Stochastic Policy Gradient Methods: Improved Sample Complexity for Fisher-non-degenerate Policies

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

Abstract

Recently, the impressive empirical success of policy gradient (PG) methods has catalyzed the development of their theoretical foundations. Despite the huge efforts directed at the design of efficient stochastic PG-type algorithms, the understanding of their convergence to a globally optimal policy is still limited. In this work, we develop improved global convergence guarantees for a general class of Fisher-non-degenerate parameterized policies which allows to address the case of continuous state action spaces. First, we propose a Normalized Policy Gradient method with Implicit Gradient Transport (N-PG-IGT) and derive a $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ sample complexity of this method for finding a global ε-optimal policy. Improving over the previously known $\tilde{\mathcal{O}}(\varepsilon^{-3})$ complexity, this algorithm does not require the use of importance sampling or second-order information and samples only one trajectory per iteration. Second, we further improve this complexity to $\tilde{\mathcal{O}}(\varepsilon^{-2})$ by considering a Hessian-Aided Recursive Policy Gradient ((N)-HARPG) algorithm enhanced with a correction based on a Hessian-vector product. Interestingly, both algorithms are (i) simple and easy to implement: single-loop, do not require large batches of trajectories and sample at most two trajectories per iteration; (ii) computationally and memory efficient: they do not require expensive subroutines at each iteration and can be implemented with memory linear in the dimension of parameters.

1. Introduction

Dating back to the works of Williams (1992); Sutton et al. (1999), Policy Gradient (PG) methods constitute one of the most popular and efficient classes of Reinforcement Learning (RL) algorithms. Combined with deep neural networks, modern PG algorithms (Silver et al., 2014; Schulman et al., 2015; 2017) have shown impressive empirical success in several challenging tasks involving large and even continuous state and action spaces.

In this class of methods, in order to learn a parameterized policy, the agent performs stochastic gradient ascent in the policy parameter space to maximize its expected return $J(\theta)$, where $\theta \in \mathbb{R}^d$ are parameters of the policy parameterization. Given the non-concavity of the policy optimization objective, several works in the literature have focused on the convergence of the policy parameter to an ε-approximate first-order stationary point (FOSP), i.e., $\|\nabla J(\theta)\| \leq \varepsilon$. This performance measure allows to define the sample complexity of PG methods as the number of random trajectory simulations (or samples) required to find an ε-approximate FOSP of the expected return function. In particular, using a similar analysis to stochastic gradient descent in nonconvex smooth optimization (Ghadimi and Lan, 2013), it has been shown that Vanilla-PG (REINFORCE) enjoys a $\tilde{\mathcal{O}}(\varepsilon^{-4})$ sample complexity (see for e.g., Yuan et al. (2022b)). In the last few years, spurred by the development of variance-reduction techniques in stochastic optimization, there has been a tremendous amount of work in designing variance-reduced variants of the PG method to improve sample efficiency. Unlike in classical supervised learning, the sampling distribution is non-stationary in RL due to policy update. Therefore, adapting variance-reduction techniques to PG methods requires to carefully address this non-stationarity. The majority of works address this issue by introducing importance sampling (IS) weights to correct the distribution shift. However, the use of the IS mechanism often requires a strong unverifiable assumption stating that the variance of IS weights is bounded. Recently, Shen et al. (2019); Salehkaleybar et al. (2022) proposed a promising alternative to IS, which incorporates the second-order information to correct the distribution shift due to the policy update. In terms of sample complexity, both approaches guarantee to find an ε-FOSP with $\tilde{\mathcal{O}}(\varepsilon^{-3})$ samples. For stochastic optimization, this complexity was recently shown to be optimal and unimprovable even when higher order stochastic oracle and smoothness are available (Arjevani et al., 2020). There-

[1] Department of Computer Science, ETH Zurich, Switzerland [2] Department of Mathematics, ETH Zurich, Switzerland. Correspondence to: I.F. <ilyas.fn979@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

fore, one needs to utilize the special structure of the policy optimization problem to improve the sample complexity of PG methods.

	Vanilla-PG	N-PG-IGT	(N)-HARPG
FOSP	$\tilde{\mathcal{O}}(\varepsilon^{-4})$	$\tilde{\mathcal{O}}(\varepsilon^{-3.5})$ (new)[1]	$\tilde{\mathcal{O}}(\varepsilon^{-3})$
Global	$\tilde{\mathcal{O}}(\varepsilon^{-3})$	$\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ (new)	$\tilde{\mathcal{O}}(\varepsilon^{-2})$ (new)

[1] The method based on implicit gradient transport was previously studied in stochastic optimization for finding FOSP (Cutkosky and Mehta, 2020), however, we are not aware of its application in RL.

Table 1: Summary of complexity results for *computationally efficient* PG methods. The first row of the table reports the number of samples for finding an ε-FOSP, and the second one is for ε-global optimum under Assumptions 3 and 4.

More recently, a line of research focused on establishing global optimality convergence rates and deriving the sample complexity to reach an ε-approximate global optimum of the expected return function (i. e., $J^* - J(\theta) \leq \varepsilon$ where J^* is the optimal return) (Agarwal et al., 2021; Fazel et al., 2018; Fatkhullin and Polyak, 2021; Bhandari and Russo, 2019; Mei et al., 2020; Zhang et al., 2020b; 2021b) instead of a mere ε-approximate FOSP. This growing body of work leverages additional structure of the policy optimization problem under the form of gradient dominance (Agarwal et al., 2021), Łojasiewicz-like inequalities (Mei et al., 2020; 2021b) or hidden convexity (Zhang et al., 2021a). As previously exploited in deterministic optimization (Lojasiewicz, 1963; Polyak, 1963; Kurdyka, 1998) (see also for e.g., Attouch and Bolte (2009)) and later on in stochastic optimization (Fontaine et al., 2021; Fatkhullin et al., 2022; Scaman et al., 2022; Masiha et al., 2022), these structural properties guarantee that the non-concave policy optimization objective has no suboptimal stationary points and allow to derive convergence rates for the function value. Most of these works assume the access to exact policy gradients to show fast convergence rates (Mei et al., 2020; 2021b; Xiao, 2022).

In contrast, the practically important case of *stochastic* gradients, is addressed only in a few recent works (Liu et al., 2020; Ding et al., 2022; Yuan et al., 2022b; Masiha et al., 2022). This line of works establishes global convergence guarantees of the expected return (up to a bias induced by the policy parameterization). A key tool for their analysis is a relaxed weak gradient dominance inequality satisfied by the expected return function for the class of *Fisher-non-degenerate parameterized policies* (FND) which includes Gaussian policies as a canonical example. Drawing on this structural property, Liu et al. (2020); Ding et al. (2022); Yuan et al. (2022b) achieve a $\tilde{\mathcal{O}}(\varepsilon^{-3})$ global optimality rate up to a bias error term due to policy parameterization. Very recently, Masiha et al. (2022) further improve this complexity to $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ by analyzing a stochastic second-order method. However, their algorithm needs to approximately solve a cubic sub-problem, which requires additional $\tilde{\mathcal{O}}(\varepsilon^{-3})$ operations at each iteration and introduces extra computational burden. A natural question that emerges from these recent results is the following:

> *Is it possible to improve $\tilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity for general FND parameterized policies using a computationally efficient PG algorithm?*

In this work, we answer this question in the affirmative. Our contributions are summarized as follows.

1.1. Summary of contributions

- First, we propose a normalized momentum-based PG method with implicit gradient transport (N-PG-IGT) and establish a $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ global convergence sample complexity. Interestingly, this algorithm does not require the IS mechanism, does not use curvature information such as Hessian estimates and only samples a single trajectory at each iteration.

- Next, we consider two variants of Hessian-aided momentum-based PG method: (i) N-HARPG – with normalization, and (ii) HARPG – without normalization. We offer a refined convergence analysis for both variants and establish the improved global convergence achieving a $\tilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity. Unlike most prior work, our analysis does not require the strong assumption of boundedness of IS weights since HARPG does not use IS. Moreover, these algorithms are single-loop, only require to sample two trajectories per iteration and can be implemented efficiently with memory and computational complexity similar to Hessian-free methods.

We highlight that our improved sample complexities are derived under the same (or less restrictive) assumptions compared to previous related work in this setting (Liu et al., 2020; Ding et al., 2021; Yuan et al., 2022b; Masiha et al., 2022). We refer the reader to Tables 1 and 2 for a contextualization of our contributions in the literature.

1.2. Related work

We briefly review the literature most closely related to our results about the global convergence of stochastic PG methods. We defer a more detailed literature review including a discussion of first-order stationarity of variance-reduced PG methods and global optimality of exact PG methods to Appendix D. For convenience, we refer the reader to Table 2 for an overview of related works and our main contributions.

Only a few recent works provide global convergence guarantees for the case of stochastic policy gradients (Liu et al., 2020; Zhang et al., 2021a; Ding et al., 2021; 2022; Yuan et al., 2022b; Masiha et al., 2022).

Existing results under our setting. For the class of *Fisher-non-degenerate* parameterized policies, Liu et al. (2020) propose a stochastic variance-reduced and Natural PG methods for which they establish a $\tilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity for finding an ε-globally optimal policy (up to a compatible function approximation error due to policy parameterization). Later, Ding et al. (2022) achieve the same sample complexity using a single-loop and finite-batch momentum-based PG algorithm. However, the above works require the strong assumption of bounded IS weights variance. Recently, Yuan et al. (2022b) showed the same $\tilde{\mathcal{O}}(\varepsilon^{-3})$ sample complexity for stochastic Vanilla-PG for the same policy class under similar assumptions. Very recently, Masiha et al. (2022) propose a stochastic cubic regularized method (SCRN) achieving a $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ sample complexity under similar assumptions. Nevertheless, at each iteration, their method needs a heavy subroutine to solve the cubic regularized sub-problem and requires large batches of samples to estimate the gradient and the Hessian (where batch-sizes depend on the dimension d and the desired accuracy ε, see Lemma 11 therein). Masiha et al. (2022) further improve the complexity to a $\tilde{\mathcal{O}}(\varepsilon^{-2})$ considering a variance-reduced variant of SCRN (see Algorithm 2 therein) for stochastic optimization. However, this algorithm also requires large batch-sizes. Moreover, adapting this method to RL is not trivial, since it would require introducing IS weights for both stochastic gradients and Hessians and assuming a bound on their variance to account for non-stationarity inherent to RL.

Global sample complexity for tabular and soft-max policies. In the tabular setting, Lan (2022) shows a $\tilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity for a stochastic policy mirror descent method. However, their analysis is specific to the tabular setting with discrete state action spaces where no parameterization is used. Under the soft-max policy parameterization, Zhang et al. (2021a) obtain a global ε-optimal policy using $\tilde{\mathcal{O}}(\varepsilon^{-2})$ samples via a double-loop variance-reduced PG algorithm with large batch sizes. Interestingly, this algorithm implements a gradient truncation mechanism to relax the boundedness assumption on the IS weights variance by utilizing the special structure of the softmax policy parameterization. We highlight that Zhang et al. (2021a) focus on the softmax policy parameterization only and consider a hard-to-verify overparameterization assumption (Assumption 5.11) to exploit hidden convexity. Very recently, when finalizing the preparation of this work, we came across the works of Alfano and Rebeschini (2022); Yuan et al. (2022a) analyzing the convergence of NPG and its Q-version (Q-NPG) for log-linear policies. In particular, Yuan et al. (2022a) achieves a $\tilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity using the framework of compatible function approximation. Furthermore, at each iteration, NPG and Q-NPG require using SGD as a subroutine to solve a regression problem to find the update direction and rely on complex sampling subprocedures (Algorithms 3, 4 p. 28-30). Complementing all these works considering direct and softmax parameterization for discrete and finite state action spaces, our analysis addresses the case of continuous state action spaces via the FND policy class, which is a fundamentally different problem setting.

2. Preliminaries

2.1. Problem formulation

We consider a discrete-time discounted Markov Decision Process (MDP) (Puterman, 2014) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are (possibly continuous) state and action spaces respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$ is the state transition density, $r : \mathcal{S} \times \mathcal{A} \to [-r_{\max}, r_{\max}]$ is the reward function which is bounded by $r_{\max} > 0$, ρ is the initial state distribution, and $\gamma \in (0,1)$ is the discount factor. A parametrized policy is defined for every parameter $\theta \in \mathbb{R}^d$ and state $s \in \mathcal{S}$ as a probability density function $\pi_\theta(\cdot|s)$ over the action space $\mathcal{A}$. At each time step $t \in \mathbb{N}$ in a state $s_t \in \mathcal{S}$, the RL agent chooses an action $a_t \in \mathcal{A}$ according to the distribution $\pi_\theta(\cdot|s_t)$, observes a reward $r(s_t,a_t)$ and the environment transitions to a state s_{t+1} in a region $\mathcal{S}^{'} \subset \mathcal{S}$ according to $\mathcal{P}(\cdot|s_t, a_t)$.

The goal of the RL agent is to find a policy π_θ maximizing the expected return defined by:

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

where the expectation is taken with respect to the distribution of the Markov chain $(s_t,a_t)_{t\in\mathbb{N}}$.[1] The agent only has access to trajectories of finite length H generated from the MDP, whereas the state transition kernel $\mathcal{P}$ and the reward function $r(\cdot,\cdot)$ are unknown. The probability distribution induced by the initial distribution ρ and a policy π_θ on the space of trajectories of length $H \geq 1$ has a density:

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

where $\tau = (s_0, a_0, \cdots, s_{H-1}, a_{H-1})$. We also define the truncated expected return given the horizon H by $J_H(\theta) \overset{\text{def}}{=} \mathbb{E}_{\rho,\pi_\theta}\left[\sum_{t=0}^{H-1} \gamma^t r(s_t,a_t)\right]$.

[1] In the rest of the paper, we will often omit the subscript in this expectation where it is clear from the context.

Algorithm	**Sample complexity***	**Last iterate**	**No IS**	**No 2nd smooth.**	**No batch**	**Comp. efficient**
Vanilla-PG (Yuan et al., 2022b)	$\frac{1}{(1-\gamma)^5}\varepsilon^{-3}$	✔	✔	✔	✔	✔
Natural-PG (Liu et al., 2020)	$\frac{1}{(1-\gamma)^6}\varepsilon^{-3}$	✗	✔	✔	✗	✗
SRVR-PG (Liu et al., 2020)	$\frac{1+W(1-\gamma)^3}{(1-\gamma)^{5.5}}\varepsilon^{-3}$	✗	✗	✔	✗	✔
STORM-PG-F (Ding et al., 2022)	$\frac{1+W^{3/2}}{(1-\gamma)^{12}}\varepsilon^{-3}$	✗	✗	✔	✔	✔
SCRN (Masiha et al., 2022)	$\frac{1}{(1-\gamma)^{7.5}}\varepsilon^{-2.5}$	✔	✔	✗	✗	✗
N-PG-IGT This paper	$\frac{1}{(1-\gamma)^{7.5}}\varepsilon^{-2.5}$	✔	✔	✗	✔	✔
(N)-HARPG This paper	$\frac{1}{(1-\gamma)^6}\varepsilon^{-2}$	✔	✔	✔	✔	✔

Table 2: Overview of related works for global optimality convergence for FND policies. "**Sample complexity**" reports the number of sample trajectories (up to a logarithmic factor) required to achieve the desired accuracy, i.e., sample complexity to achieve $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}$ when the guarantee is for the "**Last iterate**" and $\min_{t=0,\cdots,T} J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma}$ for the "best" iterate (see Assumption 4 for a definition of $\varepsilon_{\text{bias}}$). "**No IS**" means that importance sampling (IS) is not used in the method. If the algorithm uses IS, then it is additionally assumed that the variance of IS weights is bounded, i.e., $\mathbb{V}\left(\frac{p(\tau|\theta)}{p(\tau|\theta')}\right) \leq W, \forall\, \theta, \theta' \in \mathbb{R}^d$. "**No 2nd smooth.**" means second-order smoothness of the policy parameterization (Assumption 2) is not needed. "**No batch**" means that the algorithm does not require mini-batch. We refer to a method as "**Comp. efficient**" if at each iteration it requires only $\mathcal{O}\left(d\right)$ number of arithmetic operations. In contrast, SCRN and Natural-PG require additional subroutines, see Masiha et al. (2022, Remark 6) and Liu et al. (2020, Theorem 4.9).

* Sample complexity hides the dependence on constants M_g, M_h, l_2, r_{max}, μ_F and poly-logarithmic factors in ε^{-1}.

2.2. First order stationarity and global optimality

The policy optimization problem is differentiable and smooth (under some regularity conditions we precise later on), but non-concave in general. Therefore, several prior works (see Appendix for a review) focused on showing convergence to an ε-approximate FOSP of the expected return function J, i.e. to a parameter θ such that (s.t.) $\|\nabla J(\theta)\| \leq \varepsilon$. In this paper, we focus on finding a parameter θ s.t. $J^* - J(\theta) \leq \varepsilon + \mathcal{O}(\sqrt{\varepsilon_{\text{bias}}})$, where J^* is the optimal expected return (maximizing $J(\pi)$) and the $\mathcal{O}(\sqrt{\varepsilon_{\text{bias}}})$ error term is due to the approximation power of the policy parameterization (see Assumption 4).

2.3. Policy gradient and Hessian

Under mild regularity assumptions on the policy parameterization, it can be shown that the gradient of the truncated expected return function J_H w.r.t. (with respect to) the policy parameter θ and the Hessian at θ (see (Shen et al., 2019)) can be written:

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

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

with

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

and τ is an H-length simulated trajectory from the MDP with policy π_θ and initial distribution ρ. Note that all the derivatives are w.r.t. θ and the full gradient of J cannot be computed due to the infinite horizon length. An unbiased estimator of the gradient $\nabla J_H(\theta)$ above is then given by $g(\tau,\theta)$ as defined in Eq. 2.

3. Normalized Momentum-Based Policy Gradient Algorithms

In this section, we present our two momentum-based policy gradient algorithms: N-PG-IGT and N-HARPG. These methods share two key algorithmic features: *normalization* and *momentum*. Moreover, while N-PG-IGT uses a look-ahead mechanism to leverage second-order smoothness, N-HARPG uses variance reduction with Hessian correction in order to improve sample complexity.

Algorithm 1 N-PG-IGT
(Normalized-PG-with Implicit Gradient Transport)

```
1: Input: $\theta_0$, $\theta_1$, $d_0$, $T$, $\{\eta_t\}_{t\geq 0}$, $\{\gamma_t\}_{t\geq 0}$
2: for $t = 1, \dots, T-1$ do
3:     $\tilde{\theta}_t = \theta_t + \frac{1-\eta_t}{\eta_t}(\theta_t - \theta_{t-1})$
4:     $\tilde{\tau}_t \sim p(\cdot|\pi_{\tilde{\theta}_t})$
5:     $d_t = (1-\eta_t)d_{t-1} + \eta_t g(\tilde{\tau}_t, \tilde{\theta}_t)$
6:     $\theta_{t+1} = \theta_t + \gamma_t \frac{d_t}{\|d_t\|}$
7: end for
8: return $\theta_T$
```

3.1. Normalized PG with Implicit Gradient Transport

We start by introducing a policy gradient algorithm for policy optimization called Normalized-PG-with Implicit Gradient Transport (N-PG-IGT for short). We propose the N-PG-IGT algorithm for solving our RL policy optimization problem, inspired by the recently proposed NIGT method which was considered in the context of stochastic optimization (with i.i.d samples from a fixed distribution) (Cutkosky and Mehta, 2020; Arnold et al., 2019).

Motivation. Before describing the N-PG-IGT algorithm, we provide some motivation. When having access to exact policy gradients, simple normalized PG enjoys a geometric global convergence rate (Mei et al., 2021b) thanks to the non-uniform smoothness of the policy optimization objective. It is then natural to investigate the properties of this algorithm in the more realistic setting where only stochastic policy gradients can be computed using samples from the MDP. In this stochastic setting, the understanding of this method is limited. Therefore, it is instructive to discuss known results from stochastic optimization regarding the normalized stochastic gradient method. As explained by Cutkosky and Mehta (2020, Sec. 2), the issue with the simple normalized stochastic gradient method is that the gradient estimate error before normalization may be larger than the true gradient. To fix this issue, several works in the literature propose to reduce the variance of the stochastic gradients either by using very large batch sizes (see for e.g. (Hazan et al., 2015)) or by using momentum without resorting to large batches as recently suggested in (Cutkosky and Mehta, 2020). Back to our policy optimization problem, we investigated the sample complexity of the normalized PG with momentum for global optimality beyond the first-order stationarity convergence rates. However, our results for this algorithm only match the performance of vanilla PG. We defer the discussion of these results to Appendix I, where we formally introduce this Algorithm 3 and prove convergence guarantees in Theorem 8. Given these elements of motivation, we can now present the N-PG-IGT algorithm which also utilizes momentum with normalization and will allow us to derive improved convergence rates.

N-PG-IGT **(Algorithm 1).** This algorithm performs a normalized gradient ascent in the policy parameter space following the direction of a normalized moment-based stochastic policy gradient. Unlike in a simple normalized momentum PG algorithm, the direction d_t is computed using the stochastic policy gradient $g(\tilde{\tau}_t, \tilde{\theta}_t)$ instead of $g(\tau_t,\theta_t)$ with $\tau_t \sim p(\cdot|\pi_{\theta_t})$. Here, the auxiliary sequence $\tilde{\theta}_t$ is crucial for faster convergence. Observe that its update rule (Step 3) can be rewritten $\theta_t = \eta_t\tilde{\theta}_t + (1-\eta_t)\theta_{t-1}$ which shows that using $\tilde{\theta}_t$ can be seen as performing a *lookahead* step extrapolating from previous iterates θ_t and θ_{t-1}. We refer the reader to (Cutkosky and Mehta, 2020, Sec. 3) for more intuition regarding this scheme. Notice that the algorithm does not implement any importance sampling mechanism to account for the non-stationarity of our RL problem (which makes RL different from standard stochastic optimization). The algorithm samples a single trajectory at each iteration and does not require large batches.

3.2. Hessian-Aided Recursive Policy Gradient

We now present two variants of Hessian-Aided policy gradient algorithms, namely Hessian-Aided Recursive Policy Gradient (HARPG) and its normalized version (N-HARPG). We remark here that N-HARPG was previously proposed in Salehkaleybar et al. (2022) under the name SHARP [2].

(N)-HARPG **(see Algorithm 2).** The algorithm is based on an idea similar to the STORM variance reduction method (Cutkosky and Orabona, 2019) and uses second-order information (Tran and Cutkosky, 2021) instead of the difference between consecutive stochastic gradients. Interestingly, this allows to bypass the use of IS and the resulting need for controlling IS weights via unverifiable assumptions. To compute the update direction d_t of the algorithm in step 7, a second-order correction $(1-\eta_t)v_t$ is added to the momentum stochastic gradient $(1-\eta_t)d_{t-1}+\eta_t g(\tau_t,\theta_t)$. Crucially, the uniform sampling procedure in steps 3-6 guarantees that v_t is an unbiased estimator of $\nabla J(\theta_t) - \nabla J(\theta_{t-1})$, a term which is reminiscent of the one in the original STORM.The last step of the algorithm (step 8) uses the direction d_t with or without normalization to update the parameter θ_t .

Hessian-vector implementation (Step 6). We insist on the fact that the Hessian-vector product defining the correction v_t can be easily and efficiently implemented in $\mathcal{O}(Hd)$ computational complexity. Observe for this that for any arbitrary vector $u \in \mathbb{R}^d$, any trajectory τ and any parameter $\theta \in \mathbb{R}^d$, we can write the Hessian-vector prod-

[2] We use a different name for this algorithm since we are also studying the algorithm without normalization (namely HARPG) and we use different step-sizes and momentum paramemeters for the normalized algorithm in our analysis.

Algorithm 2 (N)-HARPG
((Normalized)-Hessian-Aided Recursive Policy Gradient)

```
Input: $\theta_0$, $\theta_1$, $d_0$, $T$, $\{\eta_t\}_{t\geq 0}$, $\{\gamma_t\}_{t\geq 0}$
for $t = 1, \ldots, T-1$ do
    $q_t \sim \mathcal{U}([0,1])$
    $\hat{\theta}_t = q_t\theta_t + (1-q_t)\theta_{t-1}$
    $\tau_t \sim p(\cdot|\pi_{\theta_t})$; $\hat{\tau}_t \sim p(\cdot|\pi_{\hat{\theta}_t})$
    $v_t = B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})$
    $d_t = (1-\eta_t)\,(d_{t-1} + v_t) + \eta_t g(\tau_t, \theta_t)$
    $\theta_{t+1} = \begin{cases} \theta_t + \gamma_t d_t & \text{(HARPG)} \\ \theta_t + \gamma_t \frac{d_t}{\|d_t\|} & \text{(N-HARPG)} \end{cases}$
end for
return $\theta_T$
```

uct $B(\tau,\theta)\,u$ using Eq. 3 as follows:

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

The second term can be easily computed via automatic differentiation of the scalar quantity $\langle g(\tau,\theta), u\rangle$ instead of using a finite difference approximation as in prior work (Shen et al., 2019; Huang et al., 2020). Computing and storing the full Hessian estimate $B(\hat{\tau}_t,\hat{\theta}_t)$ is not needed. Exploiting curvature information from the policy Hessian does not compromise the $\mathcal{O}(Hd)$ per-iteration computation cost.

4. Global Convergence Analysis

In this section, we establish the sample complexity of our algorithms to reach global optimality, i.e., the number of samples needed to achieve an ε-approximate global optimum in expectation up to a bias due to policy parameterization. To provide such global optimality guarantees, we leverage a structural property of the non-concave policy optimization problem under the form of a relaxed weak gradient dominance condition satisfied by the expected return J.

4.1. Assumptions

Our first assumption is a standard assumption about the regularity of the policy parameterization.[3]

Assumption 1 (Policy parameterization regularity)**.** *For every $s, a \in \mathcal{S} \times \mathcal{A}$, the function $\theta \mapsto \pi_\theta(a|s)$ is positive and continuously differentiable. Moreover,*

(i) there exists $M_g > 0$ such that for every $\theta \in \mathbb{R}^d, s \in \mathcal{S}, a \in \mathcal{A}$, $\|\nabla \log \pi_\theta(a|s)\| \leq M_g$.

(ii) the function $\theta \mapsto \pi_\theta(a|s)$ is twice continuously differentiable for every $s, a \in \mathcal{S} \times \mathcal{A}$, and there exists $M_h > 0$ such that for every $\theta \in \mathbb{R}^d$ and every $s \in \mathcal{S}, a \in \mathcal{A}$, $\|\nabla^2 \log \pi_\theta(a|s)\| \leq M_h$.

Under this assumption, we collect some known results from the literature (see for e.g., (Papini et al., 2018; Shen et al., 2019; Huang et al., 2020; Xu et al., 2020a; Yuan et al., 2022b)) regarding properties of the objective function J and its stochastic policy gradient in the next proposition.

Proposition 1. *Under Assumption 1-(i),*

(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.

Items (i) and (ii) are stated in Yuan et al. (2022b, Lemma 4.2, 4.4), item (iii) in Shen et al. (2019, Lemma 4.1).

The following assumption about the second-order regularity of the policy parameterization will be useful to establish some of our results. This assumption is also used in several other works (see for e.g., (Zhang et al., 2020b; Masiha et al., 2022; Yang et al., 2021)).

Assumption 2. *For every $(s,a) \in \mathcal{S} \times \mathcal{A}$, the function $\theta \mapsto \pi_\theta(a|s)$ is positive, twice continuously differentiable and there exists $l_2 > 0$ such that for every $\theta, \theta' \in \mathbb{R}^d$ and every $s \in \mathcal{S}, a \in \mathcal{A}$,*

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

Reinforcing our regularity assumption 1 with Assumption 2, the expected return J will further have a Lipschitz Hessian. Notice that both assumptions can be satisfied by a Gaussian policy for instance.

Lemma 1 (Zhang et al. (2020b))**.** *Let Assumptions 1 and 2 hold. Then for every $\theta, \theta' \in \mathbb{R}^d$, $\|\nabla^2 J(\theta) - \nabla^2 J(\theta')\| \leq L_h \|\theta - \theta'\|$, with $L_h = \mathcal{O}\left((1-\gamma)^{-3}\right)$.*

In this paper, we focus on the general class of Fisher-non-degenerate parameterized policies for our convergence analysis. To state our assumption, we define the Fisher information matrix $F_\rho(\theta)$ induced by a policy π_θ for some $\theta \in \mathbb{R}^d$ and an initial distribution ρ as follows:

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

where $d_\rho^{\pi_\theta}(\cdot) \overset{\text{def}}{=} (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\rho,\pi_\theta}(s_t \in \cdot)$ is the discounted state visitation measure.

Assumption 3 (Fisher-non-degenerate policy)**.** *There exists $\mu_F > 0$ s.t. for all $\theta \in \mathbb{R}^d$, $F_\rho(\theta) \succcurlyeq \mu_F \cdot I_d$.*

[3] See Appendix B for a more detailed discussion of all our assumptions.

Assumption 3 is satisfied by Gaussian policies $\pi_\theta(\cdot|s) = \mathcal{N}(\mu_\theta(s),\Sigma)$ when the parameterized mean $\mu_\theta(s)$ has a full-row rank Jacobian and the covariance matrix Σ is fixed. It also holds for a broader subclass of exponential family parameterized policies (see Ding et al. (2022, Sec. 8) for more details) and for certain neural policies (see also Liu et al. (2020, Sec. B.2)). This assumption is common in the literature (Zhang et al., 2020b; Liu et al., 2020; Ding et al., 2022; Yuan et al., 2022b; Masiha et al., 2022). We highlight that the global convergence results we will derive in this paper may not hold for softmax parameterization which does not satisfy Assumption 3 as soon as it gets closer to a deterministic policy.

Following prior work (Agarwal et al., 2021; Liu et al., 2020; Ding et al., 2022; Yuan et al., 2022b), we introduce an assumption reflecting the expressivity of our policy parameterization class via the framework of compatible function approximation (Sutton et al., 1999; Agarwal et al., 2021).

Assumption 4. *There exists $\varepsilon_{bias} \geq 0$ s.t. for every $\theta \in \mathbb{R}^d$, the transfer*[4] *error satisfies:*

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

where A^{π_θ} is the advantage function[5]*, $w^*(\theta) \overset{def}{=} F_\rho(\theta)^\dagger \nabla J(\theta)$ where $F_\rho(\theta)^\dagger$ is the pseudo-inverse of the matrix $F_\rho(\theta)$ and expectation is taken over $s \sim d_\rho^{\pi^*}, a \sim \pi^*(\cdot|s)$ where π^* is an optimal policy (maximizing $J(\theta)$).*

The common Assumption 4 states that the policy parameterization π_θ allows to nearly approximate the advantage function A^{π_θ} from the score function $\nabla \log \pi_\theta$. Importantly, we precise that $\varepsilon_{\text{bias}}$ is positive for a policy parameterization π_θ that does not cover the set of all stochastic policies and $\varepsilon_{\text{bias}}$ is small for a rich neural policy (Wang et al., 2020). We refer the reader to the aforementioned references for further details regarding Assumption 4.

Combining Assumptions 3 and 4, and following the derivations of Ding et al. (2022), we can obtain the following relaxed weak gradient dominance inequality.

Lemma 2 (Relaxed weak gradient domination, (Ding et al., 2022))**.** *Let Assumptions 1-(i), 3 and 4 hold. Then*

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

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

This result is the key tool in convergence analysis of our algorithms. See Appendix C for more details and the proof sketch.

4.2. Convergence analysis of N-PG-IGT

Our first main result characterizes the sample complexity of the N-PG-IGT algorithm to produce an ε-globally optimal policy up to a bias error term due to the approximation power of the policy parameterization.[6]

Theorem 1 (**Global convergence of** N-PG-IGT)**.** *Let Assumptions 1 to 4 hold. Set $\gamma_t = \frac{6M_g}{\mu_F(t+2)}, \eta_t = \left(\frac{2}{t+2}\right)^{4/5}$ and $H = (1-\gamma)^{-1}\log(T+1)$. Then for every integer $T \geq 1$, the output θ_T of* N-PG-IGT *satisfies*

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

where σ_g, L_h are defined in Proposition 1 and Lemma 1 respectively. Hence, for every $\varepsilon > 0$, the output θ_T satisfies $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + \sqrt{\varepsilon_{bias}}/(1-\gamma)$ for $T = \mathcal{O}(\varepsilon^{-2.5})$, i.e., the total sample complexity is of the order $\widetilde{\mathcal{O}}(\varepsilon^{-2.5})$.

Our analysis of N-PG-IGT crucially relies on second-order smoothness of $J(\theta)$, which stems from Lemma 1. N-PG-IGT improves over the best known sample complexity of (stochastic) Vanilla-PG (Yuan et al., 2022b) and STORM-PG-F (Ding et al., 2022) by a factor of $\mathcal{O}(\varepsilon^{-0.5})$ for Fisher-non-degenerate policies. The recently proposed SCRN algorithm (Masiha et al., 2022) also achieves a $\widetilde{\mathcal{O}}(\varepsilon^{-2.5})$ sample complexity under the similar set of assumptions. However, their algorithm needs a subroutine to approximately solve a cubic regularized subproblem. Moreover, SCRN requires large batch sizes and makes use of second-order information. In constrast, N-PG-IGT is single-loop, only requires to sample a single trajectory per iteration and does not need any second-order information to be implemented.

4.3. Convergence analysis of (N)-HARPG

In this section, we further improve the sample complexity for global convergence to $\tilde{\mathcal{O}}(\varepsilon^{-2})$. We show this result for both HARPG and its normalized version N-HARPG.

Theorem 2 (**Global convergence of** HARPG)**.** *Let Assumptions 1, 3 and 4 hold. Set $\gamma_t = \gamma_0\eta_t^{1/2}$, $\eta_t = \frac{2}{t+2}$ with $\gamma_0 = \min\left\{\frac{1}{8\sqrt{6}(L_g+\sigma_g+D_h\gamma^H)}, \frac{\sqrt{2}M_g}{\sqrt{3}\sigma_g\mu_F}\right\}$, $H = (1-\gamma)^{-1}\log(T+1)$. Then for every integer $T \geq 1$, the output θ_T of* HARPG *(see Algorithm 2) satisfies*

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

where σ_g, σ_h, L_g are defined in Proposition 1 and D_h is defined in Lemma 4 (see Appendix E).

[4] in the sense that the compatible function approximation error is evaluated under the *shifted* distribution induced by $d_\rho^{\pi^*}$ and π^*.

[5] We delay its definition to the appendix, it is only used here.

[6] We also analyze convergence of this method to a FOSP for a wider class of policy parameterizations (under Assumptions 1 and 2) in Theorem 5 in Appendix F.

Theorem 3 (**Global convergence of** N-HARPG). *Let Assumptions 1, 3 and 4 hold. Set* $\gamma_t = \frac{6M_g}{\mu_F(t+2)}$ *and* $\eta_t = \frac{2}{t+2}$, $H = (1-\gamma)^{-1}\log(T+1)$. *Then for every integer* $T \geq 1$, *the output* θ_T *of* N-HARPG *(see Algorithm 2) satisfies*

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

where σ_g, σ_h, L_g *are defined in Proposition 1.*

Corollary 1. *In the setting of Theorem 2 (Theorem 3), for every* $\varepsilon > 0$, *the output* θ_T *of* (N)-HARPG *satisfies* $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + 2\sqrt{\varepsilon_{bias}}/(1-\gamma)$ *for* $T = \mathcal{O}(\varepsilon^{-2})$, *i.e., the total sample complexity is of the order* $\widetilde{\mathcal{O}}(\varepsilon^{-2})$.

We remark that while (N)-HARPG uses Hessian information, our analysis does not require the second-order smoothness of $J(\theta)$ and, thus, Assumption 2. This is achieved by the additional uniform sampling in steps 3-5 of the method.

5. Experiments

We empirically test the performance of the methods on benchmark RL tasks with continuous state-action spaces. We use the class of diagonal Gaussian policies where actions are sampled from

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

for every $\theta \in \mathbb{R}^d, s \in \mathcal{S} = \mathbb{R}^p, a \in \mathcal{A} = \mathbb{R}^q$, where $\mathcal{N}(0, I)$ is the multivariate normal distribution, $\odot$ denotes the elementwise product of two vectors, $\mu_\theta(s)$ and $\sigma_\theta(s)$ are parameterized by fully-connected neural networks. We implement the algorithms based on Vanilla-PG (REINFORCE) implementation in the garage library (garage contributors, 2019) and test the methods on the commonly used MuJoCo environments. In this section, we present the results on `Humanoid` and `Reacher` environments and defer the results on other tasks to Appendix A. For all methods, at each iteration we sample 20 trajectories per iteration, and each trajectory has a maximum length $H = 500$. To ensure the same per-iteration cost for all methods, in (N)-HARPG we sample half of the trajectories to compute the stochastic gradient and another half to estimate the Hessian-vector product. For a fair comparison of the different methods, we start all our runs from the same initial policy π_{θ_0}, where θ_0 is randomly initialized. For all methods, we use time-varying step-sizes γ_t and momentum parameters η_t, which are chosen according to theory for each method, see Appendix A.

5.1. Comparison with tuned initial step-sizes

In the first experiment, we compare the methods with individually tuned initial step-sizes on `Humanoid` task. We run each algorithm with 13 different values of initial step-size γ_0 in the range $\left[10^{-3}, 4\right]$ and select the one with the best *average reward* (averaged over 5 independent runs for each algorithm and initial step-size) after $2 \cdot 10^7$ *system probes* (i.e., state transitions). In Figure 1 (left), we show the median as well as the $1/4$ and $3/4$ quantiles (shaded area) of the *average reward* depending on the number of *system probes*. We observe that N-PG-IGT outperforms other methods and quickly increases the reward above 450. Compared to Vanilla-PG, it stabilizes better around level 500. Surprisingly, theoretically better (N)-HARPG methods perform consistently worse in our experiments with tuned step-sizes. We believe this phenomenon could happen because the stochastic Hessian vector products sampled in (N)-HARPG have much higher variance than stochastic gradients in practice, which prevents us from choosing larger step-sizes, see Remark 1.

5.2. Robustness to initial step-sizes

In this experiment, we test the robustness of the methods to the choice of initial step-size on `Reacher` environment. In Figure 1 (right), we report the number of *system probes* in order to achieve *average reward* above -11 points depending on the initial step-size γ_0. We plot the median together with the $1/4$ and $3/4$ quantiles based on 5 independent runs. We observe that (variance-reduced) HARPG can reach the desired accuracy faster than Vanilla-PG for small initial step-sizes γ_0. In contrast, N-PG-IGT tolerates larger step-sizes than other methods and exhibits more robustness to initial step-sizes.

Remark 1. *It was previously reported that variance-reduced methods including* N-HARPG *(named* SHARP *in (Salehkaleybar et al., 2022)) outperform* Vanilla-PG *(Shen et al., 2019; Huang et al., 2020) without tuned step-sizes. We also observe similar behavior for small untuned step-sizes, see Figure 4 in Appendix A.*

6. Concluding Remarks

In this work, we improved the existing $\tilde{\mathcal{O}}(\varepsilon^{-3})$ global optimality sample complexity for Fisher-non-degenerate parametrized policies. We proposed two single-loop momentum-based PG methods respectively achieving $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ and $\tilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexities while only using one or two trajectories per iteration. Notably, our analysis does not rely on the unrealistic assumption of bounded IS weights variance unlike most prior work. To achieve our best $\tilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity, our analysis hinges on a careful combination of a structural weak gradient dominance condition with the design of variance-reduced policy gradients. Future directions of research include relaxing our policy regularity conditions (Zhang et al., 2022) and studying the convergence of more efficient policy search methods for continuous control with non-compact spaces and heavy-tailed policy parameterization, enhancing explo-

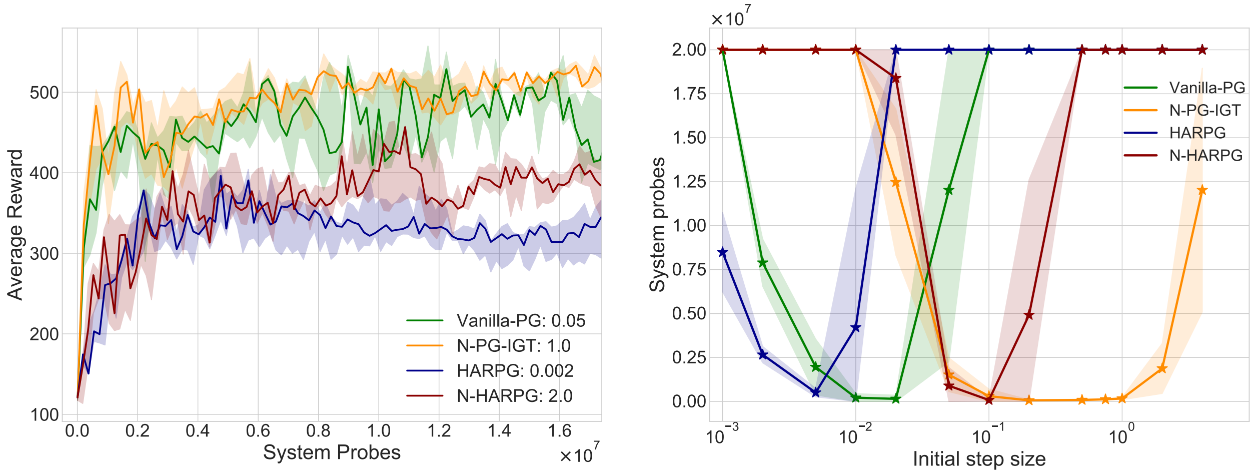

Figure 1: Left: performance of Vanilla-PG, N-PG-IGT, HARPG, N-HARPG on `Humanoid` environment. Right: robustness to initial step-sizes for proposed methods on `Reacher` environment.

ration in the light of recent results in this direction (Bedi et al., 2021).

Acknowledgements

This work was supported by ETH AI Center doctoral fellowship, ETH Foundations of Data Science (ETH-FDS), and ETH Research Grant funded through the ETH Zurich Foundation.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*, 2022.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

Sébastien Arnold, Pierre-Antoine Manzagol, Reza Babanezhad Harikandeh, Ioannis Mitliagkas, and Nicolas Le Roux. Reducing the variance in online optimization by transporting past gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1): 5–16, 2009.

Amrit Singh Bedi, Anjaly Parayil, Junyu Zhang, Mengdi Wang, and Alec Koppel. On the sample complexity and metastability of heavy-tailed policy search in continuous control. *arXiv preprint arXiv:2106.08414*, 2021.

Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian M Sadler, Pratap Tokekar, and Alec Koppel. On the hidden biases of policy mirror ascent in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1716–1731. PMLR, 17–23 Jul 2022.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 11195–11214. PMLR, 2022.

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Anirban DasGupta and Anirban DasGupta. The exponential family and statistical applications. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*, pages 583–612, 2011.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 1910–1934. PMLR, 2022.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ilyas Fatkhullin and Boris Polyak. Optimizing Static Linear Feedback: Gradient Method. *SIAM Journal on Control and Optimization*, 59(5):3887–3911, 2021.

Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global Kurdyka-łojasiewicz inequality. *Advances in Neural Information Processing Systems*, 2022.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic Heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

The garage contributors. Garage: A toolkit for reproducible reinforcement learning research. https://github.com/rlworkgroup/garage, 2019.

Matilde Gargiani, Andrea Zanelli, Andrea Martinelli, Tyler Summers, and John Lygeros. PAGE-PG: A simple and loopless variance-reduced policy gradient method with probabilistic gradient estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7223–7240. PMLR, 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Elad Hazan, Kfir Y. Levy, and Shai Shalev-Shwartz. Beyond Convexity: Stochastic Quasi-Convex Optimization. *arXiv preprint arXiv:1507.02030*, 2015.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *To appear in SIAM Journal on Optimization 2022, arXiv preprint arXiv:2007.05170*, 2020.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *International conference on machine learning*, pages 4422–4433. PMLR, 2020.

Sajad Khodadadian, Thinh T Doan, Justin Romberg, and Siva Theja Maguluri. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*, 2022a.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214, 2022b. ISSN 0167-6911.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783, 1998.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtarik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6286–6295. PMLR, 2021.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Saeed Masiha, Saber Salehkaleybar, Niao He, Negar Kiyavash, and Patrick Thiran. Stochastic second-order methods provably beat sgd for gradient-dominated functions. *To appear in Advances in Neural Information Processing Systems, arXiv preprint arXiv:2205.12856v1*, 2022.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6820–6829. PMLR, 2020.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021a.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7555–7564. PMLR, 2021b.

Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pages 4026–4035. PMLR, 2018.

Nhan Pham, Lam Nguyen, Dzung Phan, Phuong Ha Nguyen, Marten Dijk, and Quoc Tran-Dinh. A hybrid stochastic policy gradient algorithm for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 374–385. PMLR, 2020.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

Saber Salehkaleybar, Sadegh Khorasani, Negar Kiyavash, Niao He, and Patrick Thiran. Adaptive momentum-based policy gradient with second-order information. *arXiv preprint arXiv:2205.08253*, 2022.

Kevin Scaman, Cedric Malherbe, and Ludovic Dos Santos. Convergence Rates of Non-Convex Stochastic Gradient Descent Under a Generic Lojasiewicz Condition and Local Smoothness. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19310–19327. PMLR, June 2022.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International conference on machine learning*, pages 5729–5738. PMLR, 2019.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. PMLR, 2014.

Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232v2*, 2019.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Hoang Tran and Ashok Cutkosky. Better SGD using Second-order Momentum. *arXiv preprint arXiv:2103.03265*, 2021.

J.N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23 (282):1–36, 2022.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020a.

Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR, 2020b.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pages 4358–4369, 2020c.

Long Yang, Qian Zheng, and Gang Pan. Sample complexity of policy gradient finding second-order stationary points. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10630–10638, 2021.

Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic Recursive Momentum for Policy Gradient Methods, 2020.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022a.

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3332–3380. PMLR, 2022b.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020a.

Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021a.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10887–10895, 2021b.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020b.

Matthew S Zhang, Murat A Erdogdu, and Animesh Garg. Convergence and optimality of policy gradient methods in weakly smooth settings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9066–9073, 2022.

Contents

Appendix

A. Additional Experiments and Implementation Details

Complementing the empirical results presented in Section 5, in this section, we present an additional set of experiments on MuJoCo environments and provide further implementation details.

Implementation details. The parameters $\mu_\theta(s)$ and $\sigma_\theta(s)$ of Gaussian policy are parametrized with the feed forward neural network with two hidden layers of size 64 and $\tanh$ activation function. The summary of parameters for each environment is presented in Table 3.

Environments	`Walker2d`	`Hopper`	`Halfcheetah`	`Reacher`	`Humanoid`	`Cartpole`
Horizon	500	1000	500	500	500	200
Number of timesteps	2×10^7	2×10^7	2×10^7	2×10^7	2×10^7	2×10^6
Batch size	20	10	10	200	10	5
best γ_0 for Vanilla-PG	0.02	0.02	0.02	0.02	0.05	0.1
best γ_0 for N-PG-IGT	0.75	0.5	1.0	0.2	1.0	0.2
best γ_0 for HARPG	0.002	0.002	0.001	0.002	0.002	0.02
best γ_0 for NHARPG	0.1	0.05	0.2	0.1	2.0	0.1

Table 3: Hyper-parameters and step-size choice. The initial step-size is chosen from the set of 13 values by the best performance in the last iteration.

A.1. Experiment 1: Comparison with tuned initial step-sizes

We further test the algorithms with the best tuned step-sizes on `Walker2d` and `Reacher`. Similarly to the experiment in Section 5, we run each algorithm five times for every initial step-size γ_0 from the set

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

Then the reported step-size is chosen by the performance of the algorithm in the last iteration. In Table 4, we report the choice of time-varying step-sizes and momentum parameters. We refer the reader to Sections F, G and H for justification of the choice of parameters presented in the table.

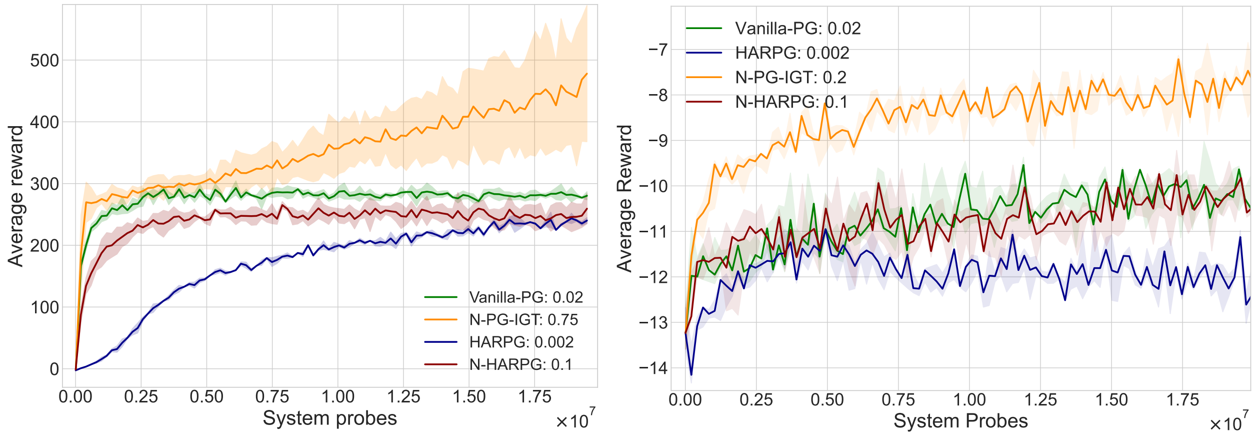

Figure 2: The performance of Vanilla-PG, N-PG-IGT, HARPG, N-HARPG on `Walker2d` (left) and `Reacher` (right) environment.

Similarly to `Humanoid` environment presented in Section 5, we observe that N-PG-IGT outperforms other methods and quickly increases the reward above 450 for `Walker2d` and above -9 for `Reacher`, see Figure 2. Unfortunately, theoretically better HARPG and N-HARPG methods do not show the improvement in practice over Vanilla-PG, which might be due to large variance of Hessian-vector estimator.

	Vanilla-PG	N-PG-IGT	HARPG	N-HARPG
Step-size, γ_t	$\gamma_0\left(\frac{2}{t+2}\right)^{2/3}$	$\frac{2\gamma_0}{t+2}$	$\gamma_0\left(\frac{2}{t+2}\right)^{1/2}$	$\frac{2\gamma_0}{t+2}$
Momentum parameter, η_t	N/A	$\left(\frac{2}{t+2}\right)^{4/5}$	$\frac{2}{t+2}$	$\frac{2}{t+2}$

Table 4: Summary of the sequences of step-sizes and momentum parameters for policy gradient methods. The initial step-size parameter γ_0 was tuned within the set (4). The best value of γ_0 for each algorithm and environment is reported in Table 3.

A.2. Experiment 2: Robustness to initial step-sizes

In addition to the experiment on `Reacher` environment (see Figure 1), we report the sensitivity of the studied methods to initial step-sizes on another two tasks: `Walker2d` and `Humanoid`. For these tasks, we observe a similar trend, which shows that HARPG can be faster than Vanilla-PG, which happens in the small step-size regime, see Figure 3. On the other hand, the normalized methods tolerate larger step-sizes and N-PG-IGT, in particular, demonstrates more robustness to step-size choice. Overall, we observe that, naturally, normalized and non-normalized methods have different range of optimal step-sizes, which makes it necessary to extensively tune the step-sizes for comparing between these two types of algorithms. Another observation is that for the tested environments, the variance reduced HARPG method performs better than Vanilla-PG only for small non-tuned step-sizes. Therefore, in the next experiment, we compare these methods in the small step-size regime.

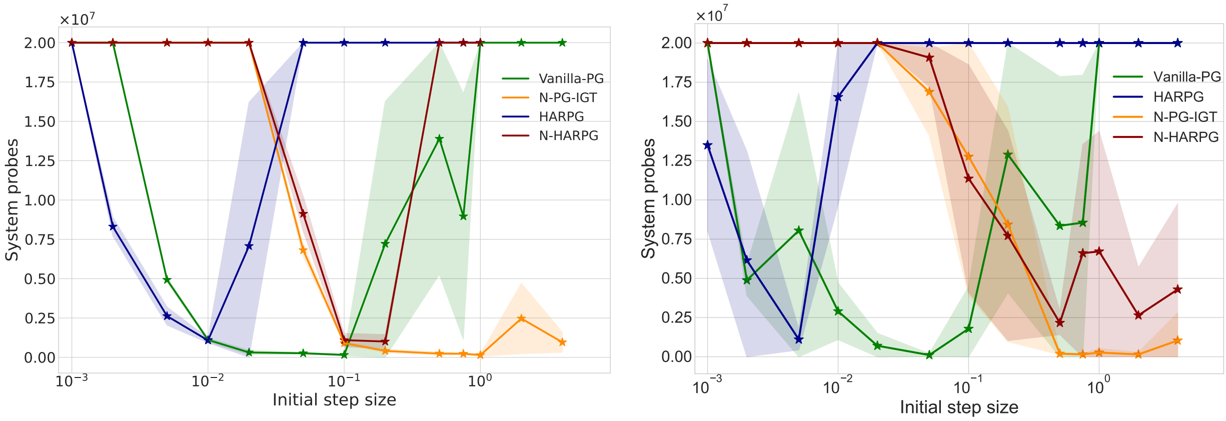

Figure 3: Robustness to initial step-sizes on `Walker2d` (left) and `Humanoid` (right) environments.

A.3. Experiment 3: Comparison of Vanilla-PG and HARPG with small untuned step-sizes

In this experiment, we demonstrate that the variance reduced HARPG method can outperform Vanilla-PG for a fixed small initial step-size γ_0, see Figure 4.[7] Our observations here are consistent with those reported in the previous work for other variance reduced policy gradient algorithms, e.g., (Shen et al., 2019; Huang et al., 2020; Salehkaleybar et al., 2022). However, when the step-sizes are tuned, Vanilla-PG preforms consistently better than variance reduced algorithms, see also Experiment 4 (Figure 5) for additional comparison to other methods with tuned step-sizes.

A.4. Experiment 4: Additional comparison to other methods with tuned step-sizes

In this experiment, we test our algorithms on `Hopper` and `Halfcheetah` tasks with tuned step-sizes. We compare the performance of presented methods with two additional algorithms. VR-PG method, also known as STORM-PG (Yuan et al.,

[7] We do not include normalized algorithms in this comparison since normalization introduces an additional scaling of the step-size.

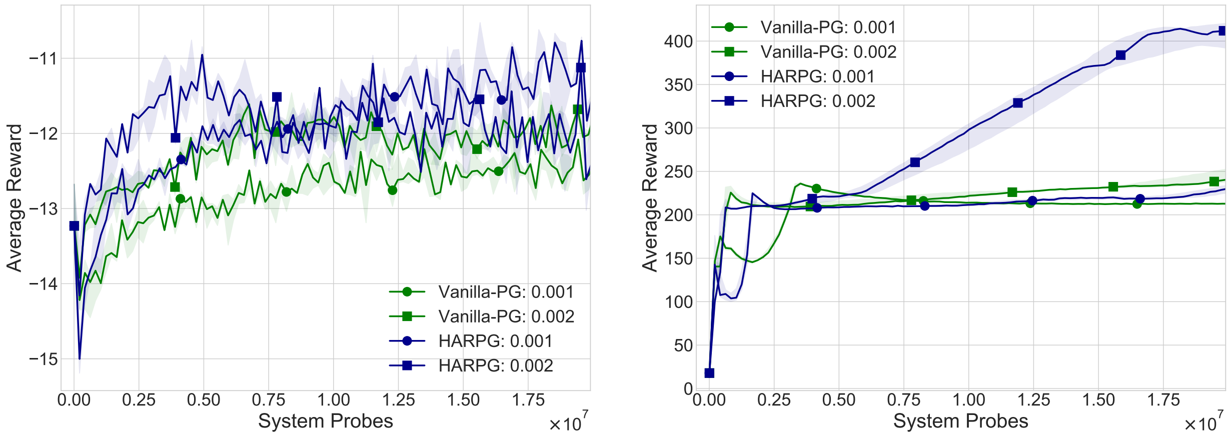

Figure 4: The performance of Vanilla-PG, and HARPG on `Reacher` (left) and `Hopper` (right) environments for non-tuned initial step-sizes.

2020) implements a variance reduced policy gradient method using the importance sampling (IS) weights instead of the Hessian correction. We also consider a normalized version of this method for a fair comparison with N-HARPG. [8] We observe in Figure 5, that while VR-PG and N-VR-PG perform similarly or better than HARPG and N-HARPG, our method N-PG-IGT still converges faster.

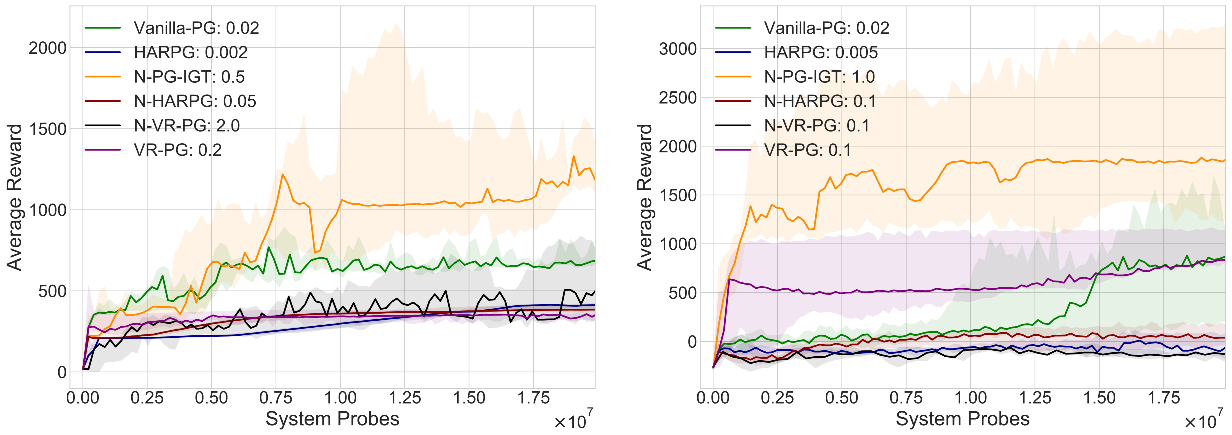

Figure 5: The performance of Vanilla-PG, N-PG-IGT, HARPG, N-HARPG, VR-PG and N-VR-PG on `Hopper` (left) and `Halfcheetah` (right) environments.

A.5. Experiment 5: Performance under the soft-max parameterization for discrete state-action space

Although our main theoretical results (except for FOSP of N-PG-IGT in Theorem 5) only hold under the Fisher-non-degenerate (FND) parameterization Assumption 3, all methods presented in this work can be applied to other policy classes without any modifications. In this experiment, we test the proposed algorithms on the popular `Cartpole` environment with soft-max policy parameterization. We observe in Figure 6 that the behavior of methods is similar to other environments.

[8] We note that theoretical analysis of these methods require an additional assumption on boundedness of IS weights.

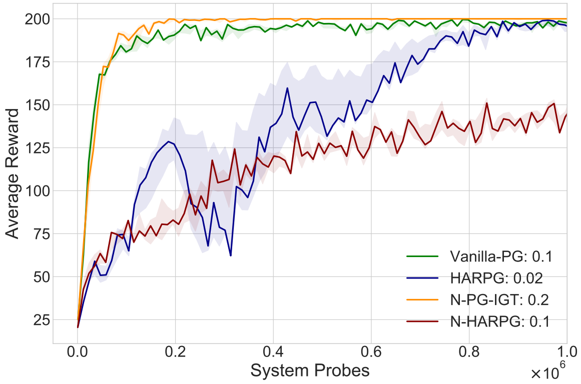

Figure 6: The performance of Vanilla-PG, N-PG-IGT, HARPG, and N-HARPG on `Cartpole` environment for tuned initial step-sizes.

B. Further Discussion of Assumptions from Section 4.1

In this section, we discuss our assumptions in the light of two simple examples while we also comment on the generality of our assumptions beyond those examples. However, we would like to highlight that our set of assumptions is general and standard in the recent studies of convergence of PG methods (see Xu et al. (2020a); Zhang et al. (2020b); Liu et al. (2020); Ding et al. (2022); Gargiani et al. (2022); Masiha et al. (2022); Salehkaleybar et al. (2022), to name a few). We believe that relaxing some of the assumptions is an interesting question that we defer for future work.

B.1. Gaussian policy

Let $\varphi : \mathcal{S} \to \mathbb{R}^d$ be a feature map and set[9] $\mathcal{A} = \mathbb{R}$. We consider the common Gaussian policy parameterized by a linear mean parameterization and a constant variance parameter $\sigma^2 > 0$ defined for every $\theta \in \mathbb{R}^d$, $(s,a) \in \mathcal{S} \times \mathcal{A}$ by:

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

More general forms of Gaussian policies use neural networks for the mean and the standard deviation parameterizations. Note that one can consider even more general parameterizations such as the exponential family or symmetric α-stable policies which include the Gaussian policy as a particular case. We refer the interested reader to the nice exposition in (Bedi et al., 2021) for a discussion around such heavy-tailed policy parameterizations. In the case of (5), the score function can be written for every $\theta \in \mathbb{R}^d$, $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows:

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

Assumption 1-(i) (Bounded score function): A few comments are in order regarding this assumption:

1. In the specific case of the Gaussian policy, we notice that strictly speaking this assumption is not satisfied for all $a \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$ even if the feature map is bounded as it was previously noted in the literature (Yuan et al., 2022b; Bedi et al., 2022). However, a number of recent works (Papini et al., 2018, Section C), (Xu et al., 2020b, Section 7), (Liu et al., 2020, Section 4), (Xu et al., 2020a, Corollary 4.8) assume boundedness of the score function. It is easy to see that such an assumption can be guaranteed if one additionally assumes the bound on sampled actions and on the mean parameterization $\varphi(s)^\top\theta$. In practice, this is usually enforced by clipping the actions selected by the policy π_θ and projecting the update of the gradient method into the ball of fixed radius.

2. It was previously shown that this standard assumption can be relaxed to hold in expectation (see Assumption 4.1 (Yuan et al., 2022b)) instead of for all state-action pairs. Such relaxed version allows to accomodate the Gaussian

[9] The more general case $\mathcal{A} = \mathbb{R}^q (q > 1)$ can be treated similarly and only needs slight adjustments in the definition of the Gaussian policy such as considering a matrix feature map (instead of a vector one) of proper dimensions. We briefly comment on this case below when discussing Assumption 3.

parameterization without additional assumptions, while it is sufficient to validate smoothness of J and the bounded variance of stochastic gradients, see (Yuan et al., 2022b, Lemma 4.2 and 4.4). However, in this work, we do not focus on such technical relaxations and work with the standard assumption of bounded score function.

3. For a policy inducing an unbounded score function, an alternative is to partition the action space for instance according to when the score function is bounded and integrable using an exploration parameter. We refer the reader to Section 5.1 in Bedi et al. (2021). Adopting this approach requires adjustments in the analysis.

Assumption 1-(ii)(Lipschitzness of the score function): We compute $\nabla^2 \log \pi_\theta(s,a) = -\frac{1}{\sigma^2}\varphi(s)\varphi(s)^\top$ which is bounded if the feature map is bounded.

Assumption 2(Second-order smoothness of the score function): This assumption trivially holds since $\nabla^2 \log \pi_\theta(s,a)$ is independent of θ. Regarding this assumption, we further highlight that it is not required for (N)-HARPG (see Theorems 2 and 3). It is only required for N-PG-IGT. While this assumption is key for the improvement for N-PG-IGT, we point out that (a) it is a mild assumption satisfied for the Gaussian policy as we just mentioned and (b) Table 2 (column 'No 2nd. smooth') highlights it as an additional assumption compared to existing results for a fair comparison. In particular, we compare to (Masiha et al., 2022) which makes the same additional assumption.

Assumption 3(Fisher non-degeneracy): The Fisher information matrix for a Gaussian policy can be written as $F_\rho(\theta) = \frac{1}{\sigma^2}\varphi(s)\varphi(s)^\top$, which is positive definite for a full row-rank feature map. More generally[10],

1. **Linear mean parameterization**: when the Gaussian policy is multivariate ($\mathcal{A} = \mathbb{R}^q$) and parametrized by $\mu_\theta(s) = \Phi(s)^T\theta$ where $\Phi : \mathcal{S} \to \mathbb{R}^{d\times q}$ (d being the dimension of the parameter θ) and Σ is a fixed covariance matrix, the Fisher information matrix is independent of θ as it is given by $\Phi(s)\Sigma^{-1}\Phi(s)^T$. Hence, Assumption 3 is satisfied if $\Phi(s)$ is full-row-rank. For the usual setting where the dimension of the parameter θ is smaller than the dimension of the action space (i.e., $d < q$), choosing the rows of $\Phi(s)$ to be linearly independent would satisfy this full-rank requirement. This is a common assumption in the RL literature when considering linear function approximation (see e.g., Tsitsiklis and Van Roy (1997); Melo et al. (2008)).

2. **Neural net mean parameterization**: when $\mu_\theta(s)$ is nonlinear in θ as in the celebrated case of neural networks, Assumption 3 can be satisfied under adequate (uniform in θ) assumptions on the Jacobian of $\mu_\theta(s)$ akin to the conditions for the linear mean parameterization.

3. Full-rank exponential family distributions (DasGupta and DasGupta, 2011) with a parameterized mean $\mu_\theta(s)$ do also satisfy Assumption 3. Let us mention that the Fisher information matrix is positive definite for many regular statistical models (Kullback, 1997).

B.2. Cauchy policy

Let $\varphi : \mathcal{S} \to \mathbb{R}^d$ be a feature map and set $\mathcal{A} = \mathbb{R}$. We suppose this feature map to be bounded (i.e., $\|\phi(s)\| \leq D$ for every $s \in \mathcal{S}$ for some positive constant D) and full-row-rank. The Cauchy distribution with linear mean parameterization $\varphi(s)^\top\theta$ and the scale parameter $\sigma > 0$ is given by:

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

for every $\theta \in \mathbb{R}^d, s \in \mathcal{S}, a \in \mathcal{A}$. In this case the score function can be written as

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

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

Following (Bedi et al., 2022, Lemma 4.7), we verify our assumptions under this policy parameterization.

[10] We report some comments from the discussion in Appendix B.2 in (Liu et al., 2020) for completeness and for the convenience of the reader.

Assumption 1-(i) (Bounded score function): Since $\frac{2x}{1+x^2} \leq 1$ for all $x \in \mathbb{R}$, we have $\|\nabla \log \pi_\theta(a|s)\| \leq M_g$ with $M_g = \frac{D}{\sigma}$ for all $\theta \in \mathbb{R}^d$ and $s \in \mathcal{S}$, $a \in \mathcal{A}$.

Assumption 1-(ii)(Lipschitzness of the score function): We compute $\nabla^2 \log \pi_\theta(a|s) = \frac{2(1-x_\theta^2)}{(1+x_\theta^2)^2} \left(\frac{\varphi(s)\varphi(s)^\top}{\sigma^2} \right)$. Then since $\frac{2(1-x^2)}{(1+x^2)^2} \leq \frac{3}{2}$ for all $x \in \mathbb{R}$, we can bound $\left\|\nabla^2 \log \pi_\theta(a|s)\right\| \leq M_h$ with $M_h = \frac{3D^2}{2\sigma^2}$.

Assumption 2(Second-order smoothness of the score function): We take an arbitrary non-zero vector $u \in \mathbb{R}^d$, and compute $\nabla(\nabla^2 \log \pi_\theta(a|s)\, u) = \frac{4x_\theta(x_\theta^2-3)}{(1+x_\theta^2)^3} \frac{1}{\sigma^3} \varphi(s)\varphi(s)^\top \varphi(s)^\top u$. Therefore, $\|\nabla^2 \log \pi_\theta(a|s) - \nabla^2 \log \pi_{\theta'}(a|s)\| \leq l_2 \|\theta - \theta'\|$ with $l_2 = \frac{4D^3}{\sigma^3}$.

Assumption 3(Fisher non-degeneracy): Using reparameterization theorem for the Fisher information of Cauchy distribution, the Fisher information matrix for $\pi_\theta(a|s)$ can be written as $F_\rho(\theta) = \frac{1}{2\sigma^2} \varphi(s)\varphi(s)^\top$, which is positive definite for full row-rank feature map.

About Assumption 4 (Bounded compatible function approximation transfer error): As previously mentioned, this assumption has been used in several works (Agarwal et al., 2021; Liu et al., 2020; Ding et al., 2022; Yuan et al., 2022b).

C. Proof Sketch of the Main Results

In this section, we provide a brief proof sketch of our results to highlight the proof technique leading to improved sample complexity compared to prior work. The sample complexity improvement comes from a combined effect of (a) exploiting the relaxed weak gradient dominance and (b) deriving a better variance error control. The complete formal proofs are presented in the subsequent sections.

Step 1. Ascent-like lemma. Define $\hat{e}_t \overset{\text{def}}{=} d_t - \nabla J_H(\theta_t)$. First, in Lemma 5 we prove the following inequality for any sequence θ_t updated via $\theta_{t+1} = \theta_t + \gamma_t d_t / \|d_t\|$ for any nonzero update direction d_t (to be given by N-PG-IGT or N-HARPG):

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

A similar lemma holds for the variant HARPG without normalization, see Lemma 8.

Step 2. Relaxed weak gradient domination inequality. Combining Assumptions 3 and 4, and following the derivations of Ding et al. (2022), we obtain the relaxed weak gradient dominance inequality, satisfied by the expected return function $J(\theta)$ (Lemma 2):

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

where $\varepsilon' = \frac{\mu_F \sqrt{\varepsilon_{\text{bias}}}}{M_g(1-\gamma)}$ and $\mu = \frac{\mu_F^2}{2M_g^2}$. This structural property of the objective function is one of the key tools of our analysis. Note that we also have $\varepsilon' = \frac{\mu_F \sqrt{\varepsilon_{\text{bias}}}}{M_g(1-\gamma)}$ and $\mu = \frac{\mu_F^2}{2M_g^2}$. In particular, the error ε' is intimately related to the expressivity of the policy parameterization quantified by $\varepsilon_{\text{bias}}$ as discussed in Assumption 4. Notice that compared to Ding et al. (2022), we do not derive a bound for the expected gradient norm and inject it in (7) to obtain our final global convergence rates. Instead, we carefully incorporate (7) into the ascent-like lemma derived for the (normalized) gradient method to obtain a recursion on $\delta_t \overset{\text{def}}{=} J^* - \mathbb{E}\left[J(\theta_t)\right]$, which leads to improved convergence rates.

Remark 2 (About gradient dominance)**.** *Notice that our relaxed weak gradient dominance condition features the norm of the gradient instead of the norm squared. Such a structural property for the standard RL setting we consider here has been unveiled and exploited in several recent works (Agarwal et al., 2021; Xiao, 2022). Note that this condition is different from the Polyak-Łojasiewicz (PŁ) like condition which features a gradient norm squared in* (7) *(with $\epsilon' = 0$). A similar PŁ condition was shown for the state feedback version of LQR problem in continuous control (Fazel et al., 2018; Fatkhullin and Polyak, 2021) which differs from our current setting.*

Step 3. Variance reduction control. This crucial step depends on the algorithm used and consists in carefully deriving a recursion on $\mathbb{E}\left[\|e_t\|\right]$ or $\mathbb{E}\left[\|e_t\|^2\right]$. For instance, in Lemma 7 for N-PG-IGT method, we derive the following recursion in vector form

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

where $\mathbb{E}\left[\|e_t\|\right] \leq \sigma_g$, and the terms S_t and Z_t correspond to the second-order Taylor approximation of J at different points. This allows us to utilize Assumption 2 together with the normalized update rule (which guarantees $\|\theta_{t+1} - \theta_t\| = \gamma_t$), and derive the control for $\|S_t\| \leq \mathcal{O}\left(\gamma_t^2 + \gamma^H\right)$, $\|Z_t\| \leq \mathcal{O}\left(\gamma_t^2 \eta_t^{-2} + \gamma^H\right)$.

By choosing $\gamma_t = \mathcal{O}\left(t^{-1}\right), \eta_t = \mathcal{O}\left(t^{-4/5}\right)$, $H = (1-\gamma)^{-1}\log(T+1)$ we guarantee

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

We emphasize the importance of using time varying step-sizes and momentum parameters for our analysis as opposed to the constant step-size depending on the total number of iterations T. In particular, selecting time varying η_t allows us to resolve (8) for all $t \geq 1$ and plug it in (6).

Step 4. Combine the variance-reduced estimate and the ascent-like lemma. Finally, we incorporate the results of Steps 2-3 into the ascent-like lemma derived in Step 1 to obtain a recursion on the sequence (δ_t). For example, for the N-PG-IGT method, we arrive at

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

It remains to resolve the resulting recursion on the sequence (δ_t) using technical lemma from Section J in order to obtain the final rate.

D. Further Related Work

In this section, we provide a brief additional discussion including first-order stationarity of variance-reduced PG methods, global optimality of exact PG methods and some additional related works complementing the main discussion in Section 1 of the paper.

D.1. First-order stationarity for variance-reduced PG methods

In the last few years, several variance-reduced PG methods were proposed in the literature to improve sample efficiency over Vanilla-PG which achieves a $\tilde{\mathcal{O}}(\varepsilon^{-4})$ sample complexity. Papini et al. (2018) proposed a stochastic variance reduced policy gradient (SVRPG) method achieving a $\tilde{\mathcal{O}}(\varepsilon^{-4})$ sample complexity. Xu et al. (2020b) revisited this algorithm and established a $\tilde{\mathcal{O}}(\varepsilon^{-10/3})$ sample complexity. This result was further improved to $\tilde{\mathcal{O}}(\varepsilon^{-3})$ by using second-order information in Shen et al. (2019) or variants of variance reduction methods such as SARAH (Nguyen et al., 2017), SPIDER (Fang et al., 2018) or PAGE (Li et al., 2021) in Xu et al. (2020a); Pham et al. (2020); Gargiani et al. (2022). Yuan et al. and Huang et al. (Yuan et al., 2020; Huang et al., 2020) proposed momentum-based policy gradient methods based on STORM (Cutkosky and Orabona, 2019). In particular, Huang et al. (2020) removed the need of large batches of trajectories for variance reduction. All the aforementioned works in this section proposing variance-reduced PG methods (except Shen et al. (2019) which uses second-order information) use importance sampling to account for the non-stationarity of the sampling distribution. As a consequence of the use of the importance sampling mechanism to address the non-stationarity issue, most of previous works require an unverifiable assumption stating that the variance of the importance sampling weights is bounded. To address this issue, Zhang et al. (2021a) provide a gradient truncation mechanism complementing importance sampling for the softmax parameterization whereas Shen et al. (2019) and Salehkaleybar et al. (2022) incorporate second-order information. Both approaches lead to the same $\tilde{\mathcal{O}}(\varepsilon^{-3})$ FOSP sample complexity.

D.2. Global optimality of exact PG methods

More recently, a line of research focused on establishing global optimality convergence rates i.e. deriving the sample complexity to reach an ε-approximate global optimum of the expected return function (Agarwal et al., 2021; Fazel et al., 2018; Bhandari and Russo, 2019; Mei et al., 2020; Zhang et al., 2020b; 2021b) instead of the mere ε-approximate first order stationary point. This line of research leverages additional structure of the policy optimization problem under the form of gradient dominance (Agarwal et al., 2021; Fazel et al., 2018), Łojasiewicz-like inequalities (Mei et al., 2020; 2021b) or hidden convexity (Zhang et al., 2021a). These structural properties guarantee that the non-concave policy optimization objective has no suboptimal stationary points. Most of these works assume the access to exact policy gradients to show fast convergence rates (Mei et al., 2020; 2021b; Xiao, 2022). Other works also establish global optimality guarantees for the Natural PG method supposing access to exact gradients (Khodadadian et al., 2022b; Bhandari and Russo, 2021). More precisely, Khodadadian et al. (2022b) and Bhandari and Russo (2021) show a linear convergence rate for the Natural PG method with adaptive step sizes by drawing a connection between policy gradient and policy iteration. These results are further sharpened in Lan (2022); Xiao (2022) in which it is shown that a general class of policy mirror descent methods, including the natural policy gradient method and a projected Q-descent method, enjoy a linear rate of convergence with geometrically increasing step sizes in the tabular setting, improving over the previously known sublinear rate of simpler PG methods (Agarwal et al., 2021; Mei et al., 2020). Zhang et al. (2020a) consider the more general problem which consists in maximizing a general concave utility function of the state-action occupancy measure. Leveraging the hidden convexity structure of the problem, they propose a globally convergent algorithm derived from a novel variational policy gradient theorem for RL with general utilities.

D.3. Further discussion about stochastic policy optimization

In the case of stochastic on-policy gradients, the acceleration brought by exploiting geometric information in the true gradient setting is not obviously exploitable as this was accurately reported in Mei et al. (2021a) for the tabular softmax parameterization. In particular, Mei et al. (2021a) exhibit a trade-off between exploiting geometry to accelerate convergence and overcoming the noise due to stochastic gradients (which is possibly infinite). We also mention that several works (see for e.g., Chen et al. (2022); Chen and Maguluri (2022); Khodadadian et al. (2022a); Hong et al. (2020); Qiu et al. (2021); Wu et al. (2020); Xu et al. (2020c)) proposed finite-time analysis for actor-critic methods to solve the policy optimization problem using (linear) function approximation for the Q-function or advantage function under ergodicity assumptions on the underlying Markov chain. Our work is focused on Monte-Carlo based PG methods (which are REINFORCE-like) for which no critic is needed (no function approximation error relative to this critic process is incurred) and no ergodicity assumption is needed. We rather rely on the compatible function approximation framework. Besides this line of works, Zhang et al. (2021b) provide global convergence rates for the REINFORCE algorithm for episodic RL with a fixed mini-batch instead of arbitrarily large batch sizes. These results take the form of an anytime sublinear high probability regret bound as well as an almost sure global convergence of the average regret with an asymptotically sublinear rate. In particular, for their regret analysis, they consider the softmax parameterization with a vanishing log barrier regularization and the result translates to a $\tilde{\mathcal{O}}(\varepsilon^{-6})$ global sample complexity as also reported in (Yuan et al., 2022b). Recently, several works considered heavy-tailed policy parameterizations offering better exploration in continuous control problems (Bedi et al., 2021; 2022).

D.4. Comparison to prior work in stochastic optimization

In the stochastic optimization setting, several recent works consider optimizing functions satisfying a weak gradient domination inequality (see Lemma 2 (7) with $\varepsilon' = 0$). In particular, (Fontaine et al., 2021; Fatkhullin et al., 2022; Scaman et al., 2022) obtain a $\mathcal{O}(\varepsilon^{-3})$ sample complexity for finding a global optimum in expectation for Stochastic Gradient Descent[11], which is the same complexity as for (stochastic) Vanilla-PG algorithm in the RL setting (Yuan et al., 2022b). Fatkhullin et al. (2022) further improves this complexity to $\mathcal{O}(\varepsilon^{-2})$ under additional average smoothness assumption by proposing a variance reduced algorithm named PAGER. Their algorithm has a double loop structure and adopts a probabilistic switch between two different batch-size sequences (increasing over iterations), which makes it more difficult to implement and analyze. Another work by Masiha et al. (2022) considers stochastic second-order methods under similar assumptions. Their Stochastic Cubic Regularized Newton's algorithm (SCRN) achieves a $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ sample complexity under a similar set of assumptions. However, SCRN uses a subroutine to approximately solve a cubic regularized subproblem (which makes it computationally inefficient), it uses samples of stochastic Hessian and requires large batches for both stochastic gradient and Hessian at every iteration. To the best of our knowledge (Fatkhullin et al., 2022) and (Masiha et al., 2022) are the only works, which provide the sample complexity beyond $\mathcal{O}(\varepsilon^{-3})$ for stochastic optimization for functions satisfying weak gradient dominance condition. However, these optimization works consider complicated algorithms involving parameter restart, cubic regularization subproblem and/or large mini-batch sizes at each iteration. In contrast, our proposed algorithms improve over $\mathcal{O}(\varepsilon^{-3})$ sample complexity with simple algorithms (N-PG-IGT, HARPG, N-HARPG), which are computationally efficient and batch-free (only require to sample one trajectory per iteration). In addition, our analysis is much simpler than the analysis of algorithms in (Fatkhullin et al., 2022) and (Masiha et al., 2022). Our analysis is also of independent interest in the study of stochastic optimization under the weak gradient domination assumption.

[11] These works consider more general assumptions, such as α-PŁ, i.e., there exists $\alpha \in [1,2]$ such that for all $\theta \in \mathbb{R}^d$ it holds $\|\nabla J(\theta)\|^{\alpha} \geq (2\mu)^{\alpha/2}(J^* - J(\theta))$, and derive the sample complexities depending on the value of α. It turns out that when $\alpha = 1$ as in our setting, the sample complexity is the worst (in ε) compared to other $\alpha > 1$.

E. Notation and Useful Lemma

Notation. For any integer p, the euclidean space $\mathbb{R}^p$ is equipped with its usual inner product $\langle \cdot, \cdot \rangle$ and its corresponding 2-norm $\| \cdot \|$. The transpose of a vector $x \in \mathbb{R}^p$ for any integer p is denoted by $x^\top$ and we denote by I_p the identity matrix of dimension p. For two matrices $A, B \in \mathbb{R}^{p \times q}$ for two integers $p, q \geq 1$, the notation $A \succcurlyeq B$ indicates that the matrix $A - B$ is positive semi-definite. For two sequences of nonnegative reals $(a_n), (b_n)$, we use the notation $a_n = \mathcal{O}(b_n)$ if there exists $c > 0$ such that $a_n \leq c\, b_n$ for every integer n . Hence, unless otherwise specified, we use the $\mathcal{O}(\cdot)$ to hide only absolute constants which do not depend on any problem parameter and we further use the notation $\tilde{\mathcal{O}}(\cdot)$ to hide only absolute constants and logarithmic factors.

Advantage function. For completeness, we define the advantage function that appeared in the transferred compatible function approximation error in the statement of Assumption 4 in the main part of this paper. For this purpose, we first define for every policy π the state-action value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for every $s \in \mathcal{S}, a \in \mathcal{A}$ as:

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

Under the same policy π, the state-value function $V^\pi : \mathcal{S} \to \mathbb{R}$ and the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are defined for every $s \in \mathcal{S}, a \in \mathcal{A}$ as follows:

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

We now state two well-known useful lemma for our analysis. The first lemma shows that the Hessian of the expected return function is Lipschitz under a second-order regularity condition on the policy parameterization.

Lemma 3 ((A.68) in (Zhang et al., 2020b))**.** *Let Assumptions 1 and 2 hold, then we have*

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

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

The second lemma controls the truncation error due to truncating simulated trajectories to the horizon H in our infinite horizon setting. Notably, this error vanishes geometrically fast with the horizon H .

Lemma 4 (Lemma 2 in (Masiha et al., 2022))**.** *Let Assumption 1 be satisfied, then for all $\theta \in \mathbb{R}^d$ and every $H \geq 1$, we have*

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

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

F. Proof of Theorem 1 (N-PG-IGT)

F.1. Global convergence

We first state a more detailed version of Theorem 1.

Theorem 4. *Let Assumptions 1, 2, 3 and 4 hold. Set* $\gamma_t = \frac{6}{\sqrt{2\mu}(t+2)} = \frac{6M_g}{\mu_F(t+2)}$, $\eta_t = \left(\frac{2}{t+2}\right)^{4/5}$, $H = \frac{9}{5}(1-\gamma)^{-1}\log(T+1)$ *. Then for every integer* $T \geq 1$*, the output* θ_T *of* N-PG-IGT *(Algorithm 1) satisfies*

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

where σ_g, L_g, L_h *are defined in Proposition 1 and Lemma 1 respectively,* D_g *is defined in Lemma 4,* ε' *and* μ *are defined in Lemma 2. Here* $\mathcal{O}(\cdot)$ *only hides absolute numerical constants.*

We start by establishing an ascent-like lemma on the function J (we are maximizing J). The following result holds for a general normalized update rule for any sequence of update directions (d_t).

Lemma 5. *Let Assumption 1 hold. Let* (d_t) *be any sequence of vectors in* $\mathbb{R}^d$ *and consider the sequence* (θ_t) *defined by:*

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

where $\theta_0 \in \mathbb{R}^d$ *(and* $\theta_{t+1} = \theta_t$ *if* $d_t = 0$*). Then we have for every integer* $t \geq 1$*, for any positive step-size* γ_t*,*

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

Proof. By smoothness of the expected return function J (Proposition 1) and using the update rule for θ_t, we get

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

Now let us bound the second term in the above inequality. Define $\hat{e}_t = d_t - \nabla J_H(\theta_t)$. We consider two cases. First, if $\|\hat{e}_t\| \leq \frac{1}{2}\|\nabla J_H(\theta_t)\|$, then

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

Otherwise, if $\|\hat{e}_t\| \geq \frac{1}{2}\|\nabla J_H(\theta_t)\|$, we have

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

Combining the two cases concludes the proof of the lemma:

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

□

We introduce a convenient shorthand notation for the rest of the paper:

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

where $J^\star$ is the optimal expected return and $(d_t), (\theta_t)$ are the sequences computed by the algorithm (N-PG-IGT later in this section, and (N)-HARPG or others in the following sections).

In the next lemma, we incorporate the relaxed weak gradient dominance inequality into the previous ascent-like lemma to bound the gradient norm. The resulting inequality is a recursive bound in the expected return gap which is important to derive our convergence rate.

Lemma 6. *Let Assumptions 1, 3 and 4 be satisfied. In the setting of Lemma 5, we have for every integer $t \geq 1$, for any positive step-size γ_t,*

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

where we recall that (δ_t) is defined in (12).

Proof. Under Assumptions 3 and 4, we can use Lemma 2 and combine it with Lemma 5 to bound the gradient norm and obtain

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

where we have used Lemma 4 to control the truncation error in the last term of (10). Adding $J^\star$ to both sides and taking expectation, we obtain (13). □

Given Lemma 6, we now control the remaining error term $\mathbb{E}[\|d_t - \nabla J_H(\theta_t)\|]$.

Lemma 7. *Let Assumptions 1 and 2 be satisfied. Let $(\tilde{\theta}_t), (d_t)$ and (θ_t) be the sequences generated by the* N-PG-IGT *algorithm (see Algorithm 1) with $\eta_t = \left(\frac{2}{t+2}\right)^{4/5}$ and $\gamma_t = \frac{6}{\sqrt{2\mu}(t+2)}$. Then for any integer $t \geq 0$ it holds*

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

where $C(p,q)$ for $q \in [0,1)$, $p > 0$ is a numerical constant defined in Lemma 15.

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

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

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

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

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

Notice that by Lemma 1 (second-order smoothness), Lemma 4 and triangle inequality, we have

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

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

Then, we derive a recursion on the sequence $(\hat{e}_t)$ by writing

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

where (i) follows from the update rule of the sequence (d_t) and (iii) stems from recognizing S_t, Z_t and observing that the last term in (ii) is zero thanks to the update rule of the sequence $(\tilde{\theta}_t)$ in N-PG-IGT (the Hessian correction terms in the last term disappear). Notice that this last simplification is the main reason explaining the update rule of the sequence $\tilde{\theta}_t$.

Consider an integer $T \geq 1$. Unrolling the recursion and using the notation $\zeta_{t+1,T} \overset{\text{def}}{=} \prod_{\tau=t+1}^{T-1}(1-\eta_{\tau+1})$ (with $\zeta_{T,T} = 1$ and $\zeta_{0,T} = \prod_{t=0}^{T-1}(1-\eta_{t+1})$), we obtain

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

Define for every integer $t \geq 1$ the σ-field $\mathcal{F}_t \overset{\text{def}}{=} \sigma(\left\{\tilde{\theta}_0, \tilde{\tau}_0, \ldots, \tilde{\tau}_{t-1}\right\})$ where $\tilde{\tau}_s \sim p(\cdot|\pi_{\tilde{\theta}_s})$ for every $0 \leq s \leq t-1$. Notice that for any integers $t_2 > t_1 \geq 1$ we have

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

Then using triangle inequality, taking expectation and applying Jensen's inequality, we get

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

where in (i) we apply (15) and (17), and use independence $\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle\right] = 0$ for any integers $t_1 \neq t_2$ (larger than 1). In the last inequality, we use $\mathbb{E}\left[\|e_{t+1}\|^2\right] \leq \sigma_g^2$ for any $t \in \{0, \cdots, T-1\}$.

Further, using Lemma 14 and 15 we have

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

where $C(p,q)$ for $q \in [0, 1)$, $p > 0$ is defined in Lemma 15. □

End of Proof of Theorem 4 (and hence Theorem 1). We now combine the ascent-like lemma incorporating the relaxed weak gradient dominance with the estimate of the vanishing error obtained in the previous lemma.

Proof. Combining the results of Lemma 6 with Lemma 7 we get

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

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

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

□

F.2. Convergence to first order stationary point

Theorem 5. *Let Assumptions 1 and 2 hold. Set* $\gamma_t=\left(\frac{2}{t+2}\right)^{5/7}$, $\eta_t=\left(\frac{2}{t+2}\right)^{4/7}$, $H=(1-\gamma)^{-1}\log(T+1)$. *Let* $\bar{\theta}_T$ *be sampled from the iterates of* N-PG-IGT *(Algorithm 1)* $\{\theta_0,\ldots,\theta_{T-1}\}$ *with probability distribution* $\mathbb{P}(\bar{\theta}_T=\theta_t)=\frac{\gamma_t}{\sum_{t=0}^{T-1}\gamma_t}$. *Then for every integer* $T\geq 1$, $\bar{\theta}_T$ *satisfies*

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

where σ_g, L_g, L_h *are defined in Proposition 1 and Lemma 1 respectively,* D_g *is defined in Lemma 4. Here* $\mathcal{O}(\cdot)$ *only hides absolute numerical constants.*

Proof. It follows from Lemma 5 combined with Lemma 4 that

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

We now control the error term $\|\hat{e}_t\|=\|d_t-\nabla J_H(\theta_t)\|$ as follows. By (20) in the proof of Lemma 7, we have

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

where in (i) we applied Lemma 14 and 15. The value of $C(p,q)$ for $q\in[0,1)$, $p>0$ is defined in Lemma 15, $\hat{e}_t=d_t-\nabla J_H(\theta_t)$

Combining the above inequality with (22), we obtain

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

Summing up the above inequality for $t = 0, \ldots, T-1$ and rearranging the terms, we arrive at

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

□

G. Proof of Theorem 2 (HARPG)

In this section, for the proof of Theorem 2 (for HARPG), we introduce the following additional useful shorthand notations:

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

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

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

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

where $(\theta_t), (d_t)$ are the sequences generated by HARPG and (λ_t) is a sequence of positive scalars to be determined later.

First, we state a detailed version of Theorem 2 before providing a complete proof.

Theorem 6. *Let Assumptions 1, 3 and 4 hold. Set $\gamma_t = \gamma_0 \eta_t^{1/2}$, $\eta_t = \frac{5}{t+5}$ with $\gamma_0 = \min\left\{\frac{\eta_0^{1/2}}{8\sqrt{3}(L_g+\sigma_g+D_h\gamma^{2H})}, \frac{1}{\sigma_g\sqrt{3\mu}}\right\}$ and $H = 2\left(1-\gamma\right)^{-1}\log(T+4)$. Let $\lambda_t = 4\gamma_0\eta_t^{-1/2}$. Then for every integer $T \geq 1$, the output θ_T of HARPG (see Algorithm 2) satisfies*

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

where σ_g, σ_h, L_g are defined in Proposition 1 and Lemma 1 respectively, D_g and D_h are defined in Lemma 4, ε' and μ are defined in Lemma 2.

Our first step in the proof is an ascent-like lemma which is different from the previously used similar results of Lemma 5 and Lemma 6 since the HARPG algorithm does not use a normalized direction to update the policy parameter.

Lemma 8. *Let Assumptions 1, 3 and 4 be satisfied. Let (d_t) be any sequence of vectors in $\mathbb{R}^d$ and consider the sequence (θ_t) defined by:*

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

where $\theta_0 \in \mathbb{R}^d$. Then for every integer $t \geq 1$, for any positive step-size γ_t s.t. $\gamma_t \leq \frac{1}{2L_g}$, we have

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

Furthermore, we also have

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

where we use the same notations as in (25) to (28).

Proof. Using the smoothness of J (Proposition 1) and the update rule of the sequence (θ_t), we get

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

where the last inequality holds by Lemma 4. By Lemma 2 we have for any $\theta \in \mathbb{R}^d$

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

Using the above inequality and the assumption on the step-size, we arrive at (29). Adding $J^\star$ to both sides, taking expectation and using Jensen's inequality for $x \mapsto x^2$, we obtain (30). □

Given (30) in Lemma 8, we now provide a recursive bound for the sequence (V_t) as defined in (26).

Lemma 9. *Let Assumption 1 be satisfied and suppose $\eta_t \leq 1$. Let the sequence (d_t) be updated for every integer $t \geq 1$ via*

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

as in Algorithm 2. Then, for every integer t, the sequence (V_t) satisfies the following recursion

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

Proof. Define $\mathcal{V}_t \overset{\text{def}}{=} g(\tau_t, \theta_t) - \nabla J_H(\theta_t)$, $\mathcal{W}_t \overset{\text{def}}{=} \nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t) + B(\hat{\tau}_t, \hat{\theta}_t)(\theta_t - \theta_{t-1})$. Notice that $\mathbb{E}\left[\mathcal{V}_t\right] = 0$ and $\mathbb{E}\left[\|\mathcal{V}_t\|^2\right] \leq \sigma_g^2$. Moreover, we also have $\mathbb{E}\left[\mathcal{W}_t\right] = 0$ since

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

Furthermore, we can control the variance of the sequence $(\mathcal{W}_t)$ as follows:

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

where the last inequality stems from using the smoothness of the function J, the boundedness of the hessian estimate variance (see Proposition 1 for both results) and Lemma 4 (to control truncation errors).

We now rewrite our quantity of interest $d_t - \nabla J_H(\theta_t)$ using the update rule of the sequence (d_t) together with the definitions of the sequences $(\mathcal{V}_t)$ and $(\mathcal{W}_t)$ that we introduced for the purpose of this proof. We have

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

Using the above decomposition, we can derive the desired recursive upper bound on the sequence (V_t):

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

where the third identity follows from observing that the conditional expectation of the random variable $\eta_t \mathcal{V}_t + (1-\eta_t)\mathcal{W}_t$ w.r.t. θ_{t-1} is equal to zero and using the tower property of the conditional expectation, the first inequality uses the assumption $\eta_t \le 1$ and the last inequality utilizes the variance bounds established at the beginning of the proof of this lemma. □

Our next step is to combine the results of Lemma 8 and Lemma 9. For this purpose we use the sequence (Λ_t) defined in (27) as a Lyapunov function. Notice that it combines the quantities δ_t and V_t with a suitable weighting sequence (λ_t) (to be defined in the next result). For convenience, we recall here the definition of the sequence (Λ_t):

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

Lemma 10. *Let Assumptions 1, 3 and 4 be satisfied. Set* $\eta_t = \frac{5}{t+5}$ *and* $\gamma_t \le \gamma_0 \eta_t^{1/2}$. *Set* $\lambda_t = \lambda_0 \eta_t^{-1/2}$ *with* $\lambda_0 = 4\gamma_0$ *and* $\gamma_0 \le \frac{1}{8\sqrt{3}(L_g + \sigma_h + D_h \gamma^{2H})}$. *Then for any integer* $t \ge 1$ *either* $\delta_t \le \frac{\sqrt{2}\varepsilon'}{\sqrt{\mu}}$ *or*

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

Proof. By Lemma 8, we have

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

Lemma 9 provides the following inequality:

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

Summing up (32) with a $\lambda_t = \lambda_0 \eta_t^{-1/2}$ multiple of (33), we obtain

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

where (i) follows by $\lambda_0 = 4\gamma_0 \leq \left(48(2L_g^2 + \sigma_h^2 + D_h^2\gamma^{2H})\gamma_0\right)^{-1}$. The choice $\gamma_0 = \frac{\lambda_0}{4}$ and the bound $\lambda_t - \lambda_{t-1} \leq \frac{\lambda_0}{2}\eta_t^{1/2}$ implies

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

thus (ii) holds. The last step (iii) follows by the assumption that $\delta_t \geq \frac{\sqrt{2}\varepsilon'}{\sqrt{\mu}}$.

Noticing that $\delta_t = \Lambda_t - \lambda_{t-1}V_t$, we have

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

where in (i) we use $-ax^2 + cx \leq \frac{c^2}{4a}$ with $x = V_t$, $a = \frac{\gamma_t\mu\lambda_{t-1}^2}{4}$, $c = \frac{1}{2}\gamma_t\left(\mu\lambda_{t-1}\Lambda_t - \frac{1}{6}\right)$.

□

End of the proof of Theorem 6 (and hence Theorem 2). By Lemma 10, we have that either $\delta_t \leq \frac{\sqrt{2}\epsilon'}{\sqrt{\mu}}$ or

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

Using Lemma 12 with $t_0 = 0$, $\tau = 5$, $r_t = \Lambda_t$, $\alpha_t = \eta_t$, $a = \frac{2}{5}$, $\beta_t = c_1\eta_t^{3/2} + c_2\eta_t^{-1/2}$, $c_1 = \left(\frac{1}{\mu\gamma_0} + 8\gamma_0\sigma_g^2\right)$, $c_2 = 24\lambda_0 D_g^2\gamma^{2H}$ we have for every integer $T \geq 1$

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

Selecting $\gamma_0 = \min\left\{\frac{1}{8\sqrt{3}(L_g+\sigma_g+D_h\gamma^{2H})}, \frac{1}{2\sigma_g\sqrt{2\mu}}\right\}$, we have $c_1 \leq \frac{8\sqrt{3}(L_g+\sigma_h+D_h\gamma^{2H})}{\mu} + \frac{2\sqrt{2}\sigma_g}{\sqrt{\mu}}$.

Since $\log x \leq x - 1$ for $x > 0$ we have

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

and thus $c_2 \leq 24\lambda_0 D_g^2\frac{1}{(T+4)^2}$.

Combining all the results together we obtain

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

H. Proof of Theorem 3 (N-HARPG)

We start by restating Theorem 3 in more details.

Theorem 7. *Let Assumptions 1, 3 and 4 hold. Set* $\gamma_t = \frac{3}{\sqrt{2\mu}(t+2)} = \frac{6M_g}{\mu_F(t+2)}$ *and* $\eta_t = \frac{2}{t+2}$, $H = \frac{3}{2}(1-\gamma)^{-1}\log(T+1)$. *Then for every integer* $T \geq 1$, *the output* θ_T *of* N-HARPG *(see Algorithm 2) satisfies*

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

where σ_g, σ_h, L_g *are defined in Proposition 1 and Lemma 1 respectively,* D_g *and* D_h *are defined in Lemma 4,* ε' *and* μ *are defined in Lemma 2.*

Similarly to the proof of Theorem 1 for N-PG-IGT, our starting point is the ascent-like result provided by Lemma 6 which holds independently of the update direction (d_t). In view of the bound (13), we need to control the error term

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

Recall that we have already provided a bound for this quantity in Lemma 9. We are now ready to provide the proof of Theorem 7.

End of the proof of Theorem 7 (and hence Theorem 3). Using Lemma 6 (note that this lemma does not use the specific update rule of the sequence (d_t) and can be used here) and applying Jensen's inequality for $x \mapsto x^2$, we have

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

Then, using Lemma 9 and observing that $\|\theta_{t+1} - \theta_t\| = \gamma_t$ as a consequence of the normalized update rule of Algorithm 2 (N-HARPG), we have

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

Applying Lemma 12 with $r_t = V_t$, $\alpha_t = \eta_{t+1}$, $a = 1$, $\beta_t = c\cdot\eta_t^2 + \bar{c}$, $c = 2\sigma_g^2 + 12(2L_g^2 + \sigma_h^2 + D_h^2\gamma^{2H})\gamma_0^2$, $\bar{c} = 24D_g^2\gamma^{2H}$, $\tau = 2, t_0 = 0$, we get

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

Combining inequalities (36) and (38), we obtain

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

Again, applying Lemma 12 with $r_t = \delta_t$, $\alpha_t = \eta_t$, $a = 1$, $\tau = 2$, $t_0 = 0$, $\beta_t = c_1\eta_t^2 + c_2\eta_t^{3/2} + c_3\eta_t + c_4\eta_t^{1/2}$,

$c_1 = \left(\frac{8\gamma_0\sigma_g}{3} + \frac{L_g\gamma_0^2}{2}\right)$, $c_2 = \frac{8\sqrt{2c}\gamma_0}{3}$, $c_3 = \frac{\gamma_0\varepsilon'}{3}$, $c_4 = \frac{16\sqrt{2\bar{c}}\gamma_0}{3}$ we get for every integer $T \geq 1$

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

The choice $H = \frac{3}{2}(1-\gamma)^{-1}\log(T+1)$ concludes the proof.

I. Convergence Analysis of Normalized-Momentum Policy Gradient Method (N-MPG)

In this section, we present a global convergence result for a simple Normalized Momentum Policy Gradient algorithm which we name N-MPG (see Algorithm 3). Unlike N-PG-IGT which enjoys a $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ global sample complexity (to reach a globally optimal policy up to the bias due to policy parameterization), N-MPG does not use a look-ahead step and our analysis only guarantees a $\tilde{\mathcal{O}}(\varepsilon^{-3})$ global sample complexity for this algorithm.

Algorithm 3 N-MPG (Normalized-Momentum Policy Gradient)

```
Input: $\theta_0, T, \{\eta_t\}_{t\geq 0}, \{\gamma_t\}_{t\geq 0}$
$\tau_0 \sim p(\cdot|\pi_{\theta_0})$
$d_0 = g(\tau_0, \theta_0)$
$\theta_1 = \theta_0 + \gamma_0 \frac{d_0}{\|d_0\|}$
for $t = 1, \ldots, T-1$ do
    $\tau_t \sim p(\cdot|\pi_{\theta_t})$
    $d_t = (1-\eta_t)d_{t-1} + \eta_t g(\tau_t, \theta_t)$
    $\theta_{t+1} = \theta_t + \gamma_t \frac{d_t}{\|d_t\|}$
end for
return $\theta_T$
```

Theorem 8. *Let Assumptions 1, 3 and 4 hold. Set $\gamma_t = \frac{6}{\sqrt{2\mu}(t+2)}$, $\eta_t = \left(\frac{2}{t+2}\right)^{2/3}$, $H = \frac{5}{3}(1-\gamma)^{-1}\log(T+1)$. Then for every integer $T \geq 1$, the output θ_T of* N-MPG *(Algorithm 3) satisfies*

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

where σ_g, L_g are defined in Proposition 1, D_g is defined in Lemma 4, ε' and μ are defined in Lemma 2.

The following corollary provides the sample complexity of the algorithm, i.e., the number of observed state-action pairs to achieve a global optimality precision $\varepsilon > 0$ up to the error floor $\varepsilon'/\sqrt{2\mu}$ related to the function approximation power of the policy parameterization.

Corollary 2. *In the setting of Theorem 8, for every $\varepsilon > 0$, the output θ_T of* N-MPG *satisfies $J^* - \mathbb{E}\left[J(\theta_T)\right] \leq \varepsilon + \varepsilon'/\sqrt{2\mu}$ for $T = \mathcal{O}(\varepsilon^{-3})$, i.e., the total sample complexity is of the order $\tilde{\mathcal{O}}(\varepsilon^{-3})$.*

Proof of Theorem 8. Combining Lemma 6 with Lemma 11 we obtain

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

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

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

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

Setting $H = \frac{5}{3}(1-\gamma)^{-1}\log(T+1)$ concludes the proof.

Lemma 11. *Let Assumption 1 be satisfied. Let (d_t) and (θ_t) be the sequences generated by the* N-MPG *algorithm (see Algorithm 3) with $\eta_t = \left(\frac{2}{t+2}\right)^{2/3}$ and $\gamma_t = \frac{6}{\sqrt{2\mu}(t+2)}$. Then for any integer $t \geq 0$ it holds*

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

where $C(p,q)$ for $q \in [0,1)$, $p > 0$ is defined in Lemma 15.

Proof. Define $\hat{e}_t \overset{\text{def}}{=} d_t - \nabla J_H(\theta_t)$ and $e_t \overset{\text{def}}{=} g(\tau_t, \theta_t) - \nabla J_H(\theta_t)$. Then the update rule of the sequence (d_t) implies

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

Unrolling the recursion and defining $S_t \overset{\text{def}}{=} \nabla J_H(\theta_{t-1}) - \nabla J_H(\theta_t)$, we have

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

Define for every integer $t \geq 1$ the σ-field $\mathcal{F}_t \overset{\text{def}}{=} \sigma(\{\theta_0, \tau_0, \ldots, \tau_{t-1}\})$ where $\tau_s \sim p(\cdot|\pi_{\theta_s})$ for every $0 \leq s \leq t-1$. Notice that for any integers $t_2 > t_1 \geq 1$ we have

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

Then using triangle inequality, taking expectation and applying Jensen's inequality, we get

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

where in (i) we utilize smoothness (Proposition 1), the fact that $\|\theta_{t+1}-\theta_t\| = \gamma_t$ and use independence $\mathbb{E}\left[\langle e_{t_1}, e_{t_2}\rangle\right] = 0$ for any $t_1 \neq t_2$. In the last inequality, we use $\mathbb{E}\left[\|e_{T-t}\|^2\right] \leq \sigma_g^2$.

Further, using Lemma 14 and 15 we have

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

where $C(p,q)$ for $q \in [0,1)$, $p > 0$ is defined in Lemma 15.

□

J. Technical Lemma

J.1. Lemma for solving recursions

The following auxiliary lemma is similar to (Stich, 2019, Lemma 7) and is useful to solve our recursions.

Lemma 12. *Let a be a positive real, τ a positive integer and let $\{r_t\}_{t\geq 0}$ be a non-negative sequence satisfying for every integer $t \geq 0$*

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

where $\{\alpha_t\}_{t\geq 0}, \{\beta_t\}_{t\geq 0}$ are non-negative sequences and $a\alpha_t \leq 1$ for all t. Then for $\alpha_t = \frac{2}{a(t+\tau)}$ we have for every integers $t_0, T \geq 1$

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

Proof. Notice that $1 - a\alpha_t = \frac{t+\tau-2}{t+\tau}$. Then for all $t \geq 0$

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

Multiplying both sides by $(t+\tau)^2$, we get

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

By summing this inequality from $t = t_0$ to $T-1$, we obtain

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

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

□

Lemma 13. *Let $\{r_t\}_{t\geq 0}$ be a non-negative sequence satisfying for every integer $t \geq 0$*

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

where $\alpha_t = \left(\frac{2}{t+2}\right)^{2/3}$ and $\{\beta_t\}_{t\geq 0}$ is a sequence of non-negative reals. Then for every integer T

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

Proof. Notice that $1 - \alpha_t = \frac{(t+2)^{2/3} - 2^{2/3}}{(t+2)^{2/3}}$. Then multiplying both sides of the recursion by $(t+\tau)^{1/3}$, we get for all $t \geq 0$

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

where the second inequality holds since $\left((t+2)^{2/3} - 2^{2/3}\right)(t+2)^{-1/3} = t^{2/3}(t+2)^{-1/3}\left(\left(1+\frac{2}{t}\right)^{2/3} - \frac{2^{2/3}}{t^{2/3}}\right) \leq t^{2/3}(t+2)^{-1/3}\left(1 + \frac{4}{3t} - \frac{2^{2/3}}{t^{2/3}}\right) \leq t^{2/3}(t+2)^{-1/3}$. By telescoping and dividing by $(T+1)^{1/3}$, we obtain for every integer 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. Lemma for decreasing step-sizes estimates

Lemma 14. *Let $q \in [0,1]$ and let $\eta_t = \left(\frac{2}{t+2}\right)^q$ for every integer t. Then for every integer t and any integer $T \geq 1$ we have*

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

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

Proof. For every integer t we have

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

Using the above result, we can write

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

□

The following lemma is a generalization of (Gadat et al., 2018, Proposition B.1) holding for every integer $T \geq 1$ (instead of only for T sufficiently large).

Lemma 15. *Let $q \in [0,1)$, $p \geq 0$, $\gamma_0 > 0$ and let $\eta_t = \left(\frac{2}{t+2}\right)^q$, $\gamma_t = \gamma_0\left(\frac{2}{t+2}\right)^p$ for every integer t. Then for any integers t and $T \geq 1$, it holds*

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

where $C = C(p,q) \overset{def}{=} 2^{p-q}(1-q)^{-1}t_0\exp\left(2^q(1-q)t_0^{1-q}\right) + 2^{2p+1-q}(1-q)^{-2}$ *and* $t_0 \overset{def}{=} \max\left\{\left(\frac{p}{(1-q)2^q}\right)^{\frac{1}{1-q}}, 2\left(\frac{p-q}{(1-q)^2}\right)\right\}^{\frac{1}{1-q}}$.

Proof. We start the proof by writing

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

For $t \geq \max\left\{\left(\frac{p}{(1-q)2^q}\right)^{\frac{1}{1-q}}, 2\left(\frac{p-q}{(1-q)^2}\right)^{\frac{1}{1-q}}\right\} \overset{\text{def}}{=} t_0$ the function $\phi(t) = t^{-p}\exp\left(2^q(1-q)t^{1-q}\right)$ is increasing, therefore, the corresponding sum is bounded by

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

Notice that integrating by parts, we have

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

where the last inequality is satisfied by the choice of t_0 when $p \geq q$, and holds trivially for $p \leq q$. Thus,

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

Finally, combining everything together, we have

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

where $C = C(p,q) \overset{\text{def}}{=} 2^{p-q}(1-q)^{-1}t_0 \exp\left(2^q(1-q)t_0^{1-q}\right) + 2^{2p+1-q}(1-q)^{-2}$. □