t - \lambda_{t-1}) - \lambda_t \eta_{t+1}\right) V_t + 2\lambda_t \sigma_g^2 \eta_{t+1}^2 \\ &\quad - \frac{1}{4\gamma_t} R_t + 12\lambda_t (2L_g^2 + \sigma_h^2 + D_h^2 \gamma^{2H}) R_t + \frac{\gamma_t \left(\varepsilon'\right)^2}{2} + 24 \lambda_t D_g^2 \gamma^{2H} \\ &\overset{(i)}{\leq} -\frac{\gamma_t \mu}{2} \delta_t^2 + \left(\frac{\gamma_t}{2} + (\lambda_t - \lambda_{t-1}) - \lambda_t \eta_{t+1}\right) V_t + 2\lambda_t \sigma_g^2 \eta_{t+1}^2 + \frac{\gamma_t \left(\varepsilon'\right)^2}{2} + 24 \lambda_t D_g^2 \gamma^{2H} \\ &\overset{(ii)}{\leq} -\frac{\gamma_t \mu}{2} \delta_t^2 - \frac{5}{6} \gamma_t V_t + 2\lambda_t \sigma_g^2 \eta_{t+1}^2 + \frac{\gamma_t \left(\varepsilon'\right)^2}{2} + 24 \lambda_t D_g^2 \gamma^{2H} \\ &\overset{(iii)}{\leq} -\frac{\gamma_t \mu}{4} \delta_t^2 - \frac{5}{6} \gamma_t V_t + 2\lambda_t \sigma_g^2 \eta_{t+1}^2 + 24 \lambda_t D_g^2 \gamma^{2H} \end{aligned} \tag{34}$$

其中，(*i*)由 $\lambda_0 = 4\gamma_0 \le \left(48(2L_g^2 + \sigma_h^2 + D_h^2\gamma^{2H})\gamma_0\right)^{-1}$ 得出。选择 $\gamma_0 = \frac{\lambda_0}{4}$，并结合上界 $\lambda_t - \lambda_{t-1} \le \frac{\lambda_0}{2}\eta_t^{1/2}$，可得

$$\frac{\gamma_t}{2} + (\lambda_t - \lambda_{t-1}) - \lambda_t\eta_{t+1} \le \frac{\gamma_t}{2} + \frac{\lambda_0}{2}\eta_t^{1/2} - \frac{5}{6}\lambda_0 \le -\frac{5}{6}\gamma_t,$$

因此，(*ii*)成立。最后一步(*iii*)由假设 $\delta_t \ge \frac{\sqrt{2\varepsilon'}}{\sqrt{\mu}}$ 推出。

注意到 $\delta_t = \Lambda_t - \lambda_{t-1}V_t$，我们有

$$\begin{aligned}
\Lambda_{t+1} - \Lambda_t &\le -\frac{\gamma_t\mu}{4}\left(\Lambda_t - \lambda_{t-1}V_t\right)^2 - \frac{5}{6}\gamma_t V_t + 2\lambda_t\sigma_g^2\eta_{t+1}^2 + 24\lambda_t D_g^2\gamma^{2H} \\
&\le -\frac{\gamma_t\mu}{4}\Lambda_t^2 - \frac{\gamma_t\mu\lambda_{t-1}^2}{4}V_t^2 + \frac{\gamma_t}{2}\left(\mu\lambda_{t-1}\Lambda_t - \frac{5}{3}\right)V_t + 2\lambda_t\sigma_g^2\eta_{t+1}^2 + 24\lambda_t D_g^2\gamma^{2H} \\
&\overset{(i)}{\le} -\frac{\gamma_t\mu}{4}\Lambda_t^2 + \frac{\gamma_t^2}{4}\left(\mu\lambda_{t-1}\Lambda_t - \frac{5}{3}\right)^2\frac{1}{\mu\gamma_t\lambda_{t-1}^2} + 2\lambda_t\sigma_g^2\eta_{t+1}^2 + 24\lambda_t D_g^2\gamma^{2H} \\
&= -\frac{\gamma_t\mu}{4}\Lambda_t^2 + \left(\mu\lambda_{t-1}\Lambda_t - \frac{5}{3}\right)^2\frac{\gamma_t}{4\mu\lambda_{t-1}^2} + 2\lambda_0\sigma_g^2\eta_{t+1}^{3/2} + 24\lambda_t D_g^2\gamma^{2H} \\
&= -\frac{5}{6}\frac{\gamma_t}{\lambda_{t-1}}\Lambda_t + \frac{25}{36}\frac{\gamma_t}{\mu\lambda_{t-1}^2} + 2\lambda_0\sigma_g^2\eta_{t+1}^{3/2} + 24\lambda_0\eta_t^{-1/2}D_g^2\gamma^{2H} \\
&= -\frac{5}{24}\eta_t\Lambda_t + \frac{25}{576}\frac{1}{\mu\gamma_0}\eta_t^{3/2} + 2\lambda_0\sigma_g^2\eta_{t+1}^{3/2} + 24\lambda_0\eta_t^{-1/2}D_g^2\gamma^{2H} \\
&\le -\frac{\eta_t}{5}\Lambda_t + \left(\frac{1}{\mu\gamma_0} + 2\lambda_0\sigma_g^2\right)\eta_t^{3/2} + 24\lambda_0\eta_t^{-1/2}D_g^2\gamma^{2H},
\end{aligned}$$

其中，在(*i*)中我们使用了 $-ax^2 + cx \le \frac{c^2}{4a}$，其中 $x = V_t$、$a = \frac{\gamma_t\mu\lambda_{t-1}^2}{4}$、$c = \frac{1}{2}\gamma_t\left(\mu\lambda_{t-1}\Lambda_t - \frac{1}{6}\right)$。

□

定理 6（因此定理2的证明结束）。由引理 10，我们有以下结论：要么 $\delta_t \le \frac{\sqrt{2\epsilon'}}{\sqrt{\mu}}$ 要么

$$\Lambda_{t+1} \le \left(1 - \frac{\eta_t}{5}\right)\Lambda_t + \left(\frac{1}{\mu\gamma_0} + 8\gamma_0\sigma_g^2\right)\eta_t^{3/2} + 24\lambda_0\eta_t^{-1/2}D_g^2\gamma^{2H}.$$

利用引理 12 以及 $t_0 = 0$，$\tau = 5$，$r_t = \Lambda_t$，$\alpha_t = \eta_t$，$a = \frac{2}{5}$，$\beta_t = c_1\eta_t^{3/2} + c_2\eta_t^{-1/2}$，$c_1 = \left(\frac{1}{\mu\gamma_0} + 8\gamma_0\sigma_g^2\right)$，$c_2 = 24\lambda_0 D_g^2\gamma^{2H}$，对每个整数 $T \ge 1$

$$\begin{aligned}
\Lambda_T &\le \frac{\Lambda_0}{(T+4)^2} + \frac{\sum_{t=0}^{T-1}(c_1\eta_t^{3/2} + c_2\eta_t^{-1/2})(t+5)^2}{(T+4)^2} \\
&\le \mathcal{O}\left(\frac{\Lambda_0}{(T+4)^2} + \frac{4\sum_{t=0}^{T-1}c_1(t+5)^{1/2} + c_2(t+5)^{5/2}}{(T+4)^2}\right) \\
&\le \mathcal{O}\left(\frac{\Lambda_0}{(T+4)^2} + \frac{4c_1}{(T+4)^{1/2}} + 4c_2(T+4)^{3/2}\right). \qquad (35)
\end{aligned}$$

取 $\gamma_0 = \min\left\{\frac{1}{8\sqrt{3}\ (L_g+\sigma_g+D_h\gamma^{2H})}, \frac{1}{2\sigma_g\sqrt{2\mu}}\right\}$，我们有 $c_1 \le \frac{8\sqrt{3}\ (L_g+\sigma_h+D_h\gamma^{2H})}{\mu} + \frac{2\sqrt{2}\sigma_g}{\sqrt{\mu}}$。由于对于 $x > 0$，$\log x \le x - 1$，我们有

$$\gamma^{2H} = \gamma^{4\log(T+4)(1-\gamma)^{-1}} \le \gamma^{-2\frac{\log(T+4)}{\log\gamma}} = \frac{1}{(T+4)^2}$$

因此 $c_2 \le 24\lambda_0 D_g^2\ \frac{1}{T+4()^2}$。

综合以上所有结果，我们得到

$$\delta_T \le \max\left\{\Lambda_T, \frac{\sqrt{2}\epsilon'}{\sqrt{\mu}}\right\} \le \mathcal{O}\left(\frac{\Lambda_0}{(T+4)^2} + \frac{\sigma_g\sqrt{\mu} + L_g + \sigma_h}{(T+4)^{1/2}} + \frac{\lambda_0 D_g^2}{(T+4)^{5/2}}\right) + \frac{\sqrt{2}\varepsilon'}{\sqrt{\mu}}.$$

H. 定理 3 的证明(N-HARPG)

我们先更详细地重述定理 3。

定理 7。 设假设 1、3和 4 成立。令 $\gamma_t = \frac{3}{\sqrt{2\mu}(t+2)} = \frac{6M_g}{\mu_F(t+2)}$ 以及 $\eta_t = \frac{2}{t+2}$, $H = \frac{3}{2}(1-\gamma)^{-1}\log(T+1)$。则对每个整数 $T \geq 1$，算法 θ_T 的输出 N-HARPG(见算法 2) 满足

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{J^* - J(\theta_0)}{(T+1)^2} + \frac{D_h}{\mu(T+1)^{3/2}} + \frac{D_g}{T+1} + \frac{\sigma_g\sqrt{\mu} + L_g}{\mu(T+1)} + \frac{\sigma_g\sqrt{\mu} + L_g + \sigma_h}{\mu(T+1)^{1/2}}\right) + \frac{\varepsilon'}{\sqrt{2\mu}},$$

其中 σ_g， σ_h， L_g 分别定义于命题 1 和引理 1 中； D_g 和 D_h 定义于引理 4中， ε' 和 μ 定义于引理 2中。

与针对 N-PG-IGT的定理 1 证明类似，我们从引理 6 给出的类上升结果出发，该结果与更新方向无关(d_t)。根据界（13），我们需要控制误差项

$$V_t = \mathbb{E}[\|d_t - \nabla J_H(\theta_t)\|^2]\,.$$

回顾一下，我们已经在引理 9 中给出了该量的一个界。现在我们可以给出定理 7 的证明。

定理 7 （以及定理 3）的证明结束。 利用引理6 （注意，该引理不使用序列 (d_t) 的特定更新规则，因此可在此处使用），并对 $x \mapsto x^2$应用詹森不等式，我们有

$$\delta_{t+1} \leq \left(1 - \frac{\sqrt{2\mu}\gamma_t}{3}\right)\delta_t + \frac{8\gamma_t}{3}V_t^{1/2} + \frac{L_g\gamma_t^2}{2} + \frac{\varepsilon'\gamma_t}{3} + \frac{4\gamma_t D_g}{3}\gamma^H. \tag{36}$$

然后，利用引理 9 并注意到，根据算法 2（N-HARPG）的归一化更新规则， $\|\theta_{t+1} - \theta_t\| = \gamma_t$ 成立，我们有

$$V_{t+1} - V_t \leq -\eta_{t+1}V_t + 2\sigma_g^2\eta_{t+1}^2 + 12(2L^2 + \sigma_H^2 + D_h^2\gamma^{2H})\gamma_t^2 + 24D_g^2\gamma^{2H}. \tag{37}$$

应用引理 12，令 $r_t = V_t$， $\alpha_t = \eta_{t+1}$， $a = 1$， $\beta_t = c\cdot\eta_t^2 + \bar{c}$， $c = 2\sigma_g^2 + 12(2L_g^2 + \sigma_h^2 + D_h^2\gamma^{2H})\gamma_0^2$， $\bar{c} = 24D_g^2\gamma^{2H}$， $\tau = 2$， $t_0 = 0$，我们得到

$$V_t \leq \frac{V_0}{(t+1)^2} + \frac{\sum_{\tau=0}^{t-1}(c\eta_\tau^2 + \bar{c})(\tau+2)^2}{(t+1)^2} \leq \sigma_g^2\eta_t^2 + 2c\eta_t + 2\bar{c}\eta_t^{-1}\,. \tag{38}$$

结合不等式（36）和（38），可得

$$\begin{aligned}
\delta_{t+1} &\leq (1-\eta_t)\,\delta_t + \frac{8\gamma_0\eta_t}{3}V_t^{1/2} + \frac{L_g\gamma_0^2\eta_t^2}{2} + \frac{\gamma_0\varepsilon'\eta_t}{3} + \frac{4\gamma_t D_g}{3}\gamma^H \\
&\leq (1-\eta_t)\,\delta_t + \frac{8\gamma_0\eta_t}{3}\left(\sigma_g\eta_t + \sqrt{2c}\eta_t^{1/2} + \sqrt{2\bar{c}}\eta_t^{-1/2}\right) + \frac{L_g\gamma_0^2\eta_t^2}{2} + \frac{\gamma_0\varepsilon'\eta_t}{3} + \frac{4\gamma_t D_g}{3}\gamma^H \\
&\leq (1-\eta_t)\,\delta_t + \left(\frac{8\gamma_0\sigma_g}{3} + \frac{L_g\gamma_0^2}{2}\right)\eta_t^2 + \frac{8\sqrt{2c}\gamma_0}{3}\eta_t^{3/2} + \frac{\gamma_0\varepsilon'\eta_t}{3} + \frac{16\sqrt{2\bar{c}}\gamma_0}{3}\eta_t^{1/2}.
\end{aligned}$$

再次应用引理 12 并结合 $r_t = \delta_t$， $\alpha_t = \eta_t$， $a = 1$， $\tau = 2$， $t_0 = 0$， $\beta_t = c_1\eta_t^2 + c_2\eta_t^{3/2} + c_3\eta_t + c_4\eta_t^{1/2}$，

$c_1 = \left(\frac{8\gamma_0\sigma_g}{3} + \frac{L_g\gamma_0^2}{2}\right), c_2 = \frac{8\sqrt{2c}\gamma_0}{3}, c_3 = \frac{\gamma_0\varepsilon'}{3}, c_4 = \frac{16\sqrt{2\bar{c}}\gamma_0}{3}$对每个整数 $T \geq$，我们都有

$$
\begin{aligned}
\delta_T &\leq \frac{\delta_0}{(T+1)^2} + \frac{\sum_{t=0}^{T-1}\beta_t(t+2)^2}{(T+1)^2} \\
&= \frac{\delta_0}{(T+1)^2} + \frac{c_1\sum_{t=0}^{T-1}\eta_t^2(t+2)^2}{(T+1)^2} + \frac{c_2\sum_{t=0}^{T-1}\eta_t^{3/2}(t+2)^2}{(T+1)^2} + \frac{c_3\sum_{t=0}^{T-1}\eta_t(t+2)^2}{(T+1)^2} + \frac{c_4\sum_{t=0}^{T-1}\eta_t^{1/2}(t+2)^2}{(T+1)^2} \\
&\leq \frac{\delta_0}{(T+1)^2} + \frac{4c_1}{T+1} + \frac{4c_2}{(T+1)^{1/2}} + c_3 + 4c_4(T+1)^{1/2} \\
&\leq \frac{\delta_0}{(T+1)^2} + 4\left(\frac{8\sigma_g}{\sqrt{2\mu}} + \frac{9L_g}{4\mu}\right)\frac{1}{T+1} + 64\left(\frac{\sigma_g\sqrt{\mu} + 6L_g + 3\sigma_h + 3D_h\gamma^H}{\mu}\right)\frac{1}{(T+1)^{1/2}} + \frac{\gamma_0\varepsilon'}{3} + 4c_4(T+1)^{1/2} \\
&\leq \frac{\delta_0}{(T+1)^2} + 32\cdot\frac{\sigma_g\sqrt{\mu} + L_g}{\mu(T+1)} + 192\cdot\frac{\sigma_g\sqrt{\mu} + L_g + \sigma_h + D_h\gamma^H}{\mu(T+1)^{1/2}} + \frac{\varepsilon'}{\sqrt{2\mu}} + 150\cdot D_g\gamma^H(T+1)^{1/2}.
\end{aligned}
$$

选择 $H = \frac{3}{2}(1-\gamma)^{-1}\log(T+1)$ 即完成证明。

I. 收敛 性分析：归一化动量策略梯度方法（N-M PG)

本节中，我们给出一种简单的归一化动量策略梯度算法的全局收敛结果，并将其命名为N-MPG （见算法 3）。与具有 $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ 全局样本复杂度的 N-PG-IGT 不同（该算法能够找到一个全局最优策略，误差仅受策略参数化带来的偏差影响），N-MPG 不使用前瞻步，而我们的分析只能保证该算法具有 $\tilde{\mathcal{O}}(\varepsilon^{-3})$全局样本复杂度。

算法 3 N-MPG（归一化动量策略梯度）

1 θ_0 T $\{\eta_t\}_{t\geq 0}$ $\{\gamma_t\}_{t\geq 0}$：
2

输入：，，，： $\tau_0 \sim p(\cdot|\pi_{\theta_0})$ **3:** $d_0 = g$ **(**τ_0, θ_0**) 4:** $\theta_1 = \theta_0 + \gamma_0 \frac{d_0}{\|d_0\|}$ **5: 对于** $t = 1, \ldots, T-1$ **执行 6:** $\tau_t \sim p(\cdot|\pi_{\theta_t})$ **7:** $d_t = (1-\eta_t)d_{t-1} + \eta_t g(\tau_t, \theta_t)$ **8:** $\theta_{t+1} = \theta_t + \gamma_t \frac{d_t}{\|d_t\|}$ **9: 结束循环 10: 返回** θ_T

定理 8。 设假设 *1*、*3*和 *4* 成立。令 $\gamma_t = \frac{6}{\sqrt{2\mu}(t+2)}$、$\eta_t = \left(\frac{2}{t+2}\right)^{2/3}$，$H = \frac{5}{3}(1-\gamma)^{-1}\log(T+1)$。则对每个整数 $T \geq 1$，*N-MPG* （算法 *(Algorithm 3* ） 的输出 θ_T 满足

$$J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \mathcal{O}\left(\frac{J^* - J(\theta_0)}{(T+1)^2} + \frac{D_g}{\sqrt{\mu}(T+1)} + \frac{\sigma_g\sqrt{\mu} + L_g}{\mu(T+1)^{1/3}}\right) + \frac{\varepsilon'}{\sqrt{2\mu}}. \tag{39}$$

其中 σ_g、L_g 由命题 *1*定义，D_g 由引理 *4*定义，ε' 和 μ 由引理 *2*定义。

下面的推论给出了该算法的样本复杂度，即为达到全局最优精度 $\varepsilon > 0$ 所需观测的状态-动作对数量；该精度还受到与策略参数化的函数逼近能力相关的误差下限 $\varepsilon'/\sqrt{2\mu}$ 的影响。

推论 2。 在定理 *8*的设定下，对每个 $\varepsilon > 0$，N-MPG 的输出 θ_T满足 $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + \varepsilon'/\sqrt{2\mu}$，其中 $T = \mathcal{O}(\varepsilon^{-3})$，也就是说，总样本复杂度的阶为 $\tilde{\mathcal{O}}(\varepsilon^{-3})$。

定理的证明*8*。 结合引理 6与引理 11 可得

$$\begin{aligned}
\delta_{t+1} &\leq \left(1 - \frac{\sqrt{2\mu}\gamma_t}{3}\right)\delta_t + \frac{8\gamma_t}{3}\mathbb{E}\left[\|d_t - \nabla J(\theta_t)\|\right] + \frac{L_g\gamma_t^2}{2} + \frac{\gamma_t\varepsilon'}{3} + \frac{4\gamma_t D_g}{3}\gamma^H \\
&\leq \left(1 - \frac{\sqrt{2\mu}\gamma_t}{3}\right)\delta_t + \frac{16}{3}\sqrt{C\left(\frac{4}{3}, \frac{2}{3}\right)}\sigma_g\gamma_t\eta_t^{1/2} + \frac{8}{3}C\left(1, \frac{2}{3}\right)L_g\gamma_t^2\eta_t^{-1} + \frac{L_g\gamma_t^2}{2} + \frac{\gamma_t\varepsilon'}{3} + \frac{4\gamma_t D_g}{3}\gamma^H \\
&\quad + \frac{16}{3}D_g\gamma^H C\left(0, \frac{2}{3}\right)\gamma_t\eta_t^{-1} \\
&\leq \left(1 - \frac{\sqrt{2\mu}\gamma_t}{3}\right)\delta_t + \frac{16}{3}\sqrt{C\left(\frac{4}{3}, \frac{2}{3}\right)}\sigma_g\gamma_t\eta_t^{1/2} + 8C\left(1, \frac{2}{3}\right)L_g\gamma_t^2\eta_t^{-1} + \frac{\gamma_t\varepsilon'}{3} + \frac{4\gamma_t D_g}{3}\gamma^H \\
&\quad + \frac{16}{3}D_g\gamma^H C\left(0, \frac{2}{3}\right)\gamma_t\eta_t^{-1}.
\end{aligned}$$

使用引理 12结合 $\alpha_t = \gamma_t$，$a = \frac{3}{\sqrt{2\mu}}$，$t_0 = 0$，$\tau = 2$，$\beta_t = \frac{16}{3}\sqrt{C\left(\frac{4}{3}, \frac{2}{3}\right)}\sigma_g\gamma_t\eta_t^{1/2} + 8C\left(1, \frac{2}{3}\right)L_g\gamma_t^2\eta_t^{-1} + \frac{\gamma_t\varepsilon'}{3} +$

$\frac{4\gamma_t D_g}{3}\gamma^H + \frac{16}{3}D_g\gamma^H C\left(0, \frac{2}{3}\right)\gamma_t\eta_t^{-1}$可得

$$\begin{aligned}\delta_T &\leq \frac{\delta_0}{(T+1)^2} + \frac{\sum_{t=0}^{T-1}\beta_t(t+2)^2}{(T+1)^2} \\ &\leq \frac{\delta_0}{(T+1)^2} + \frac{16}{3}\sqrt{C\left(\frac{4}{3},\frac{2}{3}\right)}\frac{\sigma_g\gamma_0}{(T+1)^{1/3}} + 16C\left(1,\frac{2}{3}\right)\frac{L_g\gamma_0^2}{(T+1)^{1/3}} + \frac{\gamma_0\varepsilon'}{3} + \frac{4\gamma_0 D_g}{3}\gamma^H \\ &\quad + \frac{16}{3}C\left(0,\frac{2}{3}\right)\gamma_0\gamma^H D_g(T+2)^{2/3}.\end{aligned}$$

令 $H = \frac{5}{3}(1-\gamma)^{-1}\log(T+1)$，即可完成证明。

引理 11. 假设 *1*成立。设 (d_t) 和 (θ_t) 为由 N-MPG 算法生成的序列（见算法 *3*），其中 $\eta_t = \left(\frac{2}{t+2}\right)^{2/3}$和 $\gamma_t = \frac{6}{\sqrt{2\mu}(t+2)}$。那么，对于任意整数 $t \geq 0$，有

$$\mathbb{E}\left[\|d_t - \nabla J_H(\theta_t)\|\right] \leq 2\cdot\sqrt{C\left(\frac{4}{3},\frac{2}{3}\right)}\eta_T^{1/2}\sigma_g + C\left(1,\frac{2}{3}\right)\gamma_T\eta_T^{-1}L_g + 2D_g\gamma^H C\left(0,\frac{2}{3}\right)\eta_T^{-1}, \tag{40}$$

其中 $C(p,q)$ 当 $q\in[0,1)$，$p>0$ 的定义见引理 *15*。

证明。 定义$\hat{e}_t \overset{\text{def}}{=} d_t - \nabla J_H(\theta_t)$ 和 $e_t \overset{\text{def}}{=} g(\tau_t,\theta_t) - \nabla J_H(\theta_t)$。那么，序列 (d_t) 的更新规则意味着

$$\begin{aligned}\hat{e}_t &= d_t - \nabla J_H(\theta_t) \\ &= (1-\eta_t)d_{t-1} + \eta_t g(\tau_t,\theta_t) - \nabla J_H(\theta_t) \\ &= (1-\eta_t)\left(\hat{e}_{t-1} + \nabla J_H(\theta_{t-1})\right) + \eta_t g(\tau_t,\theta_t) - \nabla J_H(\theta_t) \\ &= (1-\eta_t)\hat{e}_{t-1} + \eta_t e_t + (1-\eta_t)\left(\nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t)\right).\end{aligned}$$

展开递推关系并定义 $S_t \overset{\text{def}}{=} \nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t)$，可得

$$\hat{e}_T = \prod_{t=0}^{T-1}(1-\eta_{t+1})\hat{e}_0 + \sum_{t=0}^{T-1}\eta_{T-t}e_{T-t}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1}) + \sum_{t=0}^{T-1}(1-\eta_{T-t})S_{T-t}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1}). \tag{41}$$

对每个整数 $t\geq 1$ 定义 σ-场 $\mathcal{F}_t \overset{\text{定义}}{=} \sigma(\{\theta_0,\tau_0,\ldots,\tau_{t-1}\})$，其中对每个 $0\leq s\leq t-1$，$\tau_s\sim p(\cdot|\pi_{\theta_s})$。注意，对于任意整数 $t_2>t_1\geq 1$，我们有

$$\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle|\mathcal{F}_{t_2}\right]\right] = \mathbb{E}\left[\langle e_{t_1}, \mathbb{E}\left[e_{t_2}|\mathcal{F}_{t_2}\right]\rangle\right] = 0.$$

接着，利用三角不等式，取期望并应用詹森不等式，可得

$$\begin{aligned}
\mathbb{E}\left[\|\hat{e}_T\|\right] &\leq \mathbb{E}\left[\|\hat{e}_0\|\right]\prod_{t=0}^{T-1}(1-\eta_{t+1}) + \mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\eta_{T-t}e_{T-t}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})\right\|\right] \\
&\quad + \mathbb{E}\left[\left\|\sum_{t=0}^{T-1}(1-\eta_{T-t})S_{T-t}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})\right\|\right] \\
&\leq \sigma_g\prod_{t=0}^{T-1}(1-\eta_{t+1}) + \left(\mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\eta_{T-t}e_{T-t}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})\right\|^2\right]\right)^{1/2} \\
&\quad + \sum_{t=0}^{T-1}(1-\eta_{T-t})\mathbb{E}\left[\|S_{T-t}\|\right]\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1}) \\
&\overset{(i)}{\leq} \sigma_g\prod_{t=0}^{T-1}(1-\eta_{t+1}) + \left(\sum_{t=0}^{T-1}\eta_{T-t}^2\mathbb{E}\left[\|e_{T-t}\|^2\right]\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})^2\right)^{1/2} \\
&\quad + \left(\sum_{t=0}^{T-1}(1-\eta_{T-t})\gamma_{T-t-1}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})\right)L_g + 2D_g\gamma^H\left(\sum_{t=0}^{T-1}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})\right) \\
&\leq \left(\prod_{t=0}^{T-1}(1-\eta_{t+1})\right)\sigma_g + \left(\sum_{t=0}^{T-1}\eta_{T-t}^2\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})^2\right)^{1/2}\sigma_g \\
&\quad + \left(\sum_{t=0}^{T-1}(1-\eta_{T-t})\gamma_{T-t-1}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})\right)L_g + 2D_g\gamma^H\left(\sum_{t=0}^{T-1}\prod_{\tau=T-t}^{T-1}(1-\eta_{\tau+1})\right),
\end{aligned}$$

其中，在 (i) 中，我们利用光滑性（命题 1）、 $\|\theta_{t+1}-\theta_t\| = \gamma_t$ 这一事实，并利用独立性 $\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle\right] = 0$，对于任意 $t_1 \neq t_2$。在最后一个不等式中，我们使用 $\mathbb{E}\left[\|e_{T-t}\|^2\right] \leq \sigma_g^2$。

此外，利用引理14和15，我们有

$$\begin{aligned}
\mathbb{E}\left[\|\hat{e}_T\|\right] &\leq \eta_T\sigma_g + \sqrt{C\left(\frac{4}{3},\frac{2}{3}\right)}\eta_T^{1/2}\sigma_g + C\left(1,\frac{2}{3}\right)\gamma_T\eta_T^{-1}L_g + 2D_g\gamma^H C\left(0,\frac{2}{3}\right)\eta_T^{-1} \\
&\leq 2\cdot\sqrt{C\left(\frac{4}{3},\frac{2}{3}\right)}\eta_T^{1/2}\sigma_g + C\left(1,\frac{2}{3}\right)\gamma_T\eta_T^{-1}L_g + 2D_g\gamma^H C\left(0,\frac{2}{3}\right)\eta_T^{-1},
\end{aligned}$$

其中 $C(p,\ q)$ 对于 $q \in [0,1)$，$p > 0$ 定义见引理15。

□

J. 技术引理

J.1. 求解递推关系的引理

下面的辅助引理类似于（Stich，2019年，引理 7），有助于求解我们的递推关系。

引理 12. 设 a为正实数， τ 为正整数， $\{r_t\}_{t\geq 0}$ 为满足以下条件的非负序列：对每个整数 $t \geq 0$

$$r_{t+1} - r_t \leq -a\alpha_t r_t + \beta_t,$$

其中 $\{\alpha_t\}_{t\geq 0}$、 $\{\beta_t\}_{t\geq 0}$ 为非负序列，且对所有 t， $a\alpha_t \leq 1$ 。那么对于 $\alpha_t = \frac{2}{(t+\tau)}$，对每个整数 t_0、$T \geq 1$，我们有

$$r_T \leq \frac{(t_0+\tau-1)^2 r_{t_0}}{(T+\tau-1)^2} + \frac{\sum_{t=0}^{T-1}\beta_t(t+\tau)^2}{(T+\tau-1)^2}.$$

证明。 注意到 $1 - a\alpha_t = \frac{t+\tau-2}{t+\tau}$。于是对所有 $t \geq 0$

$$r_{t+1} \leq \frac{t+\tau-2}{t+\tau} r_t + \beta_t.$$

将两边乘以 $(t+\tau)^2$,，得到

$$\begin{aligned}(t+\tau)^2 r_{t+1} &\leq (t+\tau-2)(t+\tau)r_t + \beta_t(t+\tau)^2 \\ &\leq (t+\tau-1)^2 r_t + \beta_t(t+\tau)^2.\end{aligned}$$

将此不等式从 $t = t_0$ 到 $T-1$ 求和，得到

$$(T+\tau-1)^2 r_T \leq (t_0+\tau-1)^2 r_{t_0} + \sum_{t=t_0}^{T-1}\beta_t(t+\tau)^2,$$

$$r_T \leq \frac{(t_0+\tau-1)^2 r_{t_0}}{(T+\tau-1)^2} + \frac{\sum_{t=0}^{T-1}\beta_t(t+\tau)^2}{(T+\tau-1)^2}.$$

□

引理13。 设 $\{r_t\}_{t\geq 0}$ 是一个非负序列，对每个整数 $t \geq 0$都满足

$$r_{t+1} - r_t \leq -\alpha_t r_t + c\beta_t,$$

其中 $\alpha_t = \left(\frac{2}{t+2}\right)^{2/3}$ 和 $\{\beta_t\}_{t\geq 0}$ 是非负实数序列。那么对每个整数 T

$$r_T \leq \frac{r_0}{(T+1)^{1/3}} + \frac{c\sum_{t=0}^{T-1}\beta_t(t+2)^{1/3}}{(T+1)^{1/3}}.$$

证明。 注意到 $1 - \alpha_t = \frac{(t+2)^{2/3} - 2^{2/3}}{(t+2)^{2/3}}$。然后将递推关系两边乘以 $(t+\tau)^{1/3}$，对所有 $t \geq 0$，得到

$$\begin{aligned}(t+2)^{1/3} r_{t+1} &\leq \left((t+2)^{2/3} - 2^{2/3}\right)(t+2)^{-1/3} r_t + c\beta_t(t+2)^{1/3} \\ &\leq t^{2/3}(t+2)^{-1/3} r_t + c\beta_t(t+2)^{1/3} \\ &\leq (t+1)^{1/3} r_t + c\beta_t(t+2)^{1/3},\end{aligned}$$

其中，第二个不等式成立是因为 $\left((t+2)^{2/3} - 2^{2/3}\right)(t+2)^{-1/3} = t^{2/3}(t+2)^{-1/3}\left(\left(1+\frac{2}{t}\right)^{2/3} - \frac{2^{2/3}}{t^{2/3}}\right) \leq t^{2/3}(t+2)^{-1/3}\left(1 + \frac{4}{3t} - \frac{2^{2/3}}{t^{2/3}}\right) \leq t^{2/3}(t+2)^{-1/3}$。通过裂项求和并除以 $(T+1)^{1/3}$，可得对每个整数 T

$$r_T \leq \frac{r_0}{(T+1)^{1/3}} + \frac{c\sum_{t=0}^{T-1}\beta_t(t+2)^{1/3}}{(T+1)^{1/3}}.$$

□

J.2. 递减步长估计的引理

引理14. 设 $q \in [0,1]$ 且对每个整数 t都有 $\eta_t = \left(\frac{2}{t+2}\right)^q$。那么对每个整数 t 以及任意整数 $T \geq 1$，我们有

$$\eta_t(1-\eta_{t+1}) \leq \eta_{t+1}, \tag{42}$$

$$\prod_{t=0}^{T-1}(1-\eta_{t+1}) \leq \eta_T\,.$$

证明。对于每个整数 t，我们有

$$1-\eta_{t+1} \quad = \quad 1-\left(\frac{2}{t+3}\right)^q \leq 1-\frac{1}{t+3} = \frac{t+2}{t+3} \leq \frac{\eta_{t+1}}{\eta_t}.$$

利用上述结果，我们可以写成

$$\prod_{t=0}^{T-1}(1-\eta_{t+1}) \quad \leq \quad \prod_{t=0}^{T-1}\frac{\eta_{t+1}}{\eta_t} = \frac{\eta_T}{\eta_0} = \eta_T.$$

□

以下引理是（Gadat 等人，2018年，命题 B.1）的推广，对每个整数 $T \geq 1$ （而非仅对 T 充分大的情形）均成立。

引理15。 设 $q \in [0,1)$，$p \geq 0$、$\gamma_0 > 0$，并设 $\eta_t = \left(\frac{2}{t+2}\right)^q$、$\gamma_t = \gamma_0\left(\frac{2}{t+2}\right)^p$对每个整数 t均成立。那么对于任意整数 t和 $T \geq 1$，有

$$\sum_{t=0}^{T-1}\gamma_t\prod_{\tau=t+1}^{T-1}(1-\eta_\tau) \leq C\gamma_T\eta_T^{-1},$$

其中 $C = C(p,q) \overset{\text{定义}}{=} 2^{p-q}(1-q)^{-1}t_0\exp\left(2^q(1-q)t_0^{1-q}\right) + 2^{2p+1-q}(1-q)^{-2}$ 且 $t_0 \overset{\text{定义}}{=}$最大值 $\left\{\left(\frac{p}{1-q}()2^q\right)^{\frac{1}{1-q}}, 2\left(\frac{p-q}{1-q}()^2\right)^{\frac{1}{1-q}}\right\}$。

证明。我们首先将其写为

$$\begin{aligned}
\sum_{t=0}^{T-1}\gamma_t\prod_{\tau=t+1}^{T-1}(1-\eta_\tau) \quad &\leq \quad \sum_{t=0}^{T-1}\gamma_t\exp\left(-\sum_{\tau=t+1}^{T-1}\eta_\tau\right) \\
&= \quad \exp\left(-\sum_{\tau=0}^{T-1}\eta_\tau\right)\sum_{t=0}^{T-1}\gamma_t\exp\left(\sum_{\tau=0}^{t}\eta_\tau\right) \\
&\leq \quad \exp\left(-\sum_{\tau=0}^{T-1}\eta_\tau\right)\sum_{t=0}^{T-1}\gamma_t\exp\left(\int_0^t\eta_{\tau-1}d\tau\right) \\
&\leq \quad \gamma_0 2^p\exp\left(-\sum_{\tau=0}^{T-1}\eta_\tau\right)\sum_{t=0}^{T-1}(t+2)^{-p}\exp\left(2^q(1-q)(t+2)^{1-q}\right).
\end{aligned}$$

对于 $t \geq \max\left\{\left(\frac{p}{1-q}()2^q\right)^{\frac{1}{1-q}}, 2\left(\frac{p-q}{1-q}()^2\right)^{\frac{1}{1-q}}\right\} \overset{\text{def}}{=} t_0$，函数 $\phi(t) = t^{-p}\exp\left(2^q(1-q)t^{1-q}\right)$ 是递增的，因此，相应的和有界于

$$2^{-p}t_0\exp\left(2^q(1-q)t_0^{1-q}\right) \quad + \quad \int_{t_0}^{T+1}\tau^{-p}\exp\left(2^q(1-q)\tau^{1-q}\right)d\tau.$$

注意，通过分部积分可得

$$\begin{aligned}
I &\overset{\text{def}}{=} \int_{t_0}^{T+1} \tau^{-p} \exp\left(2^q(1-q)\tau^{1-q}\right) d\tau \\
&= \int_{t_0}^{T+1} 2^{-q}(1-q)^{-2}\tau^{q-p} d\left(\exp\left(2^q(1-q)\tau^{1-q}\right)\right) \\
&= 2^{-q}(1-q)^{-2}\exp\left(2^q(1-q)(T+1)^{1-q}\right)(T+1)^{q-p} \\
&\quad +2^{-q}(1-q)^{-2}(p-q)\int_{t_0}^{T+1}\tau^{-p+q-1}\exp\left(2^q(1-q)\tau^{1-q}\right)d\tau \\
&= 2^{-q}(1-q)^{-2}\exp\left(2^q(1-q)(T+1)^{1-q}\right)(T+1)^{q-p} \\
&\quad +2^{-q}(1-q)^{-2}(p-q)t_0^{-(1-q)}\int_{t_0}^{T+1}\tau^{-p}\exp\left(2^q(1-q)\tau^{1-q}\right)d\tau \\
&\le 2^{-q}(1-q)^{-2}\exp\left(2^q(1-q)(T+1)^{1-q}\right)(T+1)^{q-p} + \frac{1}{2}I,
\end{aligned}$$

其中 th e 最后一个不等式在选择 t_0 时满足，条件为 $p \ge q$，而对于 $p \le q$ 则显然成立。Th us,

$$I \le 2^{1-q}(1-q)^{-2}\exp\left(2^q(1-q)(T+1)^{1-q}\right)(T+1)^{q-p}.$$

最后，将所有结果综合起来，我们得到

$$\begin{aligned}
\sum_{t=0}^{T-1}\gamma_t\prod_{\tau=t+1}^{T-1}(1-\eta_\tau) &\le \gamma_0 2^p\exp\left(-2^q(1-q)(T+1)^{1-q}\right)2^{-p}t_0\exp\left(2^q(1-q)t_0^{1-q}\right) \\
&\quad +\gamma_0 2^{p+1-q}(1-q)^{-2}(T+1)^{q-p} \\
&\le \gamma_0\left(2^{-q}(1-q)^{-1}t_0\exp\left(2^q(1-q)t_0^{1-q}\right)+2^{p+1-q}(1-q)^{-2}\right)(T+1)^{q-p} \\
&\le C\gamma_T\eta_T^{-1},
\end{aligned}$$

其中 $C = C(p,\ q)$ 定义为 $= 2^{p-q}(1-q)^{-1}t_0\exp\left(2^q(1-q)t_0^{1-q}\right) + 2^{2p+1-q}(1-q)$ $^{-2}$。 □