

Variance-Reduced Policy Gradient is Nearly Minimax-Optimal

Leonardo Cesani¹
Politecnico di Milano

Matteo Papini
Università degli Studi di Milano

Marcello Restelli
Politecnico di Milano

Abstract

Standard Policy Gradient (PG) algorithms suffer from high estimation variance because they rely on Monte Carlo updates. While off-policy variance reduction mitigates this problem, establishing optimal sample complexity currently requires either restrictive assumptions, such as uniformly bounded importance weights, or impractical second-order operations. In this work, we introduce Loopless Variance-Reduced PG (**LVR-PG**), a unified theoretical framework that abstracts a class of first-order estimators that integrate recursive updates with defensive importance sampling, including known estimators such as PAGE-PG and STORM-PG. We prove that this algorithmic class achieves $\tilde{\mathcal{O}}(\epsilon^{-2})$ last-iterate global convergence under weak gradient domination for general parametric policies. We establish this guarantee using purely first-order updates, with minimal assumptions and without gradient preconditioners. By extending state-of-the-art lower bounds to continuous parametric policies, we establish an $\Omega(\epsilon^{-2})$ sample complexity limit, confirming our framework is minimax-optimal up to logarithmic factors and horizon terms.

preferences (Ziegler et al., 2019; Bai et al., 2022). At their core, the algorithms powering these methods reduce to the Policy Gradient (PG, Deisenroth et al., 2013) framework.

Despite the empirical success, their *learning efficiency* is a matter of debate (Rafailov et al., 2023). Standard PG methods estimate the gradient of the expected return using sampled trajectories (Sutton et al., 1999), and the resulting Monte Carlo estimates can exhibit high estimation variance (Zhao et al., 2011). Actor-critic algorithms (Konda and Tsitsiklis, 2000) address this difficulty through a learned critic, whose approximation error can introduce bias. For LLMs, maintaining and updating an additional critic also entails substantial memory and computational costs, motivating actor-only methods that avoid a separate critic network (Yuan et al., 2023; Ahmadian et al., 2024). Improving trajectory efficiency without introducing a learned critic is therefore relevant to scalable policy optimization.

Reducing gradient-estimation variance has long motivated theoretical research on the *sample complexity* of PG methods, defined as the number of trajectories n required to learn an ϵ -optimal policy, where ϵ is the required accuracy. Recent actor-only applications give this established problem renewed practical relevance. A central theoretical objective is to determine whether purely first-order, actor-only methods can achieve the optimal $\Omega(\epsilon^{-2})$ sample-complexity lower bound (Azar et al., 2013). However, the existing guarantees for such methods do not match this lower bound and require additional assumptions on the Hessian of the log-policy or on the variance of the importance weights.

Within the class of actor-only PG methods, variance reduction seeks to improve sample complexity by reusing historical gradient information while retaining gradient-only updates. Unlike in standard stochastic optimization, the trajectory distribution depends on the policy parameters and therefore changes over the course of optimization. To account for this distribution shift, these methods employ Importance Sampling (IS), weighting trajectory-based estimates by likelihood ratios between the target and sampling

1 INTRODUCTION

In recent years, Reinforcement Learning (RL, Sutton et al., 1998) has been used as the central tool for aligning and scaling the reasoning capabilities of Large Language Models (LLMs, Brown et al., 2020; Achiam et al., 2023). Methods based on Proximal Policy Optimization (PPO, Schulman et al., 2017; Ouyang et al., 2022) and, more recently, Group Relative Policy Optimization (GRPO, Shao et al., 2024) have achieved enormous success on reasoning tasks (Guo et al., 2025) and aligning model outputs with human

¹Correspondence to: leonardo.cesani@polimi.it.

Table 1: Comparison of sample complexities and theoretical assumptions used for variance-reduced PG algorithms. *Complexity* denotes the number of trajectories required to reach the global optimal policy. *Convergence Type*, last-iterate or average-iterate guarantees, — denotes the necessary number of trajectories for any algorithm to return an ϵ -optimal policy. *Gradient Only* denotes an algorithm that does not use second-order or Fisher information. *No 2nd-Order Asm.* denotes algorithms which do *not* require assumptions on the Hessian matrix of the log-policy. *No Bounded IW Asm.* indicates algorithms not requiring an assumption on the boundedness of the variance of the importance weights. [†] The $\mathcal{O}(\cdot)$ hides dependencies on γ, C_1, L_S, χ and ψ , $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors. [‡] Requires an additional second-order smoothness assumption.

Algorithm	Complexity [†]	Convergence Type	Gradient Only	No 2nd-Order Asm.	No Bounded IW Asm.
GPOMDP (Montenegro et al., 2024)	$\tilde{\mathcal{O}}(\epsilon^{-3})$	Last-iterate	✓	✗	✓
SRVR-PG (Liu et al., 2020)	$\mathcal{O}(\epsilon^{-3})$	Average-iterate	✓	✗	✗
STORM-PG-S (Ding et al., 2022)	$\mathcal{O}(\epsilon^{-3})$	Average-iterate	✓	✗	✗
N-PG-IGT [†] (Fatkhullin et al., 2023)	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	Last-iterate	✓	✗	✓
HARPG [‡] (Fatkhullin et al., 2023)	$\tilde{\mathcal{O}}(\epsilon^{-2})$	Last-iterate	✗	✗	✓
ANPG (Mondal and Aggarwal, 2024)	$\mathcal{O}(\epsilon^{-2})$	Average-iterate	✗	✗	✓
LVR-PG (Theorem 5.1)	$\tilde{\mathcal{O}}(\epsilon^{-2})$	Last-iterate	✓	✓	✓
Lower Bound (Theorem 3.1)	$\Omega(\epsilon^{-2})$	—	✓	✓	✓

policies. Inspired by stochastic non-convex optimization, early methods such as SVR-PG (Papini et al., 2018) and SRVR-PG (Xu et al., 2019) employ double-loop architectures. These compute large-batch gradient estimates at a fixed anchor policy to correct subsequent inner-loop mini-batch updates, making the refresh schedule dependent on the chosen epoch length. Loopless methods instead incorporate historical information through recursive momentum updates or probabilistic refreshes, as in STORM-PG (Yuan et al., 2020; Ding et al., 2022) and PAGE-PG (Gargiani et al., 2022), respectively.

Existing analyses of these importance-sampling-based methods impose strong statistical assumptions, notably a uniform bound on the variance of Importance Weights (IW) (Liu et al., 2020; Ding et al., 2022). This condition is not generally guaranteed in continuous state-action spaces, where trajectory-level likelihood ratios accumulate the distribution shifts induced by policy updates. Paczolay et al. (2024) addresses this by using *defensive* IS (Owen and Zhou, 2000), but it only studies convergence to a first-order stationary point. In contrast, N-PG-IGT (Fatkhullin et al., 2023) circumvents the IW assumption through implicit gradient transport, retaining historical gradient information without using importance sampling. Its established sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$, however, does not attain the optimal accuracy dependence. Other approaches attain this dependence through different information mechanisms. HARPG uses Hessian-vector products (Fatkhullin et al., 2023), whereas ANPG estimates natural-gradient directions through an auxiliary stochastic optimization problem involving Fisher

geometry (Mondal and Aggarwal, 2024). Despite their theoretical appeal, these approaches require additional computational overhead beyond policy-gradient estimation, motivating the study of gradient-only methods. Their analyses also impose regularity assumptions on policy derivatives, as summarized in Table 1.

The question is therefore whether existing gradient-only PG methods can achieve last-iterate global convergence with the optimal accuracy dependence, without assuming uniformly bounded importance-weight variance or additional second-order regularity.

Original Contributions. We study a unified theoretical framework encompassing existing gradient-only PG estimators equipped with defensive IS. In the remainder of the paper, we present our main contributions as follows:

- In Section 3, we adapt the hard-instance construction of Azar et al. (2013) to continuously parameterized policies satisfying WGD, establishing that any algorithm guaranteeing expected ϵ -optimality requires $\Omega(\epsilon^{-2})$ trajectories in the worst-case.
- In Section 4, we introduce the unified **LVR-PG** framework, abstracting loopless estimators like PAGE-PG (Gargiani et al., 2022) and STORM-PG (Yuan et al., 2020). By integrating *defensive* importance sampling (Paczolay et al., 2024) to control off-policy variance, we avoid assumptions on the variance of the importance weights.
- In Section 5, we prove **LVR-PG** achieves last-iterate global convergence with a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$, matching our lower bound

up to logarithmic factors, without assuming uniformly bounded importance-weight variance or additional second-order regularity.

- In Section 6, we empirically benchmark the **LVR-PG** framework against state-of-the-art baselines, identifying the defensive variant of STORM-PG as the dominant algorithmic configuration with superior sample efficiency. The full experimental evaluation is provided in Appendix I.

2 PRELIMINARIES

Notation. For a measurable set $\mathcal{X}$, we denote by $\Delta_{\mathcal{X}}$ the set of probability measures over $\mathcal{X}$. For $P \in \Delta_{\mathcal{X}}$, if it admits a density, we denote it by p . We will use $x \sim P$ to express that the random variable x is distributed according to P . For $n \in \mathbb{N}$, we let $\llbracket n \rrbracket := \{1, \dots, n\}$. Vectors are indicated with bold letters. Given two symmetric matrices A and B , we say that $A \succeq B$ if the matrix $A - B$ is positive semidefinite.

Interaction Protocol. We consider an agent interacting with a Markov Decision Process (MDP, Puterman, 2014) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \rho_0, \gamma)$ over independent, indefinite-length episodes. In each timestep $h = 0, 1, \dots$, the agent observes a state $\mathbf{s}_h \in \mathcal{S} \subseteq \mathbb{R}^{d_{\mathcal{S}}}$, where $\mathbf{s}_0 \sim \rho_0$ is the initial state and $\rho_0 \in \Delta_{\mathcal{S}}$ is the initial-state distribution, and executes an action $\mathbf{a}_h \in \mathcal{A} \subseteq \mathbb{R}^{d_{\mathcal{A}}}$ sampled from a parametric stochastic policy $\pi_{\theta} : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}$. Here, $\theta \in \Theta \subseteq \mathbb{R}^{d_{\theta}}$ is the parameter vector defining the policy space $\Pi_{\Theta} = \{\pi_{\theta} \mid \theta \in \Theta\}$. The agent then receives a reward $r_{h+1} := r(\mathbf{s}_h, \mathbf{a}_h) \in [-R_{\max}, R_{\max}]$ and transitions to a next state $\mathbf{s}_{h+1} \sim p(\cdot \mid \mathbf{s}_h, \mathbf{a}_h) \in \Delta_{\mathcal{S}}$. For a discount factor $\gamma \in [0, 1)$, the performance of θ is assessed via the *expected discounted return* under the assumption that $\theta \mapsto \pi_{\theta}$ is differentiable. The latter is defined as:

$$J(\theta) := \mathbb{E} \left[\sum_{h \geq 0} \gamma^h r(\mathbf{s}_h, \mathbf{a}_h) \mid \pi_{\theta} \right],$$

where the expectation is taken over the trajectory distribution induced by the policy and the environment dynamics. In the following, we will refer to $J : \Theta \rightarrow \mathbb{R}$ as the *performance function*, which is generally non-convex. Because the per-step rewards are bounded, it satisfies $|J(\theta)| \leq R_{\max}/(1 - \gamma)$ for all $\theta \in \Theta$. We denote the *optimal* performance as $J^* := \max_{\theta \in \Theta} J(\theta)$.

Policy Gradients. We consider learning algorithms that iteratively update a sequence of policies $\pi_{\theta_1}, \dots, \pi_{\theta_K}$. At each iteration k , the agent collects a dataset $\mathcal{D}_k = \{\tau_1^k, \dots, \tau_{n_k}^k\}$ of n_k independent trajectories drawn from the trajectory distribution induced by the policy $p_{\theta_k}(\tau)$. To accommodate struc-

tural or safety constraints common in practical applications, we consider PG methods that restrict the parameter space to a closed convex set $\Theta \subseteq \mathbb{R}^{d_{\theta}}$ and, similarly to Xu et al. (2019), update the parameters via projected stochastic gradient ascent: $\theta_{k+1} \leftarrow \mathcal{P}_{\Theta} \left(\theta_k + \eta \widehat{\nabla}_{\theta} J(\theta_k) \right)$. Here, $\eta > 0$ is the learning rate, and $\widehat{\nabla}_{\theta} J(\theta_k)$ is an *unbiased* estimator of the true gradient $\nabla_{\theta} J(\theta_k)$. We formalize the optimization geometry over the parameter domain Θ using the generalized gradient mapping $\mathcal{G}_{\eta}(\theta) := \frac{1}{\eta} (\mathcal{P}_{\Theta}(\theta + \eta \nabla J(\theta)) - \theta)$. This mapping acts as the projected gradient at θ , recovering the standard gradient $\nabla_{\theta} J(\theta)$ in the unconstrained case $\Theta = \mathbb{R}^{d_{\theta}}$. To compute the stochastic gradient estimator $\widehat{\nabla}_{\theta} J(\theta)$ from finite interaction data without introducing approximation or truncation bias, we evaluate the infinite-horizon discounted objective using a randomized geometric stopping rule, in which the interaction with the environment is terminated after each transition with probability $1 - \gamma$, independently of the interaction history. For each trajectory τ_i , let $H_{\gamma,i} \in \{1, 2, \dots\}$ denote the number of executed transitions, with stopping randomness independent across trajectories. In this way, the expected length of each trajectory is $\mathbb{E}[H_{\gamma,i}] = (1 - \gamma)^{-1}$, and, since $\Pr(H_{\gamma,i} > h) = \gamma^h$, the expected undiscounted return accumulated before stopping matches the infinite-horizon discounted objective function $J(\theta)$. For theoretical analysis, we adopt the geometric stopping rule to obtain unbiased estimates of the infinite-horizon objective. In the empirical evaluation, we use a fixed horizon. Using the geometric horizons $H_{\gamma,i}$, the estimate $\widehat{\nabla}_{\theta} J(\theta_k)$ is computed from the on-policy dataset $\mathcal{D}_k$ via the GPOMDP (Baxter and Bartlett, 2001) estimator:

$$\begin{aligned} \widehat{\nabla}_{\theta} J(\theta_k) &= \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{g}(\tau_i \mid \theta_k), \\ \mathbf{g}(\tau_i \mid \theta) &:= \sum_{h=0}^{H_{\gamma,i}-1} \left(\sum_{l=0}^h \nabla_{\theta} \log \pi_{\theta}(\mathbf{a}_l \mid \mathbf{s}_l) \right) r_{h+1}. \end{aligned} \quad (1)$$

While the variance of this on-policy estimator is bounded under standard regularity conditions (Zhao et al., 2011; Papini et al., 2022), it is often very large in practice, compromising learning efficiency.

Off-Policy Gradient Estimation. To mitigate large variance of the estimator, *variance-reduced* algorithms reuse historical data to improve sample efficiency, requiring the evaluation of the gradient at a *target* policy θ using trajectories collected by a different *behavioral* policy θ' . To maintain an unbiased gradient estimate, these methods rely on Importance Sampling (IS, Metelli et al., 2020). Given a trajectory $\tau \sim p(\cdot \mid \theta')$, the off-policy GPOMDP estimator is

defined as $\widehat{\nabla}_{\theta} J(\theta) = \frac{1}{n} \sum_{i=1}^n \omega(\theta, \tau_i) \mathbf{g}(\tau_i | \theta)$, where $\omega(\theta, \tau_i)$ is the importance weight defined by the trajectory density ratio:

$$\omega(\theta, \tau_i) := \frac{p_{\theta}(\tau_i)}{p_{\theta'}(\tau_i)} = \prod_{h=0}^{H_{\tau,i}-1} \frac{\pi_{\theta}(\mathbf{a}_h | \mathbf{s}_h)}{\pi_{\theta'}(\mathbf{a}_h | \mathbf{s}_h)},$$

where the environment transition dynamics and the trajectory's stopping probability cancel out. While this off-policy estimator is unbiased, its variance scales with the χ^2 -divergence between the trajectory distributions induced by θ and θ' (Cortes et al., 2010). Since the χ^2 -divergence is generally unbounded, the importance-weight variance can easily diverge even for simple continuous policy classes, such as Gaussians (Zhang et al., 2021).

3 LEARNING FRAMEWORK

To study the sample complexity of learning a *globally optimal policy*, we must define the structural and geometric properties of the objective landscape and establish the statistical limits of policy optimization.

Assumptions. To study variance-reduced PG algorithms within a unified framework, we define the structural and geometric properties of the objective landscape that we will use in our analysis. First, we enforce the following standard regularity conditions on the policy parametrization.

Assumption 3.1 (Regularity). *The policy class admits finite constants $C_1, L_S > 0$ such that, for all $\theta, \theta' \in \Theta$, and for any $\mathbf{s} \in \mathcal{S}$, we have:*

- (i) $\sup_{\mathbf{a} \in \mathcal{A}} \|\nabla_{\theta} \log \pi_{\theta}(\mathbf{a} | \mathbf{s})\|_2 \leq C_1$,
- (ii) $\mathbb{E}_{\mathbf{a} \sim \pi_{\theta}} \left[\|\nabla_{\theta} \log \pi_{\theta}(\mathbf{a} | \mathbf{s}) - \nabla_{\theta'} \log \pi_{\theta'}(\mathbf{a} | \mathbf{s})\|_2^2 \right] \leq L_S^2 \|\theta - \theta'\|_2^2$.

Assumption 3.1 (ii), introduced by Paczoly et al. (2024), relaxes the uniform Lipschitz score condition standard in prior literature (Papini et al., 2018; Xu et al., 2019; Yuan et al., 2020; Gargiani et al., 2022; Fatkhullin et al., 2023) by requiring only mean-squared smoothness under the on-policy distribution, rather than a worst-case bound across all state-action pairs. Condition (i) requires bounded score gradients, which softmax policies satisfy, for example. Under these regularity conditions, the performance function $J(\theta)$ and its stochastic gradient estimator are smooth and have bounded variance, respectively, as formalized in the following proposition (proof in Appendix D).

Proposition 3.1. *Under Assumption 3.1, the following properties hold.*

- (i) $\text{Var}(\mathbf{g}(\tau | \theta)) \leq \frac{4R_{\max}^2 C_1^2}{(1-\gamma)^3} =: \nu^2$;
- (ii) $\|\nabla_{\theta}^2 J(\theta)\|_2 \leq \frac{R_{\max}(C_1^2 + L_S)}{(1-\gamma)^2} =: L$.

Finally, to ensure global convergence and prevent the learning trajectory from getting trapped in suboptimal solutions, we impose a standard geometric assumption on the objective landscape (Yuan et al., 2022; Montenegro et al., 2024).

Assumption 3.2 (Weak Gradient Domination (WGD)). *There exist $\chi > 0$ and $\psi \geq 0$ such that, for every $\theta \in \Theta$ and a step size $\eta \leq 1/L$, it holds that:*

$$J^* - J(\theta) \leq \chi \|\mathcal{G}_{\eta}(\theta)\|_2 + \psi$$

Assumption 3.2 is the standard for global convergence of constrained stochastic optimization (Yuan et al., 2022; Masiha et al., 2022; Ding et al., 2022; Fatkhullin et al., 2023; Montenegro et al., 2024). Note that in the unconstrained setting, where $\Theta = \mathbb{R}^{d_{\theta}}$, the mapping strictly recovers the true gradient $\mathcal{G}_{\eta}(\theta) = \nabla_{\theta} J(\theta)$, meaning this condition reduces to the standard WGD assumption on the gradient norm. When $\psi = 0$, we recover the (strong) *gradient domination* (GD) property: $J^* - J(\theta) \leq \chi \|\mathcal{G}_{\eta}(\theta)\|_2$. GD is stricter than WGD since it requires that the constrained performance function $J(\theta)$ has no suboptimal stationary points within the feasible domain. Instead, WGD admits local maxima as long as their performance is ψ -close to the global optimum (Montenegro et al., 2024).

Lower Bound for Policy Optimization. Having established the learning framework, we now investigate the statistical limits of the policy optimization problem.

Theorem 3.1 (Sample Complexity Lower Bound). *For any discount factor $\gamma \in [2/3, 1)$ and sufficiently small target accuracy $0 < \epsilon < \mathcal{O}\left(\frac{R_{\max}}{\gamma(1-\gamma)}\right)$, there exists a class of MDPs $\mathcal{M}$ and a policy parametrization satisfying Assumptions 3.1 and 3.2, such that any algorithm guaranteeing an expected sub-optimality gap $J^* - \mathbb{E}[J(\theta_K)] \leq \epsilon$ must collect a number of trajectories n satisfying:*

$$n = \Omega\left(\frac{R_{\max}^2}{(1-\gamma)^2} \epsilon^{-2}\right).$$

Theorem 3.1 establishes an $\Omega(\epsilon^{-2})$ statistical barrier under our regularity and WGD assumptions. In the

context of RL, however, our result shows that these geometric properties do not trivialize the learning problem. Even when WGD holds with zero bias, so that GD excludes suboptimal stationary points, it cannot eliminate the irreducible estimation variance induced by the stochastic transition dynamics, thereby preserving the $\Omega(\epsilon^{-2})$ statistical barrier. This accuracy dependence is consistent with variance-reduced stochastic optimization under GD, where an $\mathcal{O}(\epsilon^{-2})$ sample complexity is achievable Fatkhullin et al. (2022).

However, existing lower bounds do not directly show this barrier in our setting. First, lower bounds for bounded-variance stochastic-gradient oracles concern algorithms that observe stochastic gradients Masiha et al. (2026), whereas MDP interaction provides additional information through states, actions, and rewards, enabling importance-weighted gradient evaluations across different policies using the same trajectories. Second, the tabular lower bound of Azar et al. (2013) does not automatically apply after restricting the MDP and policy class to satisfy GD. Theorem 3.1 addresses these limitations by establishing an $\Omega(\epsilon^{-2})$ sample complexity lower bound for policy optimization under the stated regularity and WGD assumptions.

Finally, Theorem 3.1 establishes a statistical limit for policy optimization, independently of the update rule. Although existing methods achieve near-optimal sample complexity using Hessian or Fisher information, the lower bound does not imply that such information is necessary to reach this statistical limit.

Matching the Lower Bound. Theorem 3.1 establishes the statistical limit, leaving open whether first-order policy optimization methods attain the optimal accuracy dependence. In non-convex stochastic optimization under WGD, Masiha et al. (2026) establishes an $\Omega(\epsilon^{-3})$ lower bound for the gradient-only, bounded-variance oracle model. Improving this dependence requires additional oracle structure, which recursive variance-reduction methods exploit through “double sampling”: gradient evaluations at distinct parameters using shared randomness. In PG, implementing this coupling requires correcting the distribution shift induced by policy updates. Alternative approaches avoid importance sampling: N-PG-IGT provides a $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ guarantee, whereas HARPG and ANPG attain the ϵ^{-2} dependence through Hessian-vector products and Fisher geometry, respectively (Fatkhullin et al., 2023; Mondal and Aggarwal, 2024). We therefore investigate whether we can establish the optimal accuracy dependence of this first-order recursive framework without relying on Hessian-vector products or Fisher-based updates, potentially without assuming uniformly bounded importance-weight variance.

4 THE LVR-PG FRAMEWORK

To address the limitations identified in Section 3, we introduce *Loopless Variance-Reduced Policy Gradient* (**LVR-PG**), a unified framework for analyzing loopless, gradient-only PG estimators. By suitably setting its parameters, **LVR-PG** recovers state-of-the-art estimators such as STORM-PG (Yuan et al., 2020) and PAGE-PG (Gargiani et al., 2022), equipped with the defensive IS scheme of Paczolay et al. (2024), bounding the IW and removing the need to assume their variance to be uniformly bounded. Our goal is to establish that, with an appropriate choice of hyperparameters, an algorithm within **LVR-PG** achieves last-iterate global convergence with a sample complexity matching the $\Omega(\epsilon^{-2})$ lower bound of Theorem 3.1, up to logarithmic factors and horizon terms.

Defensive Importance Sampling Classical variance-reduced PG evaluates past policies using on-policy data, leading to IW to diverge exponentially over the trajectory horizon. To avoid this blow-up in the IW, we sample trajectories from a mixture trajectory distribution and employ *defensive IS* (Owen and Zhou, 2000; Paczolay et al., 2024). Let $p_{\theta}(\tau)$ be the probability density of a trajectory τ induced by a policy π_{θ} . During off-policy estimation at iteration k , the algorithm collects a mini-batch of B trajectories. Instead of drawing these exclusively from the current policy, each trajectory is independently drawn from the target policy $\pi_{\theta_{k+1}}$ with probability $\xi \in [0, 1]$ and from the previous policy π_{θ_k} with probability $1 - \xi$. This yields a mixture trajectory distribution, defined as $q_{\xi}(\tau) = \xi p_{\theta_{k+1}}(\tau) + (1 - \xi) p_{\theta_k}(\tau)$. To estimate the gradient at either θ_{k+1} or θ_k using trajectories sampled from q_{ξ} , we use the defensive IS weight $\omega_{\xi}(\theta, \tau) := p_{\theta}(\tau)/q_{\xi}(\tau)$, where $\theta \in \{\theta_{k+1}, \theta_k\}$. For $\xi \in (0, 1)$, each target trajectory density enters q_{ξ} with a positive coefficient, so its importance weight is bounded by the reciprocal of that coefficient. In particular, when evaluating the current policy, we have $\omega_{\xi}(\theta_{k+1}, \tau) \leq 1/\xi$, and similarly $\omega_{\xi}(\theta_k, \tau) \leq 1/(1 - \xi)$. One can easily show that the gradient estimate computed using this mixture distribution remains unbiased.

The Unified Estimator. Building upon the loopless variance reduction principles introduced in Yuan et al. (2020) and Gargiani et al. (2022), our framework regulates full-batch and mini-batch updates through a randomized mechanism governed by $p \in [0, 1]$, which we call *refresh probability* parameter. At iteration $k \in \llbracket K \rrbracket$, the evaluation of the target policy θ_{k+1} and the construction of the new gradient estimator $\hat{\nabla}_{\theta}^{\dagger} J(\theta_{k+1})$ depend on a Bernoulli realization $c_k \sim \mathcal{B}(p)$. If $c_k = 1$,

Algorithm 1: LVR-PG

Input: $K, N, B, p, \alpha, \xi, \eta, \Theta, \theta_0$
 1 Init. $\widehat{\nabla}_{\theta} J(\theta_0)$ using N trajectories from π_{θ_0} ;
 2 **for** $k = 0, \dots, K - 1$ **do**
 3 $\theta_{k+1} \leftarrow \mathcal{P}_{\Theta} \left(\theta_k + \eta \widehat{\nabla}_{\theta}^{\dagger} J(\theta_k) \right)$;
 4 Draw refresh variable $c_k \sim \mathcal{B}(p)$;
 5 **if** $c_k = 1$ **then**
 6 Sample N trajectories $\tau_{1:N} \sim \pi_{\theta_{k+1}}$;
 7 Estimate gradient with Equation (1);
 8 **else**
 9 Sample B trajectories $\tau_{1:B}$ from q_{ξ} ;
 10 Estimate gradient with Equation (2);
 11 **Return** θ_K

the framework executes a full refresh, sampling N trajectories directly from the target policy $\pi_{\theta_{k+1}}$ to compute a full-batch gradient estimate with Equation (1) with $n_k = N$. If $c_k = 0$, the framework bypasses the full-batch estimation in favor of a recursive mini-batch update. To evaluate the gradients of both $\pi_{\theta_{k+1}}$ and π_{θ_k} simultaneously, **LVR-PG** collects $B \ll N$ trajectories independently drawn from the mixture behavior distribution q_{ξ} . This allows us to compute the Defensive IS operator $\omega_{\xi}(\theta, \tau)$ via the density ratio. Using this mini-batch and the mixture IS operator, the gradient estimator is recursively updated using a *momentum* parameter $\alpha \in [0, 1]$ to control off-policy variance:

$$\begin{aligned} \widehat{\nabla}_{\theta}^{\dagger} J(\theta_{k+1}) &= g_B^{\omega_{\xi}}(\theta_{k+1}) \\ &+ (1 - \alpha) \left[\widehat{\nabla}_{\theta}^{\dagger} J(\theta_k) - g_B^{\omega_{\xi}}(\theta_k) \right], \end{aligned} \quad (2)$$

where $g_B^{\omega_{\xi}}(\theta) := \frac{1}{B} \sum_{i=1}^B \omega_{\xi}(\theta, \tau_i) g(\tau_i | \theta)$. The momentum parameter α regulates the degree to which past gradient errors are discounted, and enables to control of the variance-reduction strategy to use: setting $\alpha = 0$ and $p \in [0, 1]$ recovers the update rule of PAGE-PG (Gargiani et al., 2022), while setting $\alpha > 0$ yields the momentum update of STORM-PG (Yuan et al., 2020). In the rest of the paper, we will refer to the defensive versions of these two algorithms as DEF-PAGE-PG and DEF-STORM-PG, respectively. Finally, once we compute the variance-reduced estimator $\widehat{\nabla}_{\theta}^{\dagger} J(\theta_k)$, we update the policy parameters via projected stochastic gradient ascent. The algorithm aims to find a parameter $\theta \in \Theta$ that maximizes the performance function subject to this constraint.

5 CONVERGENCE AND OPTIMALITY

In this section, we establish the sample complexity required by an instantiation of the **LVR-PG** framework to

obtain an ϵ -approximate global optimum. Specifically, we demonstrate that the **LVR-PG** framework defines a class of minimax-optimal first-order algorithms, up to logarithmic factors. The proofs for the results presented in this section are deferred to Appendix E.

Theorem 5.1 (Global Convergence of **LVR-PG**). *Under Assumptions 3.1 and 3.2, for any target accuracy $\epsilon > 0$ and fixed mini-batch size $B \geq 1$, let the batch sizes be $N, N_0 = \Theta(\epsilon^{-2})$ and the learning rate be $\eta = \Theta(B\epsilon)$. For appropriately chosen hyperparameters satisfying $p, \alpha = \mathcal{O}(B\epsilon^2)$ and $p + \alpha = \Omega(B\epsilon^2)$ and choosing $\xi = 1/2$, **LVR-PG** guarantees an expected suboptimality gap $J^* - \mathbb{E}[J(\theta_K)] \leq \psi + \epsilon$ with a total expected sample complexity of at most:*

$$n = \tilde{\mathcal{O}} \left(\frac{\chi^3 R_{\max}^2}{(1 - \gamma)^4 \epsilon^2} \right),$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

Proof Sketch. The proof establishes the minimax-optimal sample complexity for **LVR-PG** by coupling the policy's suboptimality descent with the variance recurrence of the gradient estimator.

Part I: The Descent Mechanics. We begin by characterizing the estimation error $\Delta_k := \mathbb{E}[\|\widehat{\nabla}_{\theta}^{\dagger} J(\theta_k) - \nabla_{\theta} J(\theta_k)\|_2^2]$. Lemma D.4 quantifies how this error propagates through recursive updates. Letting $\lambda := 1 - (1 - p)(1 - \alpha)^2$ and $\Psi_k := \mathbb{E}[\|\theta_{k+1} - \theta_k\|_2^2]$ we have:

$$\Delta_{k+1} \leq (1 - \lambda)\Delta_k + \frac{2(1 - \lambda)C_{\dagger, \omega}}{B}\Psi_k + C_{\nu}.$$

Here, $C_{\dagger, \omega} > 0$ is the mean-squared smoothness constant from Lemma B.1, which controls the estimation error induced by policy updates. The additive variance term C_{ν} combines the contributions of full-batch refreshes and recursive mini-batch updates, weighted by p and α^2 , respectively. From the L -smoothness of the objective obtained in Proposition 3.1, letting $r_k := \mathbb{E}[\mathcal{J}^* - \mathcal{J}(\theta_k)]$, we obtain the recursion of the suboptimality gap:

$$r_{k+1} \leq r_k - \frac{\eta}{8} \mathbb{E}[\|\mathcal{G}_{\eta}(\theta_k)\|_2^2] + \frac{3\eta}{4} \Delta_k - \left(\frac{1}{4\eta} - \frac{L}{2} \right) \Psi_k.$$

To establish convergence, we study the potential $\Phi_k := r_k + c\Delta_k$, combining suboptimality and estimation error. Choosing $c = \frac{3\eta}{2\lambda}$ and a sufficiently small step size η yields a recursion in which both objective improvement and estimator-error decay contribute to reducing Φ_k .

Part II: WGD and Recurrence. Because $\|\mathcal{G}_{\eta}(\theta_k)\|_2^2$ provides only local stationarity guarantees, we use Assumption 3.2 to bound the global suboptimality gap directly by the gradient mapping norm.

Evaluating the invariant sequence $x_k := \max\{0, \Phi_k\}$ reveals a dual-regime dynamical system:

$$x_{k+1} \leq x_k - \min \left\{ \frac{\mu\eta}{32} x_k^2, \frac{\lambda}{4} x_k \right\} + \frac{3\eta C_\nu}{2\lambda}.$$

This recurrence yields linear contraction for large values of the Lyapunov potential and a WGD-induced quadratic decrease for smaller values. The additive term $3\eta C_\nu/(2\lambda)$ accounts for stochastic gradient-estimation error. When the additive term is sufficiently small, Lemma H.2 bounds the iterations required to reach and remain within the target accuracy:

$$K \geq \left\lceil \max \left\{ \frac{32}{\mu\eta\epsilon}, \frac{4}{\lambda} \right\} \log \left(\frac{2\Phi_0}{\epsilon} \right) \right\rceil.$$

Part III: Selection of α, p and η . We now choose the hyperparameters to satisfy the accuracy condition $C_\nu = \mathcal{O}(\mu\lambda\epsilon^2)$ while controlling the sample cost. Setting $N = \Theta(\epsilon^{-2})$ and $p = \mathcal{O}(B\epsilon^2)$ keeps the expected cost of full-batch refreshes comparable to that of mini-batch updates. Choosing $\alpha = \mathcal{O}(B\epsilon^2)$ controls the momentum-dependent contribution to C_ν , while requiring $p + \alpha = \Omega(B\epsilon^2)$ ensures sufficient estimator-error decay through $\lambda = \Omega(B\epsilon^2)$. With a suitable step size $\eta = \Theta(B\epsilon)$, these choices yield $K = \tilde{\mathcal{O}}((B\epsilon^2)^{-1})$ iterations and a total sample complexity $n = \tilde{\mathcal{O}}(\epsilon^{-2})$. $\square$

Theorem 5.1 establishes that **LVR-PG** attains the minimax-optimal sample complexity in the learning framework described in Section 3, up to logarithmic factors and horizon terms, using exclusively first-order updates. Notably, this rate is achieved using a fixed step size $\eta = \Theta(B\epsilon)$, and the convergence guarantee applies directly to the last-iterate parameter θ_K , bypassing the need for trajectory averaging or random iterate selection commonly required in previous works (Liu et al., 2020; Ding et al., 2022). Furthermore, the analysis identifies a continuous two-dimensional region of hyperparameter configurations (p, α) for which the nearly optimal sample complexity is guaranteed. As illustrated in Figure 1, the conditions $p = \mathcal{O}(B\epsilon^2)$, $\alpha = \mathcal{O}(B\epsilon^2)$, and $p + \alpha = \Omega(B\epsilon^2)$, together with the batch-size and learning-rate choices established in the proof, provide sufficient conditions for this guarantee.

Remark 5.1 (Unification of Existing Algorithms). *Existing variance-reduced estimators emerge as boundary limits of this optimal space: setting $\alpha = 0$ or $p = 0$ recovers DEF-PAGE-PG and DEF-STORM-PG, respectively, both strictly preserving the $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity. Setting $p = 1$ forces full-batch evaluations at every step, recovering vanilla GPOMDP. This choice violates the computational boundary $p = \mathcal{O}(B\epsilon^2)$, shift-*

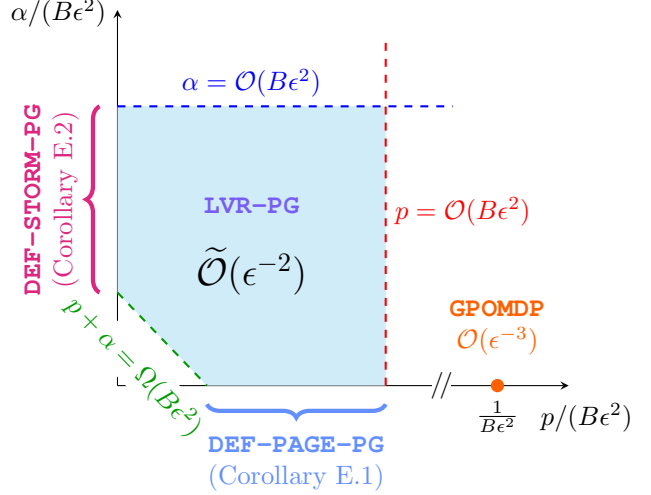


Figure 1: The continuous algorithmic design space of **LVR-PG** in normalized coordinates. The shaded region delineates the manifold of hyperparameter configurations (p, α) that satisfy the stability thresholds derived in Theorem 5.1. In this region, a configuration of **LVR-PG** is nearly optimal. GPOMDP falls out of this region, since $p = 1$.

ing the algorithm outside the optimal manifold and correctly degrading its complexity to the standard $\tilde{\mathcal{O}}(\epsilon^{-3})$ rate (Montenegro et al., 2024).

5.1 On Fisher-Non-Degeneracy

We specialize Theorem 5.1 to Fisher-Non-Degenerate (FND) policies in the unconstrained setting $\Theta = \mathbb{R}^{d_\Theta}$. This setting allows us to apply the ordinary-gradient domination inequality induced by FND under suitable approximation and concentrability assumptions (Liu et al., 2020; Ding et al., 2022; Fatkhullin et al., 2023), since $\mathcal{G}_\eta(\theta) = \nabla_\theta J(\theta)$. Although FND can also be imposed on constrained parameter sets, projection can remove gradient components, so the gradient bound does not imply the mapping-based WGD condition required by Assumption 3.2.

In the unconstrained setting $\Theta = \mathbb{R}^{d_\Theta}$, when the FIM induced by π_θ is uniformly non-degenerate, i.e., $\mathbf{F}(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(\mathbf{a} | \mathbf{s}) \nabla_\theta \log \pi_\theta(\mathbf{a} | \mathbf{s})^\top] \succeq \mu_F \mathbf{I}$ for some $\mu_F > 0$ and every $\theta \in \Theta$, WGD follows provided that the compatible function approximation error (Agarwal et al., 2021) is bounded. This condition requires $\mathbb{E}_{\pi_\theta}[(A^{\pi_\theta}(\mathbf{s}, \mathbf{a}) - (1 - \gamma)\mathbf{w}^\top \nabla_\theta \log \pi_\theta(\mathbf{a} | \mathbf{s}))^2] \leq \epsilon_{\text{bias}}$, where $\mathbf{w} = \mathbf{F}(\theta)^{-1} \nabla_\theta J(\theta)$ and A^{π_θ} is the advantage function of the policy π_θ . Under these conditions, WGD holds with $\chi = C_1/\mu_F$ and $\psi = \sqrt{\epsilon_{\text{bias}} \kappa}/(1 - \gamma)$. Here, $\kappa = \sup_{\theta \in \mathbb{R}^{d_\Theta}} \|\nu^{\pi_\theta^*}/\nu^{\pi_\theta}\|_\infty$

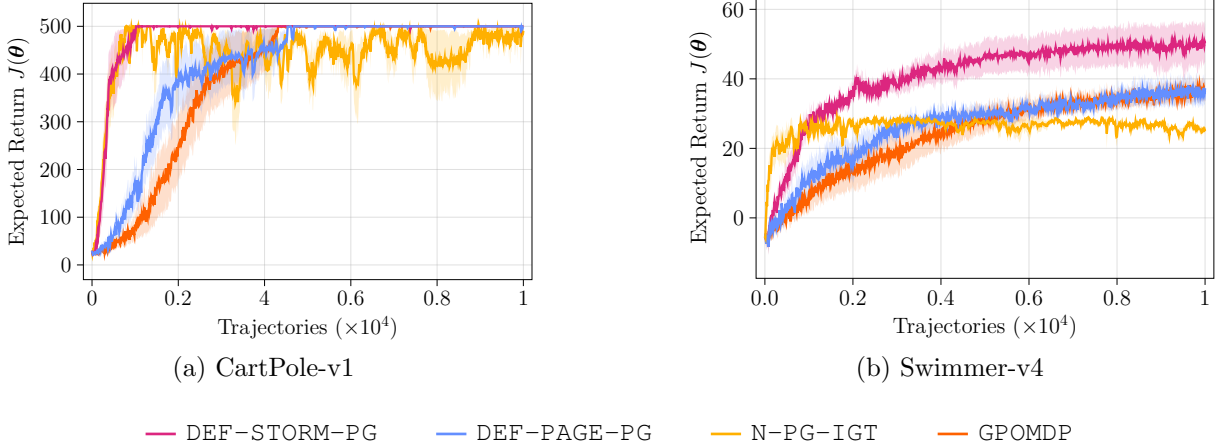


Figure 2: Comparison of DEF-STORM-PG, DEF-PAGE-PG, N-PG-IGT, and GPOMDP. Shaded regions indicate the estimated standard error of the mean.

is a concentrability coefficient, and ν^{π_θ} denotes the state-action distribution induced by π_θ . Note that, differently from previous works (Ding et al., 2022; Fatkhullin et al., 2023), we define the expected compatible function approximation error with respect to the current state-action distribution ν^{π_θ} rather than the optimal distribution ν^{π^*} . Bounding the error under ν^{π^*} is unverifiable because the optimal policy is unknown a priori. Conversely, evaluating this error under ν^{π_θ} requires the parametrization to be expressive along the realized optimization trajectory, a verifiable formulation that inherently incurs the statistical penalty κ within the irreducible bias.

Corollary 5.1 (Sample complexity under FND). *In the unconstrained setting where $\Theta = \mathbb{R}^{d_\Theta}$, letting $\chi = C_1/\mu_F$ and $\psi = \sqrt{\epsilon_{bias}\kappa}/(1-\gamma)$, assuming $\kappa < \infty$, **LVR-PG** achieves a suboptimality gap $J^* - \mathbb{E}[J(\theta_K)] \leq \frac{\sqrt{\epsilon_{bias}\kappa}}{1-\gamma} + \epsilon$ with a total sample complexity of:*

$$n = \tilde{\mathcal{O}}\left(\frac{R_{\max}^2 C_1^3}{(1-\gamma)^4 \mu_F^3 \epsilon^2}\right).$$

Corollary 5.1 confirms that when policies satisfy FND, the minimax optimal sample complexity established by Theorem 5.1 is robustly preserved without degradation. We refer the reader to Appendix C for a detailed discussion.

6 EXPERIMENTAL RESULTS

To evaluate the empirical sample efficiency of our framework, we consider CartPole-v1 and Swimmer-v4, covering classic control and MuJoCo locomotion tasks Todorov et al. (2012). The comparisons, hyperparameter ablations, and grid search reported in Appendix I motivate selecting the STORM gradient estimator as the primary representative of **LVR-PG**, given

its better sample efficiency. This configuration uses momentum-based recursive gradient estimates with defensive IS and no full-batch refreshes ($p = 0, \alpha > 0$). Figure 2 also includes DEF-PAGE-PG for completeness, showing its lower sample efficiency in the tested settings. Hybrid configurations are examined in Appendix I. We compare DEF-STORM-PG with GPOMDP, which uses fresh on-policy gradient estimates, and N-PG-IGT, which combines normalized updates with implicit gradient transport without importance sampling. These comparisons assess whether recursive variance reduction with defensive IS improves empirical sample efficiency relative to the two baselines. Across both environments, DEF-STORM-PG reaches high returns with fewer trajectories than GPOMDP and DEF-PAGE-PG. On CartPole-v1 DEF-STORM-PG exhibits fewer oscillations than N-PG-IGT, while in Swimmer-v4, the latter converges quickly to a local optimum.

7 CONCLUSIONS

We established that purely first-order PG methods can achieve minimax-optimal sample complexity up to logarithmic factors and horizon terms, with minimal assumptions and without second-order assumptions. Our findings suggest several directions for future research. First, the defensive IS mechanism requires evaluating trajectories under a mixture distribution, which inherently requires storing both the current and previous policy parameters. Developing variance-reduction mechanisms that avoid this explicit double sampling would significantly reduce memory overhead. Second, extending the **LVR-PG** theoretical framework to encompass adaptive gradient methods would bridge the gap between our convergence guarantees and the practical optimization dynamics prevalent in deep RL.

AI use statement

In this work, we used generative AI tools for text polishing, proof assistance, and code generation. We have not used generative AI tools to generate research ideas or to conduct the literature review, and generating figures or data synthesis is not applicable to this work. Additionally, we used generative AI tools to correct grammar and improve readability. We have reviewed all AI-assisted work. Specifically, the core ideas and initial proofs were developed manually by the authors, and AI was subsequently used to polish the presentation of the mathematical proofs. All AI-assisted proofs were checked line by line, and the authors thoroughly reviewed and tested any AI-assisted code for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- 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.
- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, 2024.
- Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.*, 91(3):325–349, 2013.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *journal of artificial intelligence research*, 15:319–350, 2001.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Leonardo Cesani, Matteo Papini, and Marcello Restelli. How log-barrier helps exploration in policy optimization, 2026. URL <https://arxiv.org/abs/2603.15001>.
- Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.
- Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1-2):1–142, 2013.
- Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient, 2022. URL <https://arxiv.org/abs/2110.10116>.
- Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global kurdyka-lojasiewicz inequality. *Advances in Neural Information Processing Systems*, 35:15836–15848, 2022.
- Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869. PMLR, 2023.
- 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 *International conference on machine learning*, pages 7223–7240. PMLR, 2022.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shiron Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.
- Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms, 2000. URL <http://papers.nips.cc/paper/1786-actor-critic-algorithms>.
- 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.

- Saeed Masiha, Saber Salehkaleybar, Niao He, Negar Kiyavash, and Patrick Thiran. Stochastic second-order methods improve best-known sample complexity of sgd for gradient-dominated functions. *Advances in Neural Information Processing Systems*, 35:10862–10875, 2022.
- Saeed Masiha, Negar Kiyavash, and Patrick Thiran. Sharp first-order lower bounds under sublevel $\setminus \alpha$ -polyak-lojasiewicz conditions. *arXiv preprint arXiv:2606.28278*, 2026.
- Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21(141):1–75, 2020.
- Washim U Mondal and Vaneet Aggarwal. Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3097–3105. PMLR, 2024.
- Alessandro Montenegro, Marco Mussi, Alberto Maria Metelli, and Matteo Papini. Learning optimal deterministic policies with stochastic policy gradients. *arXiv preprint arXiv:2405.02235*, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.
- Gabor Paczolay, Matteo Papini, Alberto Maria Metelli, Istvan Harmati, and Marcello Restelli. Sample complexity of variance-reduced policy gradient: weaker assumptions and lower bounds. *Machine Learning*, 113:1–36, 06 2024. doi: 10.1007/s10994-024-06573-4.
- Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirodda, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pages 4026–4035. Pmlr, 2018.
- Matteo Papini, Matteo Pirodda, and Marcello Restelli. Smoothing policies and safe policy gradients. *Machine Learning*, 111(11):4081–4137, 2022.
- Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
- Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.
- Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.
- Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.
- Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.
- 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, 2021.

Tingting Zhao, Hirotaka Hachiya, Gang Niu, and Masashi Sugiyama. Analysis and improvement of policy gradient estimation. *Advances in Neural Information Processing Systems*, 24, 2011.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

CHECKLIST

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, the paper describes the framework in Sections 2, 3 and 4.
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, the paper proves the complexity properties of a class of algorithms.
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, the anonymized code is provided in the supplemental material.
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. Yes, we define the assumptions in Section 3.
 - (b) Complete proofs of all theoretical results. Yes, the proofs of the theoretical results are provided in the appendix
 - (c) Clear explanations of any assumptions. Yes, the explanation of the assumptions is provided in both the main paper and the appendix.
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, the code to run the experiments is in the supplementary material.
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, the training details are provided in Appendix I
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, each plot includes a description of the specific measures included.
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, in Appendix I
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. Yes.
 - (b) The license information of the assets, if applicable. Not Applicable.
 - (c) New assets either in the supplemental material or as a URL, if applicable. Yes, the code can be found in the supplementary materials.
 - (d) Information about consent from data providers/curators. Not Applicable.
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. Not Applicable.
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Algorithm 2: Loopless Variance-Reduced Policy Gradient (**LVR-PG**)

Input: Iterations K ; batch size $N = N_0$; mini batch size B ; algorithmic parameters: $p \in [0, 1]$, $\alpha \in [0, 1]$, mixture coefficient $\xi \in [0, 1]$, IS operator ω , discount factor γ ; initial parameters $\theta_0 \in \Theta$; step size η ; constraint set Θ .

- 1 **for** $i = 1, \dots, N$ **do**
- 2 $\tau_i \leftarrow \text{CollectTrajectory}(\pi_{\theta_0}, \gamma)$;
- 3 $\hat{\nabla} J^\dagger(\theta_0) \leftarrow \frac{1}{N_0} \sum_{i=1}^{N_0} g(\tau_i \mid \theta_0)$;
- 4 **for** $k = 0, \dots, K - 1$ **do**
- 5 $\theta_{k+1} \leftarrow \mathcal{P}_\Theta \left(\theta_k + \eta \hat{\nabla}_\theta^\dagger J(\theta_k) \right)$;
- 6 Draw $c_k \sim \mathcal{B}(p)$;
- 7 **if** $c_k = 1$ **then**
- 8 **for** $i = 1, \dots, N$ **do**
- 9 $\tau_i \leftarrow \text{CollectTrajectory}(\pi_{\theta_{k+1}}, \gamma)$;
- 10 $\hat{\nabla}_\theta^\dagger J(\theta_{k+1}) \leftarrow \frac{1}{N} \sum_{i=1}^N g(\tau_i \mid \theta_{k+1})$;
- 11 **else**
- 12 **for** $i = 1, \dots, B$ **do**
- 13 Draw mixture variable $u \sim \mathcal{B}(\xi)$;
- 14 **if** $u = 1$ **then**
- 15 $\tau_i \leftarrow \text{CollectTrajectory}(\pi_{\theta_{k+1}}, \gamma)$;
- 16 **else**
- 17 $\tau_i \leftarrow \text{CollectTrajectory}(\pi_{\theta_k}, \gamma)$;
- 18 $\hat{\nabla}_\theta^\dagger J(\theta_{k+1}) \leftarrow \frac{1}{B} \sum_{i=1}^B \omega(\theta_{k+1}, \tau_i) g(\tau_i \mid \theta_{k+1}) + (1 - \alpha) \left[\hat{\nabla}_\theta^\dagger J(\theta_k) - \frac{1}{B} \sum_{i=1}^B \omega(\theta_k, \tau_i) g(\tau_i \mid \theta_k) \right]$;
- 19 **Return** θ_K

Algorithm 3: CollectTrajectory(π, γ)

Input: Policy π , discount factor $\gamma \in (0, 1)$.

- 1 Initialize trajectory $\tau \leftarrow \emptyset$, time step $h \leftarrow 0$;
- 2 Sample initial state $s_0 \sim \rho_0$;
- 3 **while** True **do**
- 4 Sample action $a_h \sim \pi(\cdot \mid s_h)$;
- 5 Observe reward r_{h+1} and next state $s_{h+1} \sim p(\cdot \mid s_h, a_h)$;
- 6 Append (s_h, a_h, r_{h+1}) to τ ;
- 7 Draw stopping condition $z \sim \mathcal{B}(1 - \gamma)$;
- 8 **if** $z = 1$ **then Return** $\tau \cup \{s_{h+1}\}$;
- 9 $h \leftarrow h + 1$;

A LOOPLESS VARIANCE-REDUCED POLICY GRADIENT (**LVR-PG**)

Algorithm 3 details the interaction protocol used to evaluate the policies. To ensure that the gradient estimates remain unbiased with respect to the infinite-horizon objective without introducing arbitrary truncation, the algorithm relies on a randomized geometric stopping rule. At each time step, the agent samples an action, observes the corresponding reward and state transition, and terminates the episode with probability $1 - \gamma$. This guarantees that the trajectory length follows a geometric distribution with expectation $(1 - \gamma)^{-1}$.

Algorithm 2 outlines the generic **LVR-PG** template, which tracks and updates a recursive gradient estimator $\hat{\nabla}_\theta^\dagger J(\theta_k)$ at each iteration k . The mechanism allowing for switching between full-batch and mini-batch updates relies on a Bernoulli random variable $c_k \sim \mathcal{B}(p)$ that dictates the estimator's update rule. When $c_k = 1$, the historical variance is discarded, and the estimator undergoes a full-batch refresh using N trajectories. Conversely,

when $c_k = 0$, the algorithm computes a variance-reduced update using a smaller mini-batch of B trajectories, modulated by a momentum decay parameter $\alpha \in [0, 1]$ and an importance sampling operator ω . By specifically choosing the hyperparameters (p, α, ξ) , we can instantiate multiple variance-reduced PG algorithms, as summarized in Table 2.

By selecting these parameters, the framework dynamically shifts its algorithmic behavior. PAGE-PG eliminates the recursive momentum term ($\alpha = 0$), relying entirely on the probabilistic refresh rate p to bound the accumulation of estimation error alongside standard importance sampling Gargiani et al. (2022). STORM-PG avoids full-batch refreshes entirely ($p = 0$) and instead uses a momentum parameter α to incrementally discount past gradient errors Yuan et al. (2020).

Crucially, the parameter $\xi \in (0, 1)$ activates the defensive importance sampling mechanism. As detailed in the recursive branch of Algorithm 2, drawing the Bernoulli mixture variable $u \sim \mathcal{B}(\xi)$ physically constructs the defensive trajectory distribution $q_\xi(\tau) = \xi p_{\theta_{k+1}}(\tau) + (1 - \xi)p_{\theta_k}(\tau)$. Under this regime, the IS operator $\omega(\theta, \tau)$ evaluates precisely to the density ratio $p_\theta(\tau)/q_\xi(\tau)$, intrinsically bounding the variance of the off-policy estimator without requiring arbitrary structural constraints on the policy class. While the generic LVR-PG template accommodates standard importance sampling ($\xi = 0$), throughout the theoretical analysis and empirical evaluations of this paper, we strictly consider the defensive setting ($\xi \in (0, 1)$).

We utilize the LVR-PG framework to unify the theoretical analysis of these diverse algorithms. By defining an effective decay rate $\lambda := 1 - (1 - p)(1 - \alpha)^2$, we encapsulate the error discounting effects of both probabilistic refreshes and continuous momentum into the exact same mathematical recurrence. This unified rate establishes the variance bound evaluated in Lemma D.4, allowing us to prove last-iterate global convergence in Theorem 5.1.

Table 2: Instantiations of the LVR-PG Framework

Algorithm	Refresh Prob. p	Momentum α	Defensive Parameter ξ
PAGE-PG Gargiani et al. (2022)	$p \in (0, 1)$	$\alpha = 0$	$\xi = 0$
STORM-PG Yuan et al. (2020)	$p = 0$	$\alpha \in (0, 1)$	$\xi = 0$
DEF-PAGE-PG Paczoly et al. (2024)	$p \in (0, 1)$	$\alpha = 0$	$\xi \in [0, 1]$
DEF-STORM-PG Paczoly et al. (2024)	$p = 0$	$\alpha \in (0, 1)$	$\xi \in [0, 1]$

B MEAN-SQUARED SMOOTHNESS

Because algorithms recursively update their gradient estimators using historical information, they must evaluate the difference between stochastic gradients at the current policy θ_k and a previous policy θ_{k-1} using a shared batch of trajectories. To correct for the shift in the trajectory distribution, these methods evaluate the gradients using an importance sampling (IS) operator ω . The mean-squared smoothness isolates the expected squared error of this gradient difference, taking the generic form:

$$\mathbb{E}_\tau \left[\|\omega(\theta_k, \tau)g(\tau | \theta_k) - \omega(\theta_{k-1}, \tau)g(\tau | \theta_{k-1})\|_2^2 \right] \leq C_{\dagger, \omega} \|\theta_k - \theta_{k-1}\|_2^2, \quad (3)$$

where $C_{\dagger, \omega} > 0$ is a unified bounding constant. In the following, we provide the mean-squared smoothness constants for different instantiations of Algorithm 2.

To avoid the unrealistic assumption of bounded variance for standard importance weights, which is frequently violated in continuous policy spaces, our framework relies entirely on defensive importance sampling (Owen and Zhou, 2000; Paczoly et al., 2024). Consequently, we evaluate the mean-squared smoothness strictly under the defensive mixture distribution q_ξ , which intrinsically bounds the variance without requiring arbitrary structural constraints.

Lemma B.1 (Mean-Squared Smoothness for LVR-PG). *For the LVR-PG framework, sampling from a defensive mixture distribution $q_\xi(\tau) = \xi p_{\theta_k}(\tau) + (1 - \xi)p_{\theta_{k-1}}(\tau)$ with parameter $\xi \in (0, 1)$, the unified mean-squared smoothness constant $C_{DEF, \omega}$ is strictly bounded. Letting $m_\xi := \min\{\xi, 1 - \xi\}$ and $M_r := \mathbb{E}[H_\tau^r]$ denote the moments of the geometrically distributed trajectory length, under Assumption 3.1 and bounded rewards, it holds*

that:

$$\mathbb{E}_{\tau \sim q_\xi} \left[\left\| \omega(\boldsymbol{\theta}_k, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_k) - \omega(\boldsymbol{\theta}_{k-1}, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_{k-1}) \right\|_2^2 \right] \leq C_{\dagger, \omega} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|_2^2,$$

where the unified mean-squared smoothness constant is:

$$C_{\dagger, \omega} = \frac{R_{\max}^2}{m_\xi} (C_1^4 M_5 + 2L_S^2 M_4).$$

Setting $\xi = 1/2$ minimizes the variance bound, yielding $m_{1/2} = 1/2$ and the optimized mean-squared smoothness constant for both algorithms:

$$C_{\dagger, \omega} = 2R_{\max}^2 (C_1^4 M_5 + 2L_S^2 M_4) = \mathcal{O}((1 - \gamma)^{-5}).$$

Proof. We recall a result by Paczolat et al. (2024), proved here for completeness. To establish the mean-squared smoothness bound for both algorithms, we must bound the expected squared L_2 norm of the difference between the defensive gradient estimators evaluated at consecutive iterates $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k-1}$. For brevity, we drop the trajectory dependency and denote $p_{\boldsymbol{\theta}_k} = p_{\boldsymbol{\theta}_k}(\tau)$, $\mathbf{g}_k = \mathbf{g}(\tau \mid \boldsymbol{\theta}_k)$, and $\omega_{\boldsymbol{\theta}_k} = \frac{p_{\boldsymbol{\theta}_k}}{q_\xi}$. We decouple the gradient difference algebraically by adding and subtracting $\omega_{\boldsymbol{\theta}_{k-1}} \mathbf{g}_k$:

$$\omega_{\boldsymbol{\theta}_k} \mathbf{g}_k - \omega_{\boldsymbol{\theta}_{k-1}} \mathbf{g}_{k-1} = (\omega_{\boldsymbol{\theta}_k} - \omega_{\boldsymbol{\theta}_{k-1}}) \mathbf{g}_k + \omega_{\boldsymbol{\theta}_{k-1}} (\mathbf{g}_k - \mathbf{g}_{k-1}).$$

Applying the inequality $\|x + y\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2$, we can bound the expectation by isolating the importance weights from the on-policy gradients:

$$\mathbb{E}_{\tau \sim q_\xi} \left[\left\| \omega_{\boldsymbol{\theta}_k} \mathbf{g}_k - \omega_{\boldsymbol{\theta}_{k-1}} \mathbf{g}_{k-1} \right\|_2^2 \right] \leq 2\mathbb{E}_{\tau \sim q_\xi} \left[\left\| (\omega_{\boldsymbol{\theta}_k} - \omega_{\boldsymbol{\theta}_{k-1}}) \mathbf{g}_k \right\|_2^2 \right] + 2\mathbb{E}_{\tau \sim q_\xi} \left[\left\| \omega_{\boldsymbol{\theta}_{k-1}} (\mathbf{g}_k - \mathbf{g}_{k-1}) \right\|_2^2 \right].$$

We evaluate the two terms independently. For the first term, we substitute $\omega_{\boldsymbol{\theta}_k} - \omega_{\boldsymbol{\theta}_{k-1}} = \frac{p_{\boldsymbol{\theta}_k} - p_{\boldsymbol{\theta}_{k-1}}}{q_\xi}$ and use the definition of the defensive mixture $q_\xi(\tau) \geq m_\xi(p_{\boldsymbol{\theta}_k}(\tau) + p_{\boldsymbol{\theta}_{k-1}}(\tau))$ to bound the density ratio:

$$\frac{(p_{\boldsymbol{\theta}_k} - p_{\boldsymbol{\theta}_{k-1}})^2}{q_\xi} \leq \frac{1}{m_\xi} \frac{(p_{\boldsymbol{\theta}_k} - p_{\boldsymbol{\theta}_{k-1}})^2}{p_{\boldsymbol{\theta}_k} + p_{\boldsymbol{\theta}_{k-1}}} = \frac{1}{m_\xi} (\sqrt{p_{\boldsymbol{\theta}_k}} - \sqrt{p_{\boldsymbol{\theta}_{k-1}}})^2 \frac{(\sqrt{p_{\boldsymbol{\theta}_k}} + \sqrt{p_{\boldsymbol{\theta}_{k-1}}})^2}{p_{\boldsymbol{\theta}_k} + p_{\boldsymbol{\theta}_{k-1}}} \leq \frac{2}{m_\xi} (\sqrt{p_{\boldsymbol{\theta}_k}} - \sqrt{p_{\boldsymbol{\theta}_{k-1}}})^2.$$

By bounded rewards and Assumption 3.1, the pathwise GPOMDP gradient is bounded by $\|\mathbf{g}_k\|_2 \leq R_{\max} C_1 H_\gamma^2$. Expanding the expectation of the first term yields:

$$\begin{aligned} \mathbb{E}_{\tau \sim q_\xi} \left[\left\| (\omega_{\boldsymbol{\theta}_k} - \omega_{\boldsymbol{\theta}_{k-1}}) \mathbf{g}_k \right\|_2^2 \right] &= \int \frac{(p_{\boldsymbol{\theta}_k} - p_{\boldsymbol{\theta}_{k-1}})^2}{q_\xi} \|\mathbf{g}_k\|_2^2 d\tau \\ &\leq \frac{2R_{\max}^2 C_1^2}{m_\xi} \int H_\gamma^4 (\sqrt{p_{\boldsymbol{\theta}_k}} - \sqrt{p_{\boldsymbol{\theta}_{k-1}}})^2 d\tau. \end{aligned}$$

Applying Lemma D.3 with $r = 4$ we obtain:

$$\mathbb{E}_{\tau \sim q_\xi} \left[\left\| (\omega_{\boldsymbol{\theta}_k} - \omega_{\boldsymbol{\theta}_{k-1}}) \mathbf{g}_k \right\|_2^2 \right] \leq \frac{R_{\max}^2 C_1^4 M_5}{2m_\xi} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|_2^2.$$

For the second term, we use the property $q_\xi \geq (1 - \xi)p_{\boldsymbol{\theta}_{k-1}} \geq m_\xi p_{\boldsymbol{\theta}_{k-1}}$ to convert the expectation under the defensive mixture back to the behavior policy distribution:

$$\begin{aligned} \mathbb{E}_{\tau \sim q_\xi} \left[\left\| \omega_{\boldsymbol{\theta}_{k-1}} (\mathbf{g}_k - \mathbf{g}_{k-1}) \right\|_2^2 \right] &= \int \frac{p_{\boldsymbol{\theta}_{k-1}}^2}{q_\xi} \|\mathbf{g}_k - \mathbf{g}_{k-1}\|_2^2 d\tau \\ &\leq \frac{1}{m_\xi} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}_{k-1}}} \left[\|\mathbf{g}_k - \mathbf{g}_{k-1}\|_2^2 \right]. \end{aligned}$$

Applying the mean-square continuity of the policy scores (Assumption 3.1) and the Cauchy-Schwarz inequality over the sequence of rewards, the expected squared difference of the gradients is bounded by:

$$\mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}_{k-1}}} \left[\|\mathbf{g}_k - \mathbf{g}_{k-1}\|_2^2 \right] \leq R_{\max}^2 L_S^2 M_4 \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\|_2^2.$$

Substituting this back bounds the second term by:

$$\mathbb{E}_{\tau \sim q_\xi} [\|\omega_{\theta_{k-1}}(\mathbf{g}_k - \mathbf{g}_{k-1})\|_2^2] \leq \frac{R_{\max}^2 L_S^2 M_4}{m_\xi} \|\theta_k - \theta_{k-1}\|_2^2.$$

Finally, summing the two bounds with their respective factor of 2 recovers the exact unified estimation constant:

$$\begin{aligned} \mathbb{E}_{\tau \sim q_\xi} \left[\|\omega_{\theta_k} \mathbf{g}_k - \omega_{\theta_{k-1}} \mathbf{g}_{k-1}\|_2^2 \right] &\leq 2 \left(\frac{R_{\max}^2 C_1^4 M_5}{2m_\xi} \right) \|\theta_k - \theta_{k-1}\|_2^2 + 2 \left(\frac{R_{\max}^2 L_S^2 M_4}{m_\xi} \right) \|\theta_k - \theta_{k-1}\|_2^2 \\ &= \frac{R_{\max}^2}{m_\xi} (C_1^4 M_5 + 2L_S^2 M_4) \|\theta_k - \theta_{k-1}\|_2^2. \end{aligned}$$

□

DEF-PAGE-PG, DEF-STORM-PG ($C_{\text{DEF}, \omega}$). Although DEF-PAGE-PG and DEF-STORM-PG employ different outer recursive mechanisms, their consecutive estimation bounds are identical. This equivalence arises because both algorithms utilize the exact same GPOMDP gradient estimator and rely on the defensive mixture distribution $q_\xi(\tau) = \xi p_{\theta_k}(\tau) + (1 - \xi) p_{\theta_{k-1}}(\tau)$ to control the variance of the importance weights. Since the mean squared smoothness isolates only the expected squared error between gradient estimates at consecutive policies, it is strictly independent of the outer update rule. We can verify this by algebraically splitting the update in the recursive branch of Algorithm 2:

$$\begin{aligned} \hat{\nabla}_{\theta}^{\dagger} J(\theta_{k+1}) &= (1 - \alpha) \frac{1}{B} \sum_{i=1}^B \omega(\theta_{k+1}, \tau_i) \mathbf{g}(\tau_i | \theta_{k+1}) + \alpha \frac{1}{B} \sum_{i=1}^B \omega(\theta_{k+1}, \tau_i) \mathbf{g}(\tau_i | \theta_k) \\ &\quad + (1 - \alpha) \left[\hat{\nabla}_{\theta}^{\dagger} J(\theta_k) - \frac{1}{B} \sum_{i=1}^B \omega(\theta_k, \tau_i) \mathbf{g}(\tau_i | \theta_k) \right]. \end{aligned}$$

Substituting this back into the update rule allows us to group the gradient difference directly:

$$\begin{aligned} \hat{\nabla}_{\theta}^{\dagger} J(\theta_{k+1}) &= (1 - \alpha) \hat{\nabla}_{\theta}^{\dagger} J(\theta_k) + (1 - \alpha) \left[\frac{1}{B} \sum_{i=1}^B (\omega(\theta_{k+1}, \tau_i) \mathbf{g}(\tau_i | \theta_{k+1}) - \omega(\theta_k, \tau_i) \mathbf{g}(\tau_i | \theta_k)) \right] \\ &\quad + \alpha \frac{1}{B} \sum_{i=1}^B \omega(\theta_{k+1}, \tau_i) \mathbf{g}(\tau_i | \theta_{k+1}). \end{aligned}$$

Hence, adding the continuous momentum does not alter the underlying variance of the defensive gradient differences. By exploiting the mean-square continuity of the scores (Assumption 3.1), defensive importance sampling strictly bounds the mean-squared smoothness without relying on unbounded uniform assumptions or arbitrary Hessian limits.

C ON FISHER-NON-DEGENERACY

Assumption C.1 (Fisher-non-degeneracy). *There exists some constant $\mu_F > 0$, such that for every $\theta \in \mathbb{R}^d$, the FIM $F(\theta)$ induced by the policy π_{θ} satisfies:*

$$F(\theta) = \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\theta}}} [\nabla_{\theta} \log \pi_{\theta}(\mathbf{a} | \mathbf{s}) \nabla_{\theta} \log \pi_{\theta}(\mathbf{a} | \mathbf{s})^{\top}] \succeq \mu_F \cdot \mathbf{I}. \quad (4)$$

Assumption C.1 requires the FIM to be uniformly lower-bounded by a strictly positive constant $\mu_F > 0$. This means that the FIM acts as a well-behaved preconditioner in the natural PG update (Kakade, 2001; Ding et al., 2022). Geometrically, the FIM characterizes the local curvature of the policy manifold measured via the Kullback-Leibler (KL) divergence. This condition mathematically implies that the parameterization contains no redundant or "flat" directions: any non-zero update in the parameter space θ must induce a strictly positive change in the statistical behavior of the policy.

The validity of the FND assumption fundamentally depends on the chosen policy architecture and the underlying feature representation.

Gaussian Policies. In continuous action spaces, consider a Gaussian policy with a linearly parameterized mean and a fixed covariance matrix, $\pi_{\theta}(a | s) = \mathcal{N}(\phi(s)^\top \theta, \Sigma)$. The FIM is proportional to the uncentered covariance of the state features: $F(\theta) = \mathbb{E}_{\nu^{\pi_{\theta}}} [\phi(s) \Sigma^{-1} \phi(s)^\top]$. Here, FND holds globally if and only if the feature covariance matrix is strictly positive definite. This requires the basis functions $\phi(s)$ to be linearly independent and the state distribution to provide sufficiently rich coverage. When the mean of the gaussian policy $\mu_{\theta}(s)$ is nonlinear in θ , as in the case of (deep) neural networks, Assumption C.1 can be satisfied under assumptions on the Jacobian $\nabla_{\theta} \mu_{\theta}(s)$ similar to the ones for the linear case (Fatkhullin et al., 2023).

Softmax Policies. For discrete action spaces using a softmax parameterization, the elements of the FIM depend directly on the policy’s action probabilities. If the policy approaches determinism (i.e., $\pi_{\theta}(a | s) \rightarrow 0$ for suboptimal actions), the corresponding entries in the FIM vanish, driving the minimum eigenvalue μ_F toward zero (Ding et al., 2022). In this case, strict global FND cannot hold without specific architectural constraints or regularization (Cesani et al., 2026).

The FND setting contains more than (linear) Gaussians: in addition, Assumption C.1 holds for every full-rank exponential parametrization with their mean parametrized with $\mu_{\theta}(s)$ if $\mu_{\theta}(s)$ is full rank for all $\theta \in \Theta$. We refer to Section 8 of Ding et al. (2022) and Appendix B of Fatkhullin et al. (2023) for a detailed discussion of FND.

Assumption C.2 (Compatible Function Approximation Error). *There exist a constant $\epsilon_{bias} > 0$, such that for every $\theta \in \mathbb{R}^d$, the transferred compatible function approximation error satisfies*

$$\mathbb{E}_{\nu^{\pi_{\theta}}} \left[\left(A^{\pi_{\theta}}(s, a) - (1 - \gamma) \mathbf{u}^*{}^\top \nabla_{\theta} \log \pi_{\theta}(a | s) \right)^2 \right] \leq \epsilon_{bias}, \quad (5)$$

where $\nu^{\pi_{\theta}}$ is the state-action distribution induced by a policy π_{θ} and $\mathbf{u}^* := F(\theta)^{-1} \nabla_{\theta} J(\theta)$ is the solution of

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathbb{E}_{\nu^{\pi_{\theta}}} \left[\left(A^{\pi_{\theta}}(s, a) - (1 - \gamma) \mathbf{w}^\top \nabla_{\theta} \log \pi_{\theta}(a | s) \right)^2 \right].$$

Assumption C.2 bounds the error incurred when approximating the advantage function $A^{\pi_{\theta}}(s, a)$ using the score function $\nabla_{\theta} \log \pi_{\theta}(a | s)$ as a linear feature representation. The vector $\mathbf{u}^* := F(\theta)^{-1} \nabla_{\theta} J(\theta)$ is exactly the minimizer of this weighted least-squares regression problem. As established by (Agarwal et al., 2021), this assumption characterizes the expressivity of the parameterized policy class. If the policy class is sufficiently rich, e.g., highly parameterized neural networks or softmax tabular policies, this transferred approximation error ϵ_{bias} is small.

A critical theoretical distinction must be done between Assumption C.2 and analogous assumptions utilized in the recent variance-reduced policy optimization literature. For instance, the analysis of STORM-PG-S by (Ding et al., 2022) requires the compatible function approximation error to be bounded under the state-action distribution of the *optimal* policy, ν^{π^*} . In contrast, our formulation strictly evaluates the expectation under the state-action distribution of the *current* policy, $\nu^{\pi_{\theta}}$.

This difference is motivated by the fact that assuming a bounded approximation error under ν^{π^*} means requiring that the agent has information about the optimal policy, which is not possible. This is because it is not possible to guarantee that the current parametrization accurately fits the advantage function over regions of the state-action space visited by an optimal policy π^* , which is entirely unknown. By defining the expectation over $\nu^{\pi_{\theta}}$, Assumption C.2 is entirely related to the policy that is being optimized, from which we can gather information about its state-action distribution.

Lemma C.1 (Fisher-induced Gradient Domination). *Let us consider a general Fisher-non-degenerate policy π_{θ} satisfying Assumptions C.1, 3.1, and C.2 in the unconstrained parameter space $\Theta = \mathbb{R}^{d_{\theta}}$. Assume additionally that $\sup_{\theta \in \Theta} \|\nu^{\pi_{\theta}^*} / \nu_{\theta}^{\pi}\|_{\infty} < \infty$, where the ratio denotes the Radon–Nikodym derivative. Then, we have:*

$$J(\theta^*) - J(\theta) \leq \frac{\sqrt{\epsilon_{bias}}}{1 - \gamma} \left\| \frac{\nu^{\pi_{\theta}^*}}{\nu_{\theta}^{\pi}} \right\|_{\infty} + \frac{C_1}{\mu_F} \|\nabla_{\theta} J(\theta)\|_2.$$

Proof. The proof follows the one provided by Ding et al. (2022), with the expectation of Assumption C.2 defined w.r.t. the current policy π_{θ} instead of the optimal policy π^* . We begin with the performance-difference lemma

and perform a change of distribution:

$$\begin{aligned} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}) &= \frac{1}{1-\gamma} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi^*}} [A^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}}}} \left[A^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}(\mathbf{s}, \mathbf{a})}{\nu^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})} \right]. \end{aligned}$$

By introducing the optimal compatible weights $u^* := F(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ and adding and subtracting $(1-\gamma)u^{*\top} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s})$, we obtain:

$$\begin{aligned} J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}) &= \frac{1}{1-\gamma} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}}}} \left[\left(A^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) - (1-\gamma)u^{*\top} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s}) \right) \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}(\mathbf{s}, \mathbf{a})}{\nu^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})} \right] \\ &\quad + \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}^*}}} [u^{*\top} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s})]. \end{aligned}$$

To bound the right-hand side, we evaluate the two terms independently. Applying Hölder's inequality to the first term, followed by the Cauchy–Schwarz inequality, gives:

$$\begin{aligned} &\left| \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}}}} \left[\left(A^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) - (1-\gamma)u^{*\top} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s}) \right) \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}(\mathbf{s}, \mathbf{a})}{\nu^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})} \right] \right| \\ &\leq \left\| \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}}{\nu^{\pi_{\boldsymbol{\theta}}}} \right\|_{\infty} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}}}} [|A^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) - (1-\gamma)u^{*\top} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s})|] \\ &\leq \left\| \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}}{\nu^{\pi_{\boldsymbol{\theta}}}} \right\|_{\infty} \sqrt{\mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}}}} [(A^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) - (1-\gamma)u^{*\top} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s}))^2]} \\ &\leq \sqrt{\epsilon_{\text{bias}}} \left\| \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}}{\nu^{\pi_{\boldsymbol{\theta}}}} \right\|_{\infty}, \end{aligned}$$

where the final inequality follows from Assumption C.2. For the second term, the bounded score condition in Assumption 3.1 and Fisher-non-degeneracy give:

$$\begin{aligned} \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}^*}}} [u^{*\top} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s})] &\leq \|u^*\|_2 \mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \nu^{\pi_{\boldsymbol{\theta}^*}}} [\|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a} | \mathbf{s})\|_2] \\ &\leq C_1 \|u^*\|_2 \\ &= C_1 \|F(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\|_2 \\ &\leq \frac{C_1}{\mu_F} \|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\|_2. \end{aligned}$$

Combining the two bounds yields:

$$J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}) \leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} \left\| \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}}{\nu^{\pi_{\boldsymbol{\theta}}}} \right\|_{\infty} + \frac{C_1}{\mu_F} \|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\|_2.$$

□

As discussed in Agarwal et al. (2021), analyzing PG methods requires carefully analyzing distinct sources of error: the statistical regression error and the mismatch in state-action visitations. By adopting the formulation in Assumption C.2, the representational capacity of the model is captured entirely by ϵ_{bias} under the data-generating distribution $\nu^{\pi_{\boldsymbol{\theta}}}$. Consequently, the structural penalty for not visiting optimal states is explicitly isolated and paid later in the Fisher-induced Gradient Domination Lemma, manifesting as the concentrability coefficient $\kappa = \left\| \frac{\nu^{\pi_{\boldsymbol{\theta}^*}}}{\nu^{\pi_{\boldsymbol{\theta}}}} \right\|_{\infty}$.

D INTERMEDIATE RESULTS

Here, the lemmas used in the proofs of the main results are presented.

Lemma D.1 (Variance of the Geometrically-Stopped GPOMDP Estimator). *Under Assumption 3.1(i), the variance of the on-policy gradient estimator is strictly bounded. For all $\boldsymbol{\theta} \in \Theta$, it holds that:*

$$\text{Var}(g(\tau | \boldsymbol{\theta})) \leq \frac{4R_{\max}^2 C_1^2}{(1-\gamma)^3} =: \nu^2.$$

Proof. Let the gradient estimator for a single trajectory τ of length H_γ be defined as $g(\tau \mid \boldsymbol{\theta}) = \sum_{t=0}^{H_\gamma-1} r_{t+1} \sum_{h=0}^t \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_h \mid \mathbf{s}_h)$. The variance is upper-bounded by the uncentered second moment: $\mathbb{V}\text{ar}(g(\tau \mid \boldsymbol{\theta})) \leq \mathbb{E}_\tau [\|g(\tau \mid \boldsymbol{\theta})\|_2^2]$.

First, we condition on a H_γ . Using the Cauchy-Schwarz inequality, we decouple the rewards from the score functions:

$$\begin{aligned} \|g(\tau \mid \boldsymbol{\theta})\|_2^2 &= \left\| \sum_{t=0}^{H_\gamma-1} r_{t+1} \sum_{h=0}^t \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_h \mid \mathbf{s}_h) \right\|_2^2 \\ &\leq \left(\sum_{t=0}^{H_\gamma-1} r_{t+1}^2 \right) \left(\sum_{t=0}^{H_\gamma-1} \left\| \sum_{h=0}^t \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_h \mid \mathbf{s}_h) \right\|_2^2 \right). \end{aligned} \quad (6)$$

Because the rewards are uniformly bounded, the first term is bounded by $HR_{\max}^2$. For the second term, we take the expectation over the trajectory conditioned on H . By the martingale difference property of the score function, the cross-terms of the gradients vanish in expectation:

$$\begin{aligned} \mathbb{E}_\tau \left[\left\| \sum_{h=0}^t \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_h \mid \mathbf{s}_h) \right\|_2^2 \mid H_\gamma \right] &= \sum_{h=0}^t \mathbb{E}_\tau \left[\left\| \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_h \mid \mathbf{s}_h) \right\|_2^2 \mid H_\gamma \right] \\ &\leq (t+1)C_1^2, \end{aligned} \quad (7)$$

where the inequality follows directly from Assumption 3.1(i). Substituting Equation (7) back into Equation (6) yields:

$$\mathbb{E}_\tau [\|g(\tau \mid \boldsymbol{\theta})\|_2^2 \mid H_\gamma] \leq H_\gamma R_{\max}^2 \sum_{t=0}^{H_\gamma-1} (t+1)C_1^2 = R_{\max}^2 C_1^2 \frac{H_\gamma^2(H_\gamma+1)}{2}. \quad (8)$$

Finally, we take the expectation over the geometric stopping time H_γ , which follows a geometric distribution with success probability $p = 1 - \gamma$. Using the standard bounds for the moments of a geometric distribution, we have $\mathbb{E}[H_\gamma^3] \leq 6/p^3$ and $\mathbb{E}[H_\gamma^2] \leq 2/p^2$. Thus:

$$\begin{aligned} \mathbb{E}_{H_\gamma} \left[R_{\max}^2 C_1^2 \frac{H_\gamma^3 + H_\gamma^2}{2} \right] &\leq \frac{R_{\max}^2 C_1^2}{2} \left(\frac{6}{(1-\gamma)^3} + \frac{2}{(1-\gamma)^2} \right) \\ &\leq \frac{4R_{\max}^2 C_1^2}{(1-\gamma)^3}, \end{aligned}$$

where the final inequality assumes γ is sufficiently close to 1, yielding the standard $\mathcal{O}((1-\gamma)^{-3})$ variance bound for the GPOMDP estimator. $\square$

Lemma D.2 (Smoothness of the expected return). *Let $|R(\mathbf{s}, \mathbf{a})| \leq R_{\max}$ for all $\mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$ and Θ convex. Under Assumption 3.1, we have: Then, for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,*

$$\|\nabla J(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta}')\| \leq \frac{R_{\max}(L_S + C_1^2)}{(1-\gamma)^2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

Proof. For $u \in [0, 1]$, let $\boldsymbol{\theta}_u = u\boldsymbol{\theta} + (1-u)\boldsymbol{\theta}'$, which belongs to Θ by convexity. For $h \geq 0$, let $\mathbf{Z}_h(u) = \sum_{\ell=0}^h \nabla \log \pi_{\boldsymbol{\theta}_u}(\mathbf{a}_\ell \mid \mathbf{s}_\ell)$. Let $\tau_{0:h}$ denote the trajectory prefix up to timestep h included, and $p_{\boldsymbol{\theta}}^h$ its law under policy $\pi_{\boldsymbol{\theta}}$. Let:

$$\mathbf{G}_h(u) = \mathbb{E}_{\tau_{0:h} \sim p_{\boldsymbol{\theta}_u}^h} [r_{h+1} \mathbf{Z}_h(u)],$$

so that $\nabla J(\boldsymbol{\theta}_u) = \sum_{h=0}^{\infty} \gamma^h \mathbf{G}_h(u)$. Let $\epsilon > 0$ be such that $u + \epsilon \in [0, 1]$. By adding and subtracting $\mathbb{E}_{\tau_{0:h} \sim p_{\boldsymbol{\theta}_{u+\epsilon}}^h} [r_{h+1} \mathbf{Z}_h(u)]$:

$$\mathbf{G}_h(u + \epsilon) - \mathbf{G}_h(u) = \underbrace{\mathbb{E}_{\tau_{0:h} \sim p_{\boldsymbol{\theta}_{u+\epsilon}}^h} [r_{h+1}(\mathbf{Z}_h(u + \epsilon) - \mathbf{Z}_h(u))]}_{(A)} + \underbrace{\int (p_{\boldsymbol{\theta}_{u+\epsilon}}^h(\tau_{0:h}) - p_{\boldsymbol{\theta}_u}^h(\tau_{0:h})) r_{h+1} \mathbf{Z}_h(u) d\tau_{0:h}}_{(B)}. \quad (9)$$

We bound the first term using Jensen's inequality and the second assumption on the policy scores:

$$\begin{aligned}
 \| (A) \| &\leq R_{\max} \sum_{\ell=0}^h \mathbb{E}_{\tau_{0:\ell} \sim p_{\theta_{u+\epsilon}}^\ell} [\| \nabla \log \pi_{\theta_{u+\epsilon}}(\mathbf{a}_\ell | \mathbf{s}_\ell) - \nabla \log \pi_{\theta_u}(\mathbf{a}_\ell | \mathbf{s}_\ell) \|] \\
 &= R_{\max} \sum_{\ell=0}^h \mathbb{E}_{\tau_{0:\ell} \sim p_{\theta_{u+\epsilon}}^\ell} \left[\mathbb{E}_{\mathbf{a} \sim \pi_{\theta_{u+\epsilon}}(\cdot | \mathbf{s}_\ell)} [\| \nabla \log \pi_{\theta_{u+\epsilon}}(\mathbf{a}_\ell | \mathbf{s}_\ell) - \nabla \log \pi_{\theta_u}(\mathbf{a}_\ell | \mathbf{s}_\ell) \|] \right] \\
 &\leq R_{\max}(h+1)L_S \epsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|,
 \end{aligned}$$

$$\limsup_{\epsilon \downarrow 0} \frac{\|(A)\|}{\epsilon} \leq R_{\max}(h+1)L_S \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

As for the second term, note that, for $\ell < k \leq h$,

$$\begin{aligned}
 &\mathbb{E}_{\tau_{0:k} \sim p_{\theta_u}^k} [\nabla \log \pi_{\theta_u}(\mathbf{a}_\ell | \mathbf{s}_\ell)^\top \nabla \log \pi_{\theta_u}(\mathbf{a}_k | \mathbf{s}_k)] \\
 &= \mathbb{E}_{\tau_{0:k} \sim p_{\theta_u}^k} [\nabla \log \pi_{\theta_u}(\mathbf{a}_\ell | \mathbf{s}_\ell)^\top \mathbb{E}_{\mathbf{a} \sim \pi_{\theta_u}(\cdot | \mathbf{s}_k)} [\nabla \log \pi_{\theta_u}(\mathbf{a} | \mathbf{s}_k)]] = 0,
 \end{aligned}$$

so, by the first assumption on the policy scores:

$$\mathbb{E}_{\tau_{0:h} \sim p_{\theta_u}^h} [\|\mathbf{Z}_h(u)\|^2] = \sum_{\ell=0}^h \mathbb{E}_{\tau_{0:\ell} \sim p_{\theta_u}^\ell} [\|\nabla \log \pi_{\theta_u}(\mathbf{a}_\ell | \mathbf{s}_\ell)\|^2] \leq (h+1)C_1^2.$$

Then:

$$\begin{aligned}
 \lim_{\epsilon \downarrow 0} \frac{(B)}{\epsilon} &= \int \lim_{\epsilon \downarrow 0} \frac{p_{\theta_{u+\epsilon}}^h(\tau_{0:h}) - p_{\theta_u}^h(\tau_{0:h})}{\epsilon} r_{h+1} \mathbf{Z}_h(u) d\tau_{0:h} \\
 &= \int \nabla p_{\theta_u}^h(\tau_{0:h})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}') r_{h+1} \mathbf{Z}_h(u) d\tau_{0:h} \\
 &= \mathbb{E}_{\tau_{0:h} \sim p_{\theta_u}^h} [\mathbf{Z}_h(u) \mathbf{Z}_h(u)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}') r_{h+1}],
 \end{aligned}$$

and

$$\begin{aligned}
 \left\| \lim_{\epsilon \downarrow 0} \frac{(B)}{\epsilon} \right\| &\leq R_{\max} \mathbb{E}_{\tau_{0:h} \sim p_{\theta_u}^h} [\|\mathbf{Z}_h(u)\|^2] \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \\
 &\leq R_{\max}(h+1)C_1^2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.
 \end{aligned}$$

Now, from Equation (9):

$$\begin{aligned}
 \limsup_{\epsilon \downarrow 0} \frac{\|\mathbf{G}_h(u+\epsilon) - \mathbf{G}_h(u)\|}{\epsilon} &\leq \limsup_{\epsilon \downarrow 0} \frac{\|(A)\|}{\epsilon} + \limsup_{\epsilon \downarrow 0} \frac{\|(B)\|}{\epsilon} \\
 &= \limsup_{\epsilon \downarrow 0} \frac{\|(A)\|}{\epsilon} + \left\| \lim_{\epsilon \downarrow 0} \frac{(B)}{\epsilon} \right\| \\
 &\leq R_{\max}(h+1)(L_S + C_1^2) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.
 \end{aligned}$$

The same decomposition shows that $\mathbf{G}_h$ is continuous on $[0, 1]$. By Lemma H.1:

$$\|\mathbf{G}_h(1) - \mathbf{G}_h(0)\| \leq R_{\max}(h+1)(L_S + C_1^2) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

Finally:

$$\begin{aligned}
 \|\nabla J(\boldsymbol{\theta}') - \nabla J(\boldsymbol{\theta})\| &\leq \sum_{h=0}^{\infty} \gamma^h \|\mathbf{G}_h(1) - \mathbf{G}_h(0)\| \\
 &\leq R_{\max}(L_S + C_1^2) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \sum_{h=0}^{\infty} \gamma^h (h+1) \\
 &= \frac{R_{\max}(L_S + C_1^2)}{(1-\gamma)^2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.
 \end{aligned}$$

□

Lemma D.3 (Weighted Hellinger continuity). *Suppose that Assumption 3.1(1) holds. For every integer $r \geq 0$ and every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,*

$$\int_{\mathcal{T}} H_{\gamma}(\tau)^r \left(\sqrt{p(\tau \mid \boldsymbol{\theta}_1)} - \sqrt{p(\tau \mid \boldsymbol{\theta}_2)} \right)^2 d\tau \leq \frac{C_1^2 M_{r+1}}{4} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2.$$

Proof. We recall a result by Paczolat et al. (2024), proved here for completeness Let

$$\boldsymbol{\theta}_t = (1-t)\boldsymbol{\theta}_2 + t\boldsymbol{\theta}_1, \quad \boldsymbol{\Delta} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \quad t \in [0, 1].$$

Convexity of Θ guarantees that $\boldsymbol{\theta}_t \in \Theta$. Because the stopping rule and the environment are independent of the policy parameters,

$$\nabla \log p(\tau \mid \boldsymbol{\theta}_t) = \sum_{h=0}^{H_{\gamma}(\tau)-1} \nabla \log \pi_{\boldsymbol{\theta}_t}(\mathbf{a}_h \mid \mathbf{s}_h).$$

Consequently,

$$\frac{d}{dt} \sqrt{p(\tau \mid \boldsymbol{\theta}_t)} = \frac{1}{2} \sqrt{p(\tau \mid \boldsymbol{\theta}_t)} \langle \boldsymbol{\Delta}, \nabla \log p(\tau \mid \boldsymbol{\theta}_t) \rangle.$$

By the fundamental theorem of calculus and Cauchy–Schwarz,

$$\begin{aligned} & \left(\sqrt{p(\tau \mid \boldsymbol{\theta}_1)} - \sqrt{p(\tau \mid \boldsymbol{\theta}_2)} \right)^2 \\ & \leq \frac{\|\boldsymbol{\Delta}\|^2}{4} \int_0^1 p(\tau \mid \boldsymbol{\theta}_t) \|\nabla \log p(\tau \mid \boldsymbol{\theta}_t)\|^2 dt. \end{aligned}$$

Multiplying by $H_{\gamma}(\tau)^r$ and integrating over τ yields

$$\begin{aligned} & \int_{\mathcal{T}} H_{\gamma}(\tau)^r \left(\sqrt{p(\tau \mid \boldsymbol{\theta}_1)} - \sqrt{p(\tau \mid \boldsymbol{\theta}_2)} \right)^2 d\tau \\ & \leq \frac{\|\boldsymbol{\Delta}\|^2}{4} \int_0^1 \mathbb{E}_{\tau \sim p(\cdot \mid \boldsymbol{\theta}_t)} \left[H_{\gamma}^r \left\| \sum_{h=0}^{H_{\gamma}-1} \nabla \log \pi_{\boldsymbol{\theta}_t}(\mathbf{a}_h \mid \mathbf{s}_h) \right\|^2 \right] dt. \end{aligned}$$

The score increments are martingale differences. Conditionally on H_{γ} , Assumption 3.1(i) therefore gives

$$\mathbb{E}_{\tau \sim p(\cdot \mid \boldsymbol{\theta}_t)} \left[\left\| \sum_{h=0}^{H_{\gamma}-1} \nabla \log \pi_{\boldsymbol{\theta}_t}(\mathbf{a}_h \mid \mathbf{s}_h) \right\|^2 \middle| H_{\gamma} \right] \leq C_1^2 H_{\gamma}.$$

Since the distribution of H_{γ} is independent of $\boldsymbol{\theta}_t$,

$$\mathbb{E} \left[H_{\gamma}^r \|\nabla \log p(\tau \mid \boldsymbol{\theta}_t)\|^2 \right] \leq C_1^2 M_{r+1}.$$

Substitution proves the claim. $\square$

Lemma D.4 (Estimation Error for **LVR-PG**). *Under Assumption 3.1, let $\Delta_k := \mathbb{E}[\|\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)\|_2^2]$ be the estimation error of the **LVR-PG** estimator. Letting $\lambda := 1 - (1-p)(1-\alpha)^2$ and letting $C_{\dagger, \omega} > 0$ denote the bound on the variance of the (defensive) importance weights, the following recurrence holds for all $k \in \llbracket K \rrbracket$:*

$$\Delta_{k+1} \leq (1-\lambda)\Delta_k + \frac{2(1-\lambda)C_{\dagger, \omega}}{B} \mathbb{E}[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2] + C_{\nu},$$

where $C_{\nu} := \left(\frac{p}{N} + \frac{4\alpha^2(1-p)}{B} \right) \nu^2$.

Proof. We analyze the estimation error conditioned on the Bernoulli random variable $c_k \sim \mathcal{B}(p)$. By the law of total expectation, the expected estimation error Δ_{k+1} decomposes as follows:

$$\begin{aligned} \Delta_{k+1} &= p \cdot \mathbb{E} \left[\|\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1})\|_2^2 \mid c_k = 1 \right] \\ &\quad + (1-p) \cdot \mathbb{E} \left[\|\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1})\|_2^2 \mid c_k = 0 \right]. \end{aligned}$$

When $c_k = 1$, the estimator executes a full-batch refresh. Because this relies on a new batch of N on-policy trajectories, the conditional error is bounded by the variance of the standard GPOMDP estimator obtained in Lemma D.1:

$$\mathbb{E} \left[\|\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1})\|_2^2 \mid c_k = 1 \right] \leq \frac{\nu^2}{N}.$$

When $c_k = 0$, the estimator applies the momentum update. Following the proof of Lemma 8 of Yuan et al. (2020), we condition on the history $\mathcal{F}_k$ available after computing $\boldsymbol{\theta}_{k+1}$ and before drawing c_k and the fresh trajectories. Under this conditioning, $\boldsymbol{\theta}_k$, $\boldsymbol{\theta}_{k+1}$, and $\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k)$ are fixed, while the fresh trajectories $\{\tau_i\}_{i=1}^B$ are independent samples from $\xi p_{\boldsymbol{\theta}_{k+1}} + (1-\xi)p_{\boldsymbol{\theta}_k}$. Substituting the momentum update and adding and subtracting the corresponding gradients gives:

$$\begin{aligned} \widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1}) &= (1-\alpha)(\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)) \\ &\quad + (1-\alpha) \left[\frac{1}{B} \sum_{i=1}^B (\omega_{\xi}(\boldsymbol{\theta}_{k+1}, \tau_i) \mathbf{g}(\tau_i \mid \boldsymbol{\theta}_{k+1}) - \omega_{\xi}(\boldsymbol{\theta}_k, \tau_i) \mathbf{g}(\tau_i \mid \boldsymbol{\theta}_k)) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1}) + \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) \right] \\ &\quad + \alpha \left[\frac{1}{B} \sum_{i=1}^B \omega_{\xi}(\boldsymbol{\theta}_{k+1}, \tau_i) \mathbf{g}(\tau_i \mid \boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1}) \right]. \end{aligned}$$

The three terms represent the previous estimation error, the error in estimating the consecutive-gradient difference, and the current-gradient estimation error, respectively. The latter two have zero conditional expectation by unbiasedness of the weighted estimates. Consequently, their cross terms with the previous estimation error vanish in conditional expectation. Applying $\|u+v\|_2^2 \leq 2\|u\|_2^2 + 2\|v\|_2^2$ to these two fresh error terms, and using independence across trajectories, yields

$$\begin{aligned} \mathbb{E} \left[\|\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1})\|_2^2 \mid \mathcal{F}_k, c_k = 0 \right] &\leq (1-\alpha)^2 \|\widehat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)\|_2^2 \\ &\quad + \frac{2(1-\alpha)^2}{B} \mathbb{E}_{\tau \sim \xi p_{\boldsymbol{\theta}_{k+1}} + (1-\xi)p_{\boldsymbol{\theta}_k}} \left[\|\omega_{\xi}(\boldsymbol{\theta}_{k+1}, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_{k+1}) - \omega_{\xi}(\boldsymbol{\theta}_k, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_k) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1}) + \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)\|_2^2 \right] \\ &\quad + \frac{2\alpha^2}{B} \mathbb{E}_{\tau \sim \xi p_{\boldsymbol{\theta}_{k+1}} + (1-\xi)p_{\boldsymbol{\theta}_k}} \left[\|\omega_{\xi}(\boldsymbol{\theta}_{k+1}, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1})\|_2^2 \right]. \end{aligned}$$

The factors $1/B$ follow because the centered contributions from distinct trajectories are conditionally independent. For the consecutive-gradient error, removing the centering and applying Lemma B.1 gives

$$\begin{aligned} \mathbb{E}_{\tau \sim \xi p_{\boldsymbol{\theta}_{k+1}} + (1-\xi)p_{\boldsymbol{\theta}_k}} \left[\|\omega_{\xi}(\boldsymbol{\theta}_{k+1}, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_{k+1}) - \omega_{\xi}(\boldsymbol{\theta}_k, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_k) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1}) + \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)\|_2^2 \right] &\leq \mathbb{E}_{\tau \sim \xi p_{\boldsymbol{\theta}_{k+1}} + (1-\xi)p_{\boldsymbol{\theta}_k}} \left[\|\omega_{\xi}(\boldsymbol{\theta}_{k+1}, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_{k+1}) - \omega_{\xi}(\boldsymbol{\theta}_k, \tau) \mathbf{g}(\tau \mid \boldsymbol{\theta}_k)\|_2^2 \right] \\ &\leq C_{\dagger, \omega} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2. \end{aligned}$$

Here, the first inequality uses the fact that the expectation of the weighted gradient difference is $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)$. The second inequality is the mean-squared error bound provided by Lemma B.1. For the current-

gradient error, we similarly obtain:

$$\begin{aligned}
 & \mathbb{E}_{\tau \sim \xi p_{\theta_{k+1}} + (1-\xi)p_{\theta_k}} \left[\|\omega_\xi(\theta_{k+1}, \tau) \mathbf{g}(\tau \mid \theta_{k+1}) - \nabla_{\theta} J(\theta_{k+1})\|_2^2 \right] \\
 & \leq \mathbb{E}_{\tau \sim \xi p_{\theta_{k+1}} + (1-\xi)p_{\theta_k}} \left[\|\omega_\xi(\theta_{k+1}, \tau) \mathbf{g}(\tau \mid \theta_{k+1})\|_2^2 \right] \\
 & = \mathbb{E}_{\tau \sim p_{\theta_{k+1}}} \left[\|\omega_\xi(\theta_{k+1}, \tau) \mathbf{g}(\tau \mid \theta_{k+1})\|_2^2 \right] \\
 & \leq \frac{1}{\xi} \mathbb{E}_{\tau \sim p_{\theta_{k+1}}} \left[\|\mathbf{g}(\tau \mid \theta_{k+1})\|_2^2 \right] \leq \frac{\nu^2}{\xi}.
 \end{aligned}$$

The equality follows by canceling one importance weight against the mixture distribution. The final inequalities use $\omega_\xi(\theta_{k+1}, \tau) \leq 1/\xi$ and the uncentered second-moment bound established in the proof of Lemma D.1. Substituting these bounds into the conditional estimation error inequality gives:

$$\begin{aligned}
 & \mathbb{E} \left[\|\hat{\nabla}_{\theta}^{\dagger} J(\theta_{k+1}) - \nabla_{\theta} J(\theta_{k+1})\|_2^2 \mid \mathcal{F}_k, c_k = 0 \right] \\
 & \leq (1-\alpha)^2 \|\hat{\nabla}_{\theta}^{\dagger} J(\theta_k) - \nabla_{\theta} J(\theta_k)\|_2^2 + \frac{2(1-\alpha)^2 C_{\dagger, \omega}}{B} \|\theta_{k+1} - \theta_k\|_2^2 + \frac{2\alpha^2 \nu^2}{\xi B}.
 \end{aligned}$$

Taking expectation over the preceding history, and using the independence of c_k from that history, yields,

$$\begin{aligned}
 & \mathbb{E} \left[\|\hat{\nabla}_{\theta}^{\dagger} J(\theta_{k+1}) - \nabla_{\theta} J(\theta_{k+1})\|_2^2 \mid c_k = 0 \right] \\
 & \leq (1-\alpha)^2 \Delta_k + \frac{2(1-\alpha)^2 C_{\dagger, \omega}}{B} \mathbb{E} [\|\theta_{k+1} - \theta_k\|_2^2] + \frac{2\alpha^2 \nu^2}{\xi B}.
 \end{aligned}$$

Summing the two branches, we obtain,

$$\begin{aligned}
 \Delta_{k+1} & \leq p \left(\frac{\nu^2}{N} \right) + (1-p) \left[(1-\alpha)^2 \Delta_k + \frac{2(1-\alpha)^2 C_{\dagger, \omega}}{B} \mathbb{E} [\|\theta_{k+1} - \theta_k\|_2^2] + \frac{2\alpha^2 \nu^2}{\xi B} \right] \\
 & = (1-p)(1-\alpha)^2 \Delta_k + \frac{2(1-p)(1-\alpha)^2 C_{\dagger, \omega}}{B} \mathbb{E} [\|\theta_{k+1} - \theta_k\|_2^2] + \frac{p\nu^2}{N} + \frac{2\alpha^2(1-p)\nu^2}{\xi B}.
 \end{aligned}$$

Recognizing that $(1-p)(1-\alpha)^2 = 1-\lambda$, and plugging our choice of $\xi = 1/2$ from Lemma B.1, this yields the recurrence:

$$\Delta_{k+1} \leq (1-\lambda)\Delta_k + \frac{2(1-\lambda)C_{\dagger, \omega}}{B} \mathbb{E} [\|\theta_{k+1} - \theta_k\|_2^2] + \left(\frac{p}{N} + \frac{4\alpha^2(1-p)}{B} \right) \nu^2.$$

□

E CONVERGENCE

In this section, the last-iterate global-convergence guarantee for the **LVR-PG** algorithm is provided. In particular, Theorem 5.1 provides the sample-complexity for the **LVR-PG** framework, and while Corollaries E.1 and E.2 provide the sample complexities **DEF-PAGE-PG** and **DEF-STORM-PG** as instantiations of **LVR-PG**. Finally, Corollary 5.1 provides the convergence result for **LVR-PG** when instantiated for Fisher-non-degenerate policies. The lemmas used to prove the results of this section are provided in Appendix D and H.

E.1 Proof of Theorem 5.1

Here, the proof of Theorem 5.1 is provided. Leveraging the proof technique in Theorem 4.5 of Xu et al. (2019) for projected gradient ascent, we construct a recurrence that depends on both the sub-optimality gap and the variance of the estimator. We show that this recurrence achieves minimax-optimal sample complexity up to logarithmic factors and horizon terms.

Theorem 5.1 (Global Convergence of **LVR-PG**). *Under Assumptions 3.1 and 3.2, for any target accuracy $\epsilon > 0$ and fixed mini-batch size $B \geq 1$, let the batch sizes be $N, N_0 = \Theta(\epsilon^{-2})$ and the learning rate be $\eta = \Theta(B\epsilon)$. For appropriately chosen hyperparameters satisfying $p, \alpha = \mathcal{O}(B\epsilon^2)$ and $p + \alpha = \Omega(B\epsilon^2)$ and choosing $\xi = 1/2$,*

LVR-PG guarantees an expected suboptimality gap $J^* - \mathbb{E}[J(\boldsymbol{\theta}_K)] \leq \psi + \epsilon$ with a total expected sample complexity of at most:

$$n = \tilde{\mathcal{O}}\left(\frac{\chi^3 R_{\max}^2}{(1-\gamma)^4 \epsilon^2}\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

Proof. Following the proof of Theorem 4.5 by Xu et al. (2019), we start by redefining the objective function to incorporate the constraint set via the non-smooth indicator function $\mathbb{I}_{\Theta}(\boldsymbol{\theta})$, introducing $\mathcal{J}(\boldsymbol{\theta}) := J(\boldsymbol{\theta}) - \mathbb{I}_{\Theta}(\boldsymbol{\theta})$. We formally define the generalized gradient mapping as $\mathcal{G}_{\eta}(\boldsymbol{\theta}) := \frac{1}{\eta}(\mathcal{P}_{\Theta}(\boldsymbol{\theta} + \eta \nabla J(\boldsymbol{\theta})) - \boldsymbol{\theta})$. Note that the mapping $\mathcal{G}_{\eta}$ serves as a generalized projected gradient at $\boldsymbol{\theta}$. By definition, in the unconstrained setting where $\Theta = \mathbb{R}^{d_{\Theta}}$, this recovers the standard gradient, yielding $\mathcal{G}_{\eta}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. According to the definition of $\mathbb{I}_{\Theta}(\boldsymbol{\theta})$:

$$\mathbb{I}_{\Theta}(\mathbf{u}) = \begin{cases} 0, & \text{if } \mathbf{u} \in \Theta \\ +\infty & \text{otherwise} \end{cases},$$

we have $\arg \max_{\boldsymbol{\theta} \in \mathbb{R}^{d_{\Theta}}} \mathcal{J}(\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) := \boldsymbol{\theta}^*$. Letting $\tilde{\mathcal{J}}^* := \mathcal{J}^* - \psi$ and defining $\mu := 1/\chi^2$, we reformulate the WGD condition (Assumption 3.2) over the bounded domain using this gradient mapping:

$$\sqrt{\mu} \max\{0, \tilde{\mathcal{J}}^* - \mathcal{J}(\boldsymbol{\theta}_k)\} \leq \|\mathcal{G}_{\eta}(\boldsymbol{\theta}_k)\|_2$$

Considering an iteration $k \in \llbracket K \rrbracket$, the parameter update applies projected stochastic gradient ascent: $\boldsymbol{\theta}_{k+1} = \mathcal{P}_{\Theta}(\boldsymbol{\theta}_k + \eta \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k))$. By expanding the Euclidean projection and discarding terms independent of the optimization variable, this is equivalent to:

$$\boldsymbol{\theta}_{k+1} = \arg \min_{\mathbf{u} \in \mathbb{R}^d} \left\{ \mathbb{I}_{\Theta}(\mathbf{u}) + \frac{1}{2\eta} \|\mathbf{u} - \boldsymbol{\theta}_k\|_2^2 - \langle \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k), \mathbf{u} \rangle \right\}.$$

We define the stochastic gradient mapping as $\tilde{\mathcal{G}}_k := \frac{1}{\eta}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k)$. Let $\mathbf{g} \in \partial \mathbb{I}_{\Theta}(\boldsymbol{\theta}_{k+1})$ be a subgradient of the indicator function. According to the optimality condition of the projection step, we have $\mathbf{g} + \frac{1}{\eta}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k) = 0$. By the convexity of $\mathbb{I}_{\Theta}$, it follows that:

$$\begin{aligned} \mathbb{I}_{\Theta}(\boldsymbol{\theta}_{k+1}) &\leq \mathbb{I}_{\Theta}(\boldsymbol{\theta}_k) + \langle \mathbf{g}, \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle \\ &= \mathbb{I}_{\Theta}(\boldsymbol{\theta}_k) - \left\langle \frac{1}{\eta}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \right\rangle. \end{aligned}$$

From the L -smoothness of $J(\boldsymbol{\theta})$ obtained in Lemma D.2, we can write the ascent lemma as follows:

$$J(\boldsymbol{\theta}_{k+1}) \geq J(\boldsymbol{\theta}_k) + \langle \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle - \frac{L}{2} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2.$$

We construct the one-step descent for the constrained objective $\mathcal{J}(\boldsymbol{\theta})$ by subtracting the indicator function upper bound from the smoothness lower bound. Expanding the inner product isolates the quadratic penalty, yielding:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\theta}_{k+1}) &\geq \mathcal{J}(\boldsymbol{\theta}_k) + \langle \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle + \frac{1}{\eta} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 - \frac{L}{2} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \\ &= \mathcal{J}(\boldsymbol{\theta}_k) + \eta \langle \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k), \tilde{\mathcal{G}}_k \rangle + \eta \|\tilde{\mathcal{G}}_k\|_2^2 - \frac{L}{2} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2. \end{aligned}$$

Applying Young's inequality $\langle x, y \rangle \geq -\frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2$ to the inner product, we bound the relation:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\theta}_{k+1}) &\geq \mathcal{J}(\boldsymbol{\theta}_k) - \frac{\eta}{2} \|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k)\|_2^2 + \frac{\eta}{2} \|\tilde{\mathcal{G}}_k\|_2^2 - \frac{L}{2} \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \\ &= \mathcal{J}(\boldsymbol{\theta}_k) - \frac{\eta}{2} \|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^{\dagger} J(\boldsymbol{\theta}_k)\|_2^2 + \frac{\eta}{4} \|\tilde{\mathcal{G}}_k\|_2^2 + \left(\frac{1}{4\eta} - \frac{L}{2} \right) \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2. \end{aligned}$$

By expanding the definitions of the gradient mappings and applying the property of the Euclidean projection operator $\|\mathcal{P}_\Theta(x) - \mathcal{P}_\Theta(y)\|_2 \leq \|x - y\|_2$, we bound the difference between the exact and stochastic gradient mappings:

$$\begin{aligned} \|\mathcal{G}_\eta(\boldsymbol{\theta}_k) - \tilde{\mathcal{G}}_k\|_2 &= \frac{1}{\eta} \left\| \mathcal{P}_\Theta(\boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)) - \mathcal{P}_\Theta(\boldsymbol{\theta}_k + \eta \hat{\nabla}_{\boldsymbol{\theta}}^\dagger J(\boldsymbol{\theta}_k)) \right\|_2 \\ &\leq \frac{1}{\eta} \left\| (\boldsymbol{\theta}_k + \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)) - (\boldsymbol{\theta}_k + \eta \hat{\nabla}_{\boldsymbol{\theta}}^\dagger J(\boldsymbol{\theta}_k)) \right\|_2 \\ &= \|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^\dagger J(\boldsymbol{\theta}_k)\|_2. \end{aligned}$$

Utilizing the inequality $\|\mathcal{G}_\eta(\boldsymbol{\theta}_k)\|_2^2 \leq 2\|\tilde{\mathcal{G}}_k\|_2^2 + 2\|\mathcal{G}_\eta(\boldsymbol{\theta}_k) - \tilde{\mathcal{G}}_k\|_2^2$, which follows from adding and subtracting the stochastic gradient mapping and applying Young's inequality, we bound the remaining mapping term:

$$\frac{\eta}{4} \|\tilde{\mathcal{G}}_k\|_2^2 \geq \frac{\eta}{8} \|\mathcal{G}_\eta(\boldsymbol{\theta}_k)\|_2^2 - \frac{\eta}{4} \|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^\dagger J(\boldsymbol{\theta}_k)\|_2^2.$$

Substituting this lower bound yields the final single-step proximal descent inequality:

$$\mathcal{J}(\boldsymbol{\theta}_{k+1}) \geq \mathcal{J}(\boldsymbol{\theta}_k) - \frac{3\eta}{4} \|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k) - \hat{\nabla}_{\boldsymbol{\theta}}^\dagger J(\boldsymbol{\theta}_k)\|_2^2 + \frac{\eta}{8} \|\mathcal{G}_\eta(\boldsymbol{\theta}_k)\|_2^2 + \left(\frac{1}{4\eta} - \frac{L}{2} \right) \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2. \quad (10)$$

Reversing the sign, applying the total expectation, and substituting the suboptimality $r_k := \mathbb{E}[\tilde{\mathcal{J}}^* - \mathcal{J}(\boldsymbol{\theta}_k)]$ and the estimation error $\Delta_k := \mathbb{E}[\|\hat{\nabla}_{\boldsymbol{\theta}}^\dagger J(\boldsymbol{\theta}_k) - \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k)\|_2^2]$, we obtain the projected suboptimality recurrence:

$$r_{k+1} \leq r_k - \frac{\eta}{8} \mathbb{E}[\|\mathcal{G}_\eta(\boldsymbol{\theta}_k)\|_2^2] + \frac{3\eta}{4} \Delta_k - \left(\frac{1}{4\eta} - \frac{L}{2} \right) \mathbb{E}[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2]. \quad (11)$$

From Lemma D.4, the unified variance recurrence uses the effective decay rate $\lambda := 1 - (1-p)(1-\alpha)^2$ and noise floor $C_\nu := \left(\frac{p}{N} + \frac{4\alpha^2(1-p)}{B} \right) \nu^2$. The variance recurrence is:

$$\Delta_{k+1} \leq (1-\lambda)\Delta_k + \frac{(1-\lambda)C_{\dagger,\omega}}{B} \mathbb{E}[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2] + C_\nu. \quad (12)$$

We introduce the potential function $\Phi_k := r_k + c\Delta_k$ with a coupling constant $c > 0$. Because the estimation error Δ_k is intrinsically non-negative, the potential strictly upper-bounds the suboptimality ($\Phi_k \geq r_k$). Consequently, establishing the convergence of Φ_k is sufficient to guarantee the convergence of r_k . Combining equations (11) and (12), we construct the unified recurrence:

$$\begin{aligned} \Phi_{k+1} &\leq r_k - \frac{\eta}{8} \mathbb{E}[\|\mathcal{G}_\eta(\boldsymbol{\theta}_k)\|_2^2] + \frac{3\eta}{4} \Delta_k - \left(\frac{1}{4\eta} - \frac{L}{2} \right) \mathbb{E}[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2] \\ &\quad + c \left((1-\lambda)\Delta_k + \frac{2(1-\lambda)C_{\dagger,\omega}}{B} \mathbb{E}[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2] + C_\nu \right) \\ &= \Phi_k - \frac{\eta}{8} \mathbb{E}[\|\mathcal{G}_\eta(\boldsymbol{\theta}_k)\|_2^2] - \left(c\lambda - \frac{3\eta}{4} \right) \Delta_k \\ &\quad - \left(\frac{1}{4\eta} - \frac{L}{2} - c \frac{2(1-\lambda)C_{\dagger,\omega}}{B} \right) \mathbb{E}[\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2] + cC_\nu. \end{aligned}$$

To simplify the recurrence, we determine the constant c by requiring the Δ_k coefficient to retain an exact $-\frac{3\eta}{4}$ residual descent rate:

$$c\lambda - \frac{3\eta}{4} = \frac{3\eta}{4} \Rightarrow c = \frac{3\eta}{2\lambda}.$$

Substituting this choice of c , we require the sequence distance coefficient to remain non-negative to eliminate it from the upper bound:

$$\frac{1}{4\eta} - \frac{L}{2} - \frac{3\eta(1-\lambda)C_{\dagger,\omega}}{\lambda B} \geq 0.$$

To ensure the sequence distance coefficient provides sufficient slack to absorb the variance accumulation, we tighten the standard smoothness restriction. By rigorously restricting the baseline learning rate to $\eta \leq \frac{1}{4L}$, we enforce the bound $\frac{L}{2} \leq \frac{1}{8\eta}$, which guarantees the strictly positive residual $\frac{1}{4\eta} - \frac{L}{2} \geq \frac{1}{4\eta} - \frac{1}{8\eta} = \frac{1}{8\eta}$. Substituting this residual isolates the maximum stable learning rate required to absorb the variance accumulation:

$$\frac{1}{8\eta} \geq \frac{3\eta(1-\lambda)C_{\dagger,\omega}}{\lambda B} \Rightarrow \eta \leq \sqrt{\frac{\lambda B}{24(1-\lambda)C_{\dagger,\omega}}}.$$

Under these stability conditions, the recurrence simplifies to:

$$\Phi_{k+1} \leq \Phi_k - \frac{\eta}{8} \mathbb{E} [\|\mathcal{G}_\eta(\theta_k)\|_2^2] - \frac{3\eta}{4} \Delta_k + \frac{3\eta C_\nu}{2\lambda}. \quad (13)$$

Substituting the modified gradient domination condition and applying Jensen's inequality for both the maximum and square operations yields:

$$\Phi_{k+1} \leq \Phi_k - \frac{\mu\eta}{8} \max\{0, r_k\}^2 - \frac{3\eta}{4} \Delta_k + \frac{3\eta C_\nu}{2\lambda}. \quad (14)$$

To establish a last-iterate guarantee, we must prove that the algorithm does not leave the target accuracy region once it enters it. We explicitly define the non-negative sequence $x_k := \max\{0, \Phi_k\}$ and show it unconditionally satisfies a unified descent recurrence. We evaluate the transition for x_k based on the sign of Φ_k . If $\Phi_k \leq 0$, we have $x_k = 0$. From Equation (13), dropping the non-positive descent terms yields $\Phi_{k+1} \leq \Phi_k + \frac{3\eta C_\nu}{2\lambda} \leq \frac{3\eta C_\nu}{2\lambda}$. Taking the positive part of both sides guarantees $x_{k+1} \leq \frac{3\eta C_\nu}{2\lambda}$, which satisfies the target recurrence evaluated at $x_k = 0$. Conversely, if $\Phi_k > 0$, we have $x_k = \Phi_k$. Because $\Phi_k \leq \max\{0, r_k\} + c\Delta_k$, it follows that at least one of the following conditions must hold: either $\max\{0, r_k\} \geq \frac{\Phi_k}{2}$, or $c\Delta_k \geq \frac{\Phi_k}{2}$. If the first case holds ($\max\{0, r_k\} \geq \frac{\Phi_k}{2}$), we bound the suboptimality descent while discarding the negative variance term:

$$-\frac{\mu\eta}{8} \max\{0, r_k\}^2 \leq -\frac{\mu\eta}{8} \left(\frac{x_k}{2}\right)^2 = -\frac{\mu\eta}{32} x_k^2.$$

If the second case holds ($c\Delta_k \geq \frac{\Phi_k}{2}$), utilizing our choice of $c = 3\eta/2\lambda$ yields $\Delta_k \geq \frac{\lambda x_k}{3\eta}$. Discarding the suboptimality term bounds the variance descent term:

$$-\frac{3\eta}{4} \Delta_k \leq -\frac{3\eta}{4} \left(\frac{\lambda x_k}{3\eta}\right) = -\frac{\lambda}{4} x_k.$$

Bounding the descent in Equation (14) by the minimum of these two rates, and observing that taking the maximum operator $x_{k+1} = \max\{0, \Phi_{k+1}\}$ on the left preserves the inequality because the right-hand side is inherently non-negative or bounded safely by 0, we obtain the final unconditional recurrence:

$$x_{k+1} \leq x_k - \min\left\{\frac{\mu\eta}{32} x_k^2, \frac{\lambda}{4} x_k\right\} + \frac{3\eta C_\nu}{2\lambda}.$$

We directly apply Lemma H.2 to the invariant sequence x_k with $a = \frac{\mu\eta}{32}$, $b = \frac{\lambda}{4}$, and $e = \frac{3\eta C_\nu}{2\lambda}$. Because $\lambda \leq 1$ structurally, we inherently satisfy the lemma's contraction requirement $\rho_\epsilon \leq b = \frac{\lambda}{4} < 1$. Hence, from Lemma H.2, the iteration complexity is described by:

$$K \geq \max\left\{\frac{32}{\mu\eta\epsilon}, \frac{4}{\lambda}\right\} \log_+ \frac{2\Phi_0}{\epsilon}. \quad (15)$$

To guarantee convergence to the target accuracy ϵ , the noise term must satisfy two conditions. The first condition $e \leq a\frac{\epsilon^2}{2}$ requires $\frac{3\eta C_\nu}{2\lambda} \leq \frac{\mu\eta\epsilon^2}{64}$, which simplifies to bounding the noise floor $C_\nu \leq \frac{\mu\lambda\epsilon^2}{96}$. The second condition $e \leq b\frac{\epsilon}{2}$ requires $\frac{3\eta C_\nu}{2\lambda} \leq \frac{\lambda\epsilon}{8}$, which dictates the maximum stable learning rate $\eta \leq \frac{\lambda^2\epsilon}{12C_\nu}$. We satisfy the first accuracy condition $C_\nu \leq \frac{\mu\lambda\epsilon^2}{96}$ by configuring the algorithmic parameters such that each of the two additive terms in $C_\nu = \left(\frac{p}{N} + \frac{4\alpha^2(1-p)}{B}\right)\nu^2$ is independently bounded by half of the target limit, $\frac{\mu\lambda\epsilon^2}{192}$. With the unified decay rate $\lambda = p + (1-p)(2\alpha - \alpha^2)$, we evaluate both terms to define the feasible geometric region.

First Term of C_ν and Batch Size. Observing that $2\alpha - \alpha^2 \geq 0$, we establish the lower bound $\lambda \geq p$. It is therefore sufficient to enforce:

$$\frac{p\nu^2}{N} \leq \frac{\mu p \epsilon^2}{192} \implies N \geq \frac{192\nu^2}{\mu \epsilon^2}.$$

To minimize sample cost while satisfying this requirement, we select a full batch size $N = \Theta(\epsilon^{-2})$ and initialize with $N_0 = N = \Theta(\epsilon^{-2})$.

Second Term of C_ν and Momentum. Observing that $2\alpha - \alpha^2 = \alpha(2 - \alpha) \geq \alpha$ for any $\alpha \in (0, 1]$, and dropping the positive p term from λ , we obtain the lower bound $\lambda \geq (1 - p)\alpha$. Substituting this lower bound into the right-hand side, it is strictly sufficient to require:

$$\frac{4\alpha^2(1 - p)\nu^2}{B} \leq \frac{\mu(1 - p)\alpha\epsilon^2}{192}.$$

Because $\alpha > 0$, we divide both sides by the common factor $\alpha(1 - p)$ to directly isolate α :

$$\frac{4\alpha\nu^2}{B} \leq \frac{\mu\epsilon^2}{192} \implies \alpha \leq \frac{\mu B \epsilon^2}{768\nu^2}.$$

This establishes the strict upper bound $\alpha = \mathcal{O}(B\epsilon^2)$.

Expected Sample Cost and Refresh Probability. We choose $p, \alpha \geq 0$ such that

$$p + \alpha = \frac{\mu B \epsilon^2}{768\nu^2}.$$

This ensures the momentum constraint established above. For sufficiently small ϵ , we have $p + \alpha \leq 1/2$, and hence

$$p + \alpha \leq \lambda = p + (1 - p)(2\alpha - \alpha^2) \leq p + 2\alpha \leq 2(p + \alpha).$$

Therefore, $\lambda = \Theta(\mu B \epsilon^2 / \nu^2)$, with universal implicit constants. Moreover, since $p \leq \mu B \epsilon^2 / (768\nu^2)$,

$$\begin{aligned} pN + (1 - p)B &\leq \frac{\mu B \epsilon^2}{768\nu^2} \left(\frac{192\nu^2}{\mu \epsilon^2} + 1 \right) + B \\ &\leq \frac{5B}{4} + \frac{1}{2} = \mathcal{O}(B). \end{aligned}$$

Learning Rate. To achieve the optimal total sample complexity $n = \tilde{\mathcal{O}}(\epsilon^{-2})$, and match the lower-bound established by Theorem 3.1 up to logarithmic factors, the iteration complexity K must be bounded by $\mathcal{O}((B\epsilon^2)^{-1})$. This bound limits the total number of optimization steps to guarantee that the cumulative sample cost of the $\mathcal{O}(B)$ per-iteration updates does not breach the optimal $\tilde{\mathcal{O}}(\epsilon^{-2})$ global sample complexity. Given the initial batch $N_0 = \Theta(\epsilon^{-2})$ and an expected per-iteration sample cost bounded by $\mathcal{O}(B)$, the number of iterations must satisfy $K \cdot \mathcal{O}(B) \leq \tilde{\mathcal{O}}(\epsilon^{-2})$, which mandates $K \leq \mathcal{O}((B\epsilon^2)^{-1})$. The sufficient iteration budget in Equation (15) contains the estimator-decay term $4/\lambda$. To keep this term within the desired iteration budget $\mathcal{O}((B\epsilon^2)^{-1})$, we impose $\lambda = \Omega(B\epsilon^2)$. This condition controls the iteration complexity of our analysis; it does not establish that configurations with smaller λ necessarily require more iterations to reach the target accuracy. Expanding the definition of the decay rate yields: $\lambda = 1 - (1 - p)(1 - \alpha)^2 = p + 2\alpha - \alpha^2 - 2p\alpha + p\alpha^2$. Since the upper bounds $p = \mathcal{O}(B\epsilon^2)$ and $\alpha = \mathcal{O}(B\epsilon^2)$ are established, the higher-order terms α^2 , $p\alpha$, and $p\alpha^2$ scale as $\mathcal{O}(B^2\epsilon^4)$ or smaller. Consequently, for sufficiently small target accuracy $\epsilon \rightarrow 0$, the decay rate is dominated by its linear components, reducing asymptotically to $\lambda = p + 2\alpha - \mathcal{O}(B^2\epsilon^4)$. To satisfy the $\lambda = \Omega(B\epsilon^2)$ decay floor, these dominant terms must jointly fulfill the geometric condition $p + \alpha = \Omega(B\epsilon^2)$. For any configuration (p, α) satisfying this combined system of bounds, the decay rate and noise floor scale precisely as $\lambda = \Theta(B\epsilon^2)$ and $C_\nu = \Theta(B\epsilon^4)$, respectively. We evaluate the remaining step-size restrictions using these asymptotic rates. The last-iterate stability condition $\eta \leq \frac{\lambda^2 \epsilon}{12C_\nu}$ resolves to $\eta \leq \Theta(B\epsilon)$. Concurrently, the variance recurrence threshold evaluates to $\sqrt{\frac{\lambda B}{12(1 - \lambda)C_{\dagger, \omega}}} = \Theta(B\epsilon)$. By explicitly selecting the step size η to match this scaling alongside the L -smoothness limit, all stability requirements are unconditionally satisfied:

$$\eta = \min \left\{ \frac{1}{4L}, \frac{\lambda^2 \epsilon}{12C_\nu}, \sqrt{\frac{\lambda B}{24(1 - \lambda)C_{\dagger, \omega}}} \right\} = \Theta(B\epsilon).$$

Sample Complexity. With these conditions established, Lemma H.2 guarantees that it suffices to choose the number of iterations as:

$$K = \left\lceil \max \left\{ \frac{32}{\mu\eta\epsilon}, \frac{4}{\lambda} \right\} \log_+ \frac{2\Phi_0}{\epsilon} \right\rceil.$$

Substituting our choice of η into the iteration bound gives:

$$K = \left\lceil \max \left\{ \frac{128L}{\mu\epsilon}, \frac{384C_\nu}{\mu\lambda^2\epsilon^2}, \frac{32\sqrt{24(1-\lambda)C_{\dagger,\omega}}}{\mu\epsilon\sqrt{\lambda B}}, \frac{4}{\lambda} \right\} \log_+ \frac{2\Phi_0}{\epsilon} \right\rceil.$$

The noise constraint $C_\nu \leq \mu\lambda\epsilon^2/96$ implies

$$\frac{384C_\nu}{\mu\lambda^2\epsilon^2} \leq \frac{384}{\mu\lambda^2\epsilon^2} \frac{\mu\lambda\epsilon^2}{96} = \frac{4}{\lambda}.$$

Thus, this term is absorbed by the estimator-decay term. Moreover, the established bound $\lambda \geq \mu B\epsilon^2/(768\nu^2)$ yields

$$\frac{4}{\lambda} \leq \frac{3072\nu^2}{\mu B\epsilon^2}.$$

Using the same lower bound on λ and $1 - \lambda \leq 1$, the variance-stability term satisfies

$$\begin{aligned} \frac{32\sqrt{24(1-\lambda)C_{\dagger,\omega}}}{\mu\epsilon\sqrt{\lambda B}} &\leq \frac{32\sqrt{24C_{\dagger,\omega}}}{\mu\epsilon} \sqrt{\frac{768\nu^2}{\mu B^2\epsilon^2}} \\ &= \frac{3072\sqrt{2}\nu\sqrt{C_{\dagger,\omega}}}{\mu^{3/2}B\epsilon^2}. \end{aligned}$$

Combining these bounds, absorbing numerical constants and logarithmic factors, and accounting for the ceiling, we obtain:

$$K = \tilde{\mathcal{O}} \left(1 + \frac{L}{\mu\epsilon} + \frac{\nu^2}{\mu B\epsilon^2} + \frac{\nu\sqrt{C_{\dagger,\omega}}}{\mu^{3/2}B\epsilon^2} \right).$$

The expected total sample complexity n is determined by the initial batch size N_0 plus the sum of the expected batch sizes drawn over the K iterations:

$$n = N_0 + K(pN + (1-p)B).$$

By substituting $N_0 = \lceil 192\nu^2/(\mu\epsilon^2) \rceil$, the iteration bound above, and the bounded per-iteration cost $pN + (1-p)B = \mathcal{O}(B)$, we obtain:

$$n = \tilde{\mathcal{O}} \left(1 + B + \frac{BL}{\mu\epsilon} + \frac{\nu^2}{\mu\epsilon^2} + \frac{\nu\sqrt{C_{\dagger,\omega}}}{\mu^{3/2}\epsilon^2} \right).$$

Thus, for fixed B and fixed positive problem-dependent constants, the complexity is $n = \tilde{\mathcal{O}}(\epsilon^{-2})$ as $\epsilon \rightarrow 0$. To establish the asymptotic dependence on the reward bounds and the effective horizon, we substitute the problem-dependent constants. From Lemma D.1, the GPOMDP second-moment bound scales as $\nu^2 = \mathcal{O}(R_{\max}^2(1-\gamma)^{-3})$. From Lemma B.1, the defensive mean-squared smoothness constant scales as $C_{\dagger,\omega} = \mathcal{O}(R_{\max}^2(1-\gamma)^{-5})$. Substituting these bounds into the terms proportional to ϵ^{-2} , and using $\mu = \chi^{-2}$, yields:

$$n = \tilde{\mathcal{O}} \left(\frac{R_{\max}^2}{\epsilon^2} \left[\frac{\chi^2}{(1-\gamma)^3} + \frac{\chi^3}{(1-\gamma)^4} \right] \right),$$

where the policy-regularity and defensive-mixture constants are treated as fixed. This establishes the minimax-optimal dependence on ϵ , up to logarithmic factors, while retaining the explicit dependence on the WGD constant and the effective horizon. $\square$

E.2 Corollaries of Theorem 5.1

Here, we consider the sample complexity of two specific instantiations of **LVR-PG**: DEF-PAGE-PG and DEF-STORM-PG. First, we show how the parameter λ , defined in Lemma B.1, changes setting the α and p parameters. Then, we study the batch size N , mini-batch size B and learning rate η to recover the desired sample complexity.

Corollary E.1 (Sample Complexity of DEF-PAGE-PG). *Under Assumptions 3.1–3.2, for a fixed mini-batch size $B \geq 1$ and sufficiently small target accuracy $\epsilon > 0$, through an appropriate choice of the algorithmic parameters (full batch size N , initial batch size N_0 , refresh probability p , and learning rate η), DEF-PAGE-PG with $\alpha = 0$ guarantees $J^* - \mathbb{E}[J(\theta_k)] \leq \psi + \epsilon$ with a sample complexity of at most:*

$$\begin{aligned} n &= N_0 + K \left(1 + B - \frac{B}{N} \right) \\ &= \tilde{\mathcal{O}} \left(\frac{R_{\max}^2}{\epsilon^2} \left[\frac{\chi^2}{(1-\gamma)^3} + \frac{\chi^3}{(1-\gamma)^4} \right] \right). \end{aligned}$$

Here, $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors, and B , the policy-regularity constants, and the defensive-mixture parameter are treated as fixed.

Proof. Setting the momentum parameter to $\alpha = 0$ specializes the generic variance recurrence used in the proof of Theorem 5.1 to the DEF-PAGE-PG estimator. Under this configuration, the effective decay rate λ and C_ν simplify to:

$$\lambda = p, \quad C_\nu = \frac{p\nu^2}{N}.$$

For $p > 0$, the noise constraint $C_\nu \leq \mu\lambda\epsilon^2/96$ reduces to:

$$\frac{p\nu^2}{N} \leq \frac{\mu p \epsilon^2}{96} \implies N \geq \frac{96\nu^2}{\mu\epsilon^2}.$$

Accordingly, we select $N_0 = N = \lceil 96\nu^2/(\mu\epsilon^2) \rceil$ and $p = 1/N$. For sufficiently small ϵ , we have $96\nu^2/(\mu\epsilon^2) \geq 1$, so that

$$\frac{96\nu^2}{\mu\epsilon^2} \leq N \leq \frac{192\nu^2}{\mu\epsilon^2}.$$

Next, we evaluate the stability constraints on the learning rate. For the variance stability requirement $\eta \leq \sqrt{\lambda B/[24(1-\lambda)C_{\text{DEF},\omega}]}$, using $\lambda = 1/N$ and $1-\lambda \leq 1$, it suffices to enforce:

$$\eta \leq \sqrt{\frac{\mu B}{4608\nu^2 C_{\text{DEF},\omega}}} \epsilon \leq \sqrt{\frac{B}{24N C_{\text{DEF},\omega}}}.$$

For the accuracy requirement $\eta \leq \lambda^2\epsilon/(12C_\nu)$, substitution gives:

$$\frac{\lambda^2\epsilon}{12C_\nu} = \frac{pN\epsilon}{12\nu^2} = \frac{\epsilon}{12\nu^2}.$$

Defining

$$c_\eta := \min \left\{ \frac{1}{12\nu^2}, \sqrt{\frac{\mu B}{4608\nu^2 C_{\text{DEF},\omega}}} \right\},$$

we therefore select $\eta = \min\{1/(4L), c_\eta\epsilon\}$. The sufficient iteration budget established in the proof of Theorem 5.1 is $K \geq \max\{32/(\mu\eta\epsilon), 4/\lambda\} \log_+(2\Phi_0/\epsilon)$. Using $4/\lambda = 4N \leq 768\nu^2/(\mu\epsilon^2)$ and substituting the full learning-rate choice, we select:

$$K = \left\lceil \max \left\{ \frac{128L}{\mu\epsilon}, \frac{768\nu^2}{\mu\epsilon^2}, \frac{1536\sqrt{2}\nu\sqrt{C_{\text{DEF},\omega}}}{\mu^{3/2}\sqrt{B}\epsilon^2} \right\} \log_+ \frac{2\Phi_0}{\epsilon} \right\rceil.$$

Since an iteration draws N trajectories with probability $p = 1/N$ and B trajectories otherwise, the expected per-iteration cost is $1 + (1 - 1/N)B$. Consequently:

$$n = N_0 + K \left(1 + B - \frac{B}{N} \right).$$

Substituting the selected values of N_0 and K , and using $1 + B - B/N \leq 1 + B$, yields:

$$n = \tilde{\mathcal{O}} \left(1 + B + \frac{(1+B)L}{\mu\epsilon} + \frac{(1+B)\nu^2}{\mu\epsilon^2} + \frac{(1+B)\nu\sqrt{C_{\text{DEF},\omega}}}{\mu^{3/2}\sqrt{B}\epsilon^2} \right).$$

For fixed B and fixed positive problem-dependent constants, the terms proportional to ϵ^{-2} dominate as $\epsilon \rightarrow 0$. Substituting $\nu^2 = \mathcal{O}(R_{\max}^2(1-\gamma)^{-3})$, $C_{\text{DEF},\omega} = \mathcal{O}(R_{\max}^2(1-\gamma)^{-5})$, and $\mu = \chi^{-2}$ gives:

$$n = \tilde{\mathcal{O}} \left(\frac{R_{\max}^2}{\epsilon^2} \left[\frac{\chi^2}{(1-\gamma)^3} + \frac{\chi^3}{(1-\gamma)^4} \right] \right),$$

where B , the policy-regularity constants, and the defensive-mixture parameter are treated as fixed. $\square$

Corollary E.2 (Sample Complexity of DEF-STORM-PG). *Under Assumptions 3.1–3.2, for a fixed mini-batch size $B \geq 1$ and sufficiently small target accuracy $\epsilon > 0$, through an appropriate choice of the algorithmic parameters (momentum α , initial batch size N_0 , and learning rate η), DEF-STORM-PG with $p = 0$ and defensive-mixture parameter $\xi = 1/2$ guarantees $J^* - \mathbb{E}[J(\theta_k)] \leq \psi + \epsilon$ with a sample complexity of at most:*

$$\begin{aligned} n &= N_0 + KB \\ &= \tilde{\mathcal{O}} \left(\frac{R_{\max}^2}{\epsilon^2} \left[\frac{\chi^2}{(1-\gamma)^3} + \frac{\chi^3}{(1-\gamma)^4} \right] \right). \end{aligned}$$

Here, $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors, and B and the policy-regularity constants are treated as fixed.

Proof. Setting the refresh probability to $p = 0$ specializes the generic variance recurrence used in the proof of Theorem 5.1 to the DEF-STORM-PG estimator. Under the defensive-mixture choice $\xi = 1/2$, the decay rate λ and the corrected noise floor C_ν simplify to:

$$\lambda = 2\alpha - \alpha^2, \quad C_\nu = \frac{4\alpha^2\nu^2}{B}.$$

Since $\lambda \geq \alpha$ for $\alpha \in [0, 1]$, a sufficient condition for $C_\nu \leq \mu\lambda\epsilon^2/96$ is:

$$\frac{4\alpha^2\nu^2}{B} \leq \frac{\mu\alpha\epsilon^2}{96} \implies \alpha \leq \frac{\mu B\epsilon^2}{384\nu^2}.$$

We therefore select $\alpha = \mu B\epsilon^2/(384\nu^2)$ and $N_0 = \lceil 96\nu^2/(\mu\epsilon^2) \rceil$. For sufficiently small ϵ , this choice satisfies $\alpha \in (0, 1)$. For the variance stability requirement $\eta \leq \sqrt{\lambda B/[24(1-\lambda)C_{\text{DEF},\omega}]}$, using $\lambda \geq \alpha$ and $1 - \lambda \leq 1$, we enforce:

$$\eta \leq \sqrt{\frac{\alpha B}{24C_{\text{DEF},\omega}}} = \sqrt{\frac{\mu B^2}{9216\nu^2 C_{\text{DEF},\omega}}} \epsilon.$$

For the last-iterate accuracy requirement, $\lambda \geq \alpha$ gives:

$$\frac{\lambda^2\epsilon}{12C_\nu} \geq \frac{\alpha^2\epsilon}{12(4\alpha^2\nu^2/B)} = \frac{B\epsilon}{48\nu^2}.$$

Consequently, defining

$$c_\eta := \min \left\{ \frac{B}{48\nu^2}, \sqrt{\frac{\mu B^2}{9216\nu^2 C_{\text{DEF},\omega}}} \right\},$$

we select $\eta = \min\{1/(4L), c_\eta\epsilon\}$. Using the sufficient iteration budget established in the proof of Theorem 5.1, together with $4/\lambda \leq 4/\alpha = 1536\nu^2/(\mu B\epsilon^2)$, we select:

$$K = \left\lceil \max \left\{ \frac{128L}{\mu\epsilon}, \frac{1536\nu^2}{\mu B\epsilon^2}, \frac{3072\nu\sqrt{C_{\text{DEF},\omega}}}{\mu^{3/2}B\epsilon^2} \right\} \log_+ \frac{2\Phi_0}{\epsilon} \right\rceil.$$

Here, the terms proportional to ϵ^{-1} and ϵ^{-2} follow by substituting the full learning-rate choice into $32/(\mu\eta\epsilon)$. Because $p = 0$, the algorithm draws a fixed mini-batch of size B at every iteration. Consequently, the total sample complexity is:

$$n = N_0 + KB.$$

Substituting the selected values of N_0 and K gives:

$$n = \tilde{\mathcal{O}} \left(1 + B + \frac{BL}{\mu\epsilon} + \frac{\nu^2}{\mu\epsilon^2} + \frac{\nu\sqrt{C_{\text{DEF},\omega}}}{\mu^{3/2}\epsilon^2} \right).$$

For fixed B and fixed positive problem-dependent constants, the terms proportional to ϵ^{-2} dominate as $\epsilon \rightarrow 0$. Substituting $\nu^2 = \mathcal{O}(R_{\max}^2(1-\gamma)^{-3})$, $C_{\text{DEF},\omega} = \mathcal{O}(R_{\max}^2(1-\gamma)^{-5})$, and $\mu = \chi^{-2}$ yields:

$$n = \tilde{\mathcal{O}} \left(\frac{R_{\max}^2}{\epsilon^2} \left[\frac{\chi^2}{(1-\gamma)^3} + \frac{\chi^3}{(1-\gamma)^4} \right] \right),$$

where the policy-regularity constants are treated as fixed. $\square$

F SAMPLE COMPLEXITY UNDER FISHER-NON-DEGENERACY

While Theorem 5.1 establishes the sample complexity bounds under the abstract weak gradient domination condition, we apply Assumptions C.2 and C.1 to apply Lemma C.1. Note that, in the FND setting, we consider the unconstrained parameter space $\Theta = \mathbb{R}^d$.

Corollary 5.1 (Sample complexity under FND). *In the unconstrained setting where $\Theta = \mathbb{R}^d$, letting $\chi = C_1/\mu_F$ and $\psi = \sqrt{\epsilon_{\text{bias}}}\kappa/(1-\gamma)$, assuming $\kappa < \infty$, **LVR-PG** achieves a suboptimality gap $J^* - \mathbb{E}[J(\theta_K)] \leq \frac{\sqrt{\epsilon_{\text{bias}}}\kappa}{1-\gamma} + \epsilon$ with a total sample complexity of:*

$$n = \tilde{\mathcal{O}} \left(\frac{R_{\max}^2 C_1^3}{(1-\gamma)^4 \mu_F^3 \epsilon^2} \right).$$

Proof. The result follows directly from Theorem 5.1 by applying the Fisher-induced Gradient Domination Lemma. Specifically, the lemma establishes that under Assumptions C.2 and C.1, the Weak Gradient Domination condition holds with the constants $\chi = C_1/\mu_F$ and $\psi = \sqrt{\epsilon_{\text{bias}}}\kappa/(1-\gamma)$. Substituting these specific values into the optimality gap and the general sample complexity bound of Theorem 5.1 immediately yields the stated result. $\square$

G LOWER-BOUND

Theorem 3.1 (Sample Complexity Lower Bound). *For any discount factor $\gamma \in [2/3, 1)$ and sufficiently small target accuracy $0 < \epsilon < \mathcal{O}\left(\frac{R_{\max}}{\gamma(1-\gamma)}\right)$, there exists a class of MDPs $\mathcal{M}$ and a policy parametrization satisfying Assumptions 3.1 and 3.2, such that any algorithm guaranteeing an expected sub-optimality gap $J^* - \mathbb{E}[J(\theta_K)] \leq \epsilon$ must collect a number of trajectories n satisfying:*

$$n = \Omega \left(\frac{R_{\max}^2}{(1-\gamma)^2} \epsilon^{-2} \right).$$

Before starting with the proof of Theorem 3.1, we describe it and its difference from previous works. The proof relies on the information-theoretic framework for minimax lower bounds and uses Le Cam's two-point

method. While the hard instance construction shares topological similarities with the absorbing MDPs introduced by Azar et al. (2013), we accommodate the specific geometry of policy optimization. While Azar et al. (2013) establishes lower bounds for unconstrained tabular reinforcement learning, in our proof, we consider a constrained continuous policy parametrization, $\theta \in [\alpha, 1 - \alpha]$. This domain restriction is the critical mechanism that enforces Assumption 3.1. Furthermore, we demonstrate that this constrained geometry satisfies Assumption 3.2 with $\chi = 4/3$ and $\psi = 0$, avoiding vanishing gradients at the optimal boundaries.

To establish a fundamental lower bound on the sample complexity of *any* learning algorithm, we must construct an adversarial scenario in which making the correct decision is statistically difficult, yet an incorrect guess incurs a penalty. We achieve this by designing two nearly identical MDPs, $\mathcal{M}_1$ and $\mathcal{M}_2$. These instances differ strictly by a small perturbation δ in the transition probabilities, making them difficult to distinguish, yet they require mutually exclusive optimal policies. We first quantify the sub-optimality gap to show that if an algorithm $\mathcal{A}$ fails to identify the true underlying MDP and selects a parameter favouring the wrong action, it suffers a performance penalty of at least $\Omega(\delta(1 - \gamma)^{-1})$. By tuning this "wrong-action" penalty to perfectly match our target accuracy threshold ϵ , we isolate the required perturbation scale to $\delta = \mathcal{O}(\epsilon(1 - \gamma)^2)$.

Because a valid $\mathcal{A}$ must guarantee ϵ -optimality across *all* instances, it is forced to determine whether it is interacting with $\mathcal{M}_1$ or $\mathcal{M}_2$. If it collects too few samples, the trajectory distributions of the two MDPs remain virtually identical, and the algorithm will inevitably guess wrong with high probability, thereby suffering the tuned ϵ penalty. We formalize this through Le Cam's method, which implies that, to reliably distinguish the two instances, the total Kullback-Leibler divergence between their n -trajectory distributions must exceed a given threshold.

Finally, we establish how much data is required to cross this threshold by analytically upper-bounding the KL divergence. We show that the divergence accumulates to at most $\mathcal{O}(n\delta^2(1 - \gamma)^{-2})$. Substituting our chosen δ into this upper bound and solving for the sample size n reveals that the minimum amount of data required to distinguish the instances, and thus avoid the sub-optimality penalty, is exactly $\Omega((1 - \gamma)^{-2}R_{\max}^2\epsilon^{-2})$.

Proof. The proof is subdivided into several parts.

Part I: Construction of the Hard Instance. We consider a class of MDPs consisting of two instances, $\mathcal{M} \in \{\mathcal{M}_1, \mathcal{M}_2\}$, with state space $\mathcal{S} = \{s, \bar{s}\}$ and action space $\mathcal{A} = \{a_0, a_1\}$. The state $\bar{s}$ is an absorbing state where any action $a \in \mathcal{A}$ yields a deterministic self-transition with zero reward, meaning $P(\bar{s}|\bar{s}, a) = 1$ and $R(\bar{s}, a, \bar{s}) = 0$. In the initial state s , taking an action a triggers a self-transition to s with probability p_a , yielding a reward of $R_{\max}$, and transitions to the absorbing state $\bar{s}$ with probability $1 - p_a$, yielding $R_{\max}$. The two instances differ strictly in the transition probabilities from the initial state s . For the instance $\mathcal{M}_1$, we set $p_{a_0} = p + \delta$ and $p_{a_1} = p$, making a_0 the optimal action. Conversely, for the instance $\mathcal{M}_2$, we set $p_{a_0} = p$ and $p_{a_1} = p + \delta$, making a_1 the optimal action. To ensure the transition probabilities remain valid while capturing the lower-bound complexity dependence on the effective horizon, we define the base transition probability as $p = 1 - \frac{1-\gamma}{\gamma}$. Consequently, to satisfy the requirement that $p + \delta \leq 1$, the perturbation parameter δ is strictly bounded by $\delta \leq \frac{1-\gamma}{2\gamma}$.

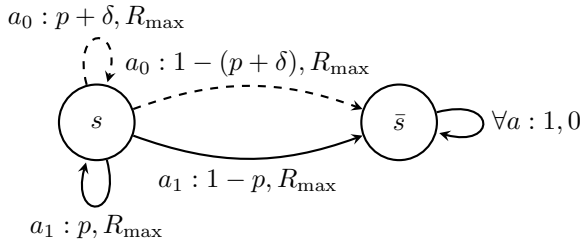


Figure 3: Transition dynamics for instance $\mathcal{M}_1$.

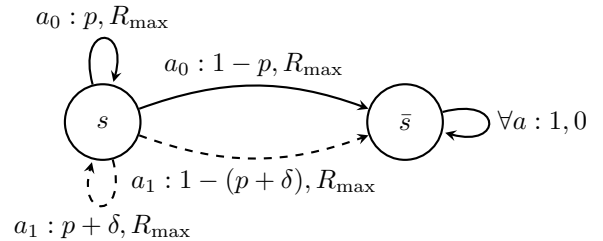


Figure 4: Transition dynamics for instance $\mathcal{M}_2$.

We consider a direct policy parametrization in which the scalar parameter θ determines the probability of

selecting the action a_0 in the initial state s . For the binary action space $\mathcal{A} = \{a_0, a_1\}$, the policy is defined as:

$$\pi_\theta(a \mid s) = \begin{cases} \theta & \text{if } a = a_0, \\ 1 - \theta & \text{otherwise.} \end{cases}$$

Part II: Verification of the Assumptions. Here, we check that the policy class we have chosen satisfies Assumptions 3.1 and 3.2. To ensure the policy space satisfies the regularity constraints of Assumption 3.1, we restrict the parameter domain to $\Theta = [\alpha, 1 - \alpha]$ for a strictly positive constant $\alpha \in (0, 1/2)$. The score function is given by:

$$\nabla_\theta \log \pi_\theta(a \mid s) = \begin{cases} \frac{1}{\theta} & \text{if } a = a_0, \\ -\frac{1}{1-\theta} & \text{otherwise.} \end{cases}$$

Condition (i) requires the score norm to be bounded. Since $\theta \in [\alpha, 1 - \alpha]$, the probability of any action is bounded strictly away from zero, uniformly bounding the score. We establish the constant C_1 as the supremum of the score function over the parameter space and action space:

$$\sup_{a \in \mathcal{A}} \|\nabla_\theta \log \pi_\theta(a \mid s)\|_2 = \max_{\theta \in [\alpha, 1-\alpha]} \max \left\{ \frac{1}{\theta}, \frac{1}{1-\theta} \right\} = \frac{1}{\alpha}.$$

Thus, condition (i) is satisfied with $C_1 \leq 1/\alpha$. To verify condition (ii), we must establish that the score function is Lipschitz continuous in expectation. We first calculate the Hessian of the log-policy:

$$\nabla_\theta^2 \log \pi_\theta(a \mid s) = \begin{cases} -\frac{1}{\theta^2} & \text{if } a = a_0, \\ -\frac{1}{(1-\theta)^2} & \text{otherwise.} \end{cases}$$

The restricted domain Θ guarantees that the Hessian norm is uniformly bounded by $1/\alpha^2$. By the Mean Value Theorem, the squared difference in the score function for any two parameters $\theta, \theta' \in \Theta$ is bounded by the squared supremum of the Hessian. For any action $a \in \mathcal{A}$:

$$\begin{aligned} \|\nabla_\theta \log \pi_\theta(a \mid s) - \nabla_{\theta'} \log \pi_{\theta'}(a \mid s)\|_2^2 &\leq \left(\sup_{\tilde{\theta} \in \Theta} \|\nabla_{\tilde{\theta}}^2 \log \pi_{\tilde{\theta}}(a \mid s)\|_2 \right)^2 \|\theta - \theta'\|_2^2 \\ &\leq \left(\frac{1}{\alpha^2} \right)^2 \|\theta - \theta'\|_2^2 \\ &= \frac{1}{\alpha^4} \|\theta - \theta'\|_2^2. \end{aligned}$$

Because this upper bound holds deterministically for all actions $a \in \mathcal{A}$, taking the expectation over $a \sim \pi_\theta(\cdot \mid s)$ preserves the inequality. Thus, condition (ii) is satisfied with the constant $L_S = 1/\alpha^2$:

$$\mathbb{E}_{a \sim \pi_\theta} \left[\|\nabla_\theta \log \pi_\theta(a \mid s) - \nabla_{\theta'} \log \pi_{\theta'}(a \mid s)\|_2^2 \right] \leq \frac{1}{\alpha^4} \|\theta - \theta'\|_2^2.$$

We now verify the validity of Assumption 3.2 with $\psi = 0$ in our setting. We start computing the expected return $J^{(i)}(\theta)$, its optimal value $J^{*,(i)}$ within the restricted domain $\Theta = [\alpha, 1 - \alpha]$, and the gradient norm $\|\nabla_\theta J^{(i)}(\theta)\|_2$ for both instances $i \in \{1, 2\}$. By assigning a reward of $R_{\max}$ to any action taken in state s prior to reaching the zero-reward absorbing state $\bar{s}$, the expected immediate reward at time t , given $s_t = s$, is exactly $R_{\max}$. The expected discounted return resolves to a standard geometric series:

$$J(\theta) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid s_0 = s) R_{\max} = R_{\max} \sum_{t=0}^{\infty} (\gamma p_\theta)^t = \frac{R_{\max}}{1 - \gamma p_\theta}.$$

For the specific instances $\mathcal{M}_1$ and $\mathcal{M}_2$, the effective self-transition probabilities under the policy π_θ are:

$$\begin{aligned} p_\theta^{(1)} &= \theta(p + \delta) + (1 - \theta)p = p + \theta\delta, \\ p_\theta^{(2)} &= \theta p + (1 - \theta)(p + \delta) = p + (1 - \theta)\delta. \end{aligned}$$

Table 3: Values for $J^{(i)}(\theta)$, $J^{*,(i)}$, and $\|\nabla_\theta J^{(i)}(\theta)\|_2$ for the $\mathcal{M}$ instances.

Instance	$J^{(i)}(\theta)$	$J^{*,(i)}$	$\ \nabla_\theta J^{(i)}(\theta)\ _2$
$\mathcal{M}_1$	$\frac{R_{\max}}{1-\gamma(p+\theta\delta)}$	$\frac{R_{\max}}{1-\gamma(p+(1-\alpha)\delta)}$	$\frac{\gamma\delta R_{\max}}{(1-\gamma(p+\theta\delta))^2}$
$\mathcal{M}_2$	$\frac{R_{\max}}{1-\gamma(p+(1-\theta)\delta)}$	$\frac{R_{\max}}{1-\gamma(p+(1-\alpha)\delta)}$	$\frac{\gamma\delta R_{\max}}{(1-\gamma(p+(1-\theta)\delta))^2}$

Because the expected return monotonically increases with the self-transition probability, the optimal returns $J^{*,(i)}$ over the constrained domain $\Theta = [\alpha, 1-\alpha]$ are achieved by maximizing $p_\theta^{(i)}$. For $\mathcal{M}_1$, this occurs at $\theta = 1-\alpha$, and for $\mathcal{M}_2$, this occurs at $\theta = \alpha$. In both cases, the maximum self-transition probability evaluates to $p+(1-\alpha)\delta$. Differentiating the objective function with respect to θ cleanly isolates the gradient of the transition probability:

$$\nabla_\theta J^{(i)}(\theta) = \frac{\gamma R_{\max}}{(1-\gamma p_\theta^{(i)})^2} \nabla_\theta p_\theta^{(i)}.$$

Applying this to both instances yields the values defined in Table 3. We establish the constant χ by first evaluating the ratio with respect to the unconstrained gradient, $\frac{J^{*,(i)} - J^{(i)}(\theta)}{\|\nabla_\theta J^{(i)}(\theta)\|_2} \leq \chi$. As we will formalize in Remark G.1, bounding this unconstrained ratio is sufficient to satisfy Assumption 3.2 for the generalized mapping $\mathcal{G}_\eta$. Starting with $\mathcal{M}_1$, the sub-optimality gap reduces to:

$$\begin{aligned} J^{*,(1)} - J^{(1)}(\theta) &= \frac{R_{\max}}{1-\gamma(p+(1-\alpha)\delta)} - \frac{R_{\max}}{1-\gamma(p+\theta\delta)} \\ &= \frac{\gamma R_{\max}((p+(1-\alpha)\delta) - (p+\theta\delta))}{(1-\gamma(p+(1-\alpha)\delta))(1-\gamma(p+\theta\delta))} \\ &= \frac{\gamma R_{\max}(1-\alpha-\theta)\delta}{(1-\gamma(p+(1-\alpha)\delta))(1-\gamma(p+\theta\delta))}. \end{aligned}$$

Dividing this gap by the gradient norm causes both γ and $R_{\max}$ to cancel out:

$$\begin{aligned} \frac{J^{*,(1)} - J^{(1)}(\theta)}{\|\nabla_\theta J^{(1)}(\theta)\|_2} &= \frac{\gamma R_{\max}(1-\alpha-\theta)\delta}{(1-\gamma(p+(1-\alpha)\delta))(1-\gamma(p+\theta\delta))} \cdot \frac{(1-\gamma(p+\theta\delta))^2}{\gamma\delta R_{\max}} \\ &= (1-\alpha-\theta) \frac{1-\gamma(p+\theta\delta)}{1-\gamma(p+(1-\alpha)\delta)}. \end{aligned}$$

Applying the definitions $p = \frac{2\gamma-1}{\gamma}$ (requiring $\gamma > 1/2$) and $\delta \leq \frac{1-\gamma}{2\gamma}$, the baseline effective horizon evaluates to $1-\gamma p = 2(1-\gamma)$. We bound the denominator from below:

$$1-\gamma(p+(1-\alpha)\delta) \geq 1-\gamma p - \gamma\delta \geq 2(1-\gamma) - \frac{1-\gamma}{2} = \frac{3}{2}(1-\gamma).$$

Leveraging $\theta\delta > 0$, the numerator is bounded from above:

$$1-\gamma(p+\theta\delta) \leq 1-\gamma p = 2(1-\gamma).$$

Because the domain restriction ensures $(1-\alpha-\theta) \leq 1$, we obtain the uniform bound for $\mathcal{M}_1$:

$$\frac{J^{*,(1)} - J^{(1)}(\theta)}{\|\nabla_\theta J^{(1)}(\theta)\|_2} \leq 1 \cdot \frac{2(1-\gamma)}{\frac{3}{2}(1-\gamma)} = \frac{4}{3}.$$

We repeat the exact procedure for $\mathcal{M}_2$. The optimal parameter for $\mathcal{M}_2$ is $\theta^* = \alpha$. The sub-optimality gap is:

$$\begin{aligned} J^{*,(2)} - J^{(2)}(\theta) &= \frac{R_{\max}}{1 - \gamma(p + (1 - \alpha)\delta)} - \frac{R_{\max}}{1 - \gamma(p + (1 - \theta)\delta)} \\ &= \frac{\gamma R_{\max}((p + (1 - \alpha)\delta) - (p + (1 - \theta)\delta))}{(1 - \gamma(p + (1 - \alpha)\delta))(1 - \gamma(p + (1 - \theta)\delta))} \\ &= \frac{\gamma R_{\max}(\theta - \alpha)\delta}{(1 - \gamma(p + (1 - \alpha)\delta))(1 - \gamma(p + (1 - \theta)\delta))}. \end{aligned}$$

Dividing by the gradient norm for $\mathcal{M}_2$ yields:

$$\frac{J^{*,(2)} - J^{(2)}(\theta)}{\|\nabla_{\theta} J^{(2)}(\theta)\|_2} = (\theta - \alpha) \frac{1 - \gamma(p + (1 - \theta)\delta)}{1 - \gamma(p + (1 - \alpha)\delta)}.$$

The denominator bound remains identical to $\mathcal{M}_1$. For the numerator, since $(1 - \theta)\delta > 0$:

$$1 - \gamma(p + (1 - \theta)\delta) \leq 1 - \gamma p = 2(1 - \gamma).$$

The parameter restriction $\theta \in [\alpha, 1 - \alpha]$ enforces $(\theta - \alpha) \leq 1 - 2\alpha < 1$. Substituting these bounds yields the limit for $\mathcal{M}_2$:

$$\frac{J^{*,(2)} - J^{(2)}(\theta)}{\|\nabla_{\theta} J^{(2)}(\theta)\|_2} \leq (1 - 2\alpha) \cdot \frac{2(1 - \gamma)}{\frac{3}{2}(1 - \gamma)} < \frac{4}{3}.$$

The preceding computations establish

$$J^{*,(i)} - J^{(i)}(\theta) \leq \frac{4}{3} \|\nabla_{\theta} J^{(i)}(\theta)\|_2, \quad i \in \{1, 2\}.$$

We now verify the corresponding inequality for the generalized gradient mapping.

Remark G.1 (Boundary Conditions and the Gradient Mapping). *Because the policy space is restricted to $\Theta = [\alpha, 1 - \alpha]$, we must verify that Assumption 3.2 also holds when projection clips the update. The optimal parameters are $\theta^{*,(1)} = 1 - \alpha$ and $\theta^{*,(2)} = \alpha$. For both instances, the gradient points toward the corresponding optimum, so projection gives*

$$\|G_{\eta}^{(i)}(\theta)\|_2 = \min \left\{ \|\nabla_{\theta} J^{(i)}(\theta)\|_2, \frac{|\theta^{*,(i)} - \theta|}{\eta} \right\}.$$

If the update is not clipped, the generalized mapping equals the gradient, and the previously established inequality $J^{,(i)} - J^{(i)}(\theta) \leq \frac{4}{3} \|\nabla_{\theta} J^{(i)}(\theta)\|_2$ applies directly. Otherwise, the projected update reaches $\theta^{*,(i)}$, yielding $|\theta^{*,(i)} - \theta| = \eta \|G_{\eta}^{(i)}(\theta)\|_2$. By the mean value theorem,*

$$\begin{aligned} J^{*,(i)} - J^{(i)}(\theta) &\leq \sup_{\theta' \in \Theta} \|\nabla_{\theta} J^{(i)}(\theta')\|_2 |\theta^{*,(i)} - \theta| \\ &= \frac{\eta \gamma \delta R_{\max}}{[1 - \gamma(p + (1 - \alpha)\delta)]^2} \|G_{\eta}^{(i)}(\theta)\|_2. \end{aligned}$$

The regularity constants established above, $C_1 = 1/\alpha$ and $L_S = 1/\alpha^2$, give $L = 2R_{\max}/[\alpha^2(1 - \gamma)^2]$. Using $\eta \leq 1/L$, $1 - \gamma(p + (1 - \alpha)\delta) \geq \frac{3}{2}(1 - \gamma)$, and $\gamma\delta \leq (1 - \gamma)/2$, the coefficient in the last display is at most $\alpha^2(1 - \gamma)/9 \leq 4/3$. Thus, $J^{,(i)} - J^{(i)}(\theta) \leq \frac{4}{3} \|G_{\eta}^{(i)}(\theta)\|_2$ also holds when projection clips the update. At the optimum, both sides vanish. Consequently, Assumption 3.2 holds throughout Θ , for every $0 < \eta \leq 1/L$, with $\psi = 0$ and $\chi = 4/3$.*

Part III: Value Gap and Choice of Perturbation. To establish the statistical hardness, we must ensure that failing to identify the correct MDP results in a substantial performance penalty. Specifically, we require that if an algorithm $\mathcal{A}$ outputs a policy that favors the wrong action, the sub-optimality gap is at least 4ϵ .

In instance $\mathcal{M}_1$, the optimal parameter is $\theta^* = 1 - \alpha > 1/2$. The algorithm outputs a highly sub-optimal policy if it incorrectly favors action a_1 , which occurs when $\theta \leq 1/2$. We evaluate the sub-optimality gap for $\mathcal{M}_1$ given $\theta \leq 1/2$ to find a strict lower bound. From the previous derivation, the gap is:

$$J^{*,(1)} - J^{(1)}(\theta) = \frac{\gamma R_{\max}(1 - \alpha - \theta)\delta}{(1 - \gamma(p + (1 - \alpha)\delta))(1 - \gamma(p + \theta\delta))}.$$

To lower-bound this gap, we lower-bound the numerator and upper-bound the denominator. Because $\theta \leq 1/2$, the term $(1 - \alpha - \theta)$ is bounded below by $(1/2 - \alpha)$. Furthermore, since we established the condition $\gamma \geq 1/2$ to ensure valid transition probabilities, we substitute $\gamma = 1/2$ to strictly lower-bound the numerator and remove its dependence on γ :

$$\gamma R_{\max}(1 - \alpha - \theta)\delta \geq \frac{1}{2} R_{\max} \left(\frac{1}{2} - \alpha \right) \delta = \frac{R_{\max}(1 - 2\alpha)\delta}{4}.$$

For the denominator, since $\theta\delta \geq 0$ and $(1 - \alpha)\delta \geq 0$, the effective horizon terms are strictly upper-bounded by the baseline effective horizon $1 - \gamma p = 2(1 - \gamma)$:

$$(1 - \gamma(p + (1 - \alpha)\delta))(1 - \gamma(p + \theta\delta)) \leq (2(1 - \gamma))(2(1 - \gamma)) = 4(1 - \gamma)^2.$$

Combining these bounds, the sub-optimality gap for any parameter $\theta \leq 1/2$ is bounded from below:

$$J^{*,(1)} - J^{(1)}(\theta) \geq \frac{\frac{R_{\max}(1 - 2\alpha)\delta}{4}}{4(1 - \gamma)^2} = \frac{R_{\max}(1 - 2\alpha)\delta}{16(1 - \gamma)^2}.$$

To enforce that this misidentification penalty is at least 4ϵ , we equate this lower bound to 4ϵ and solve for the perturbation parameter δ :

$$\frac{R_{\max}(1 - 2\alpha)\delta}{16(1 - \gamma)^2} = 4\epsilon \implies \delta = \frac{64\epsilon(1 - \gamma)^2}{R_{\max}(1 - 2\alpha)}.$$

This specific choice of δ enforces the required value gap while controlling the statistical indistinguishability of the instances. To satisfy the transition probability constraint $\delta \leq \frac{1 - \gamma}{2\gamma}$, the target accuracy ϵ must be sufficiently small:

$$\frac{64\epsilon(1 - \gamma)^2}{R_{\max}(1 - 2\alpha)} \leq \frac{1 - \gamma}{2\gamma} \implies \epsilon \leq \frac{R_{\max}(1 - 2\alpha)}{128\gamma(1 - \gamma)}.$$

Part IV: Statistical Test and Information-Theoretic Lower Bound. We formulate the learning process as a statistical testing problem. Suppose an algorithm $\mathcal{A}$ interacts with an unknown instance $\mathcal{M}_i$, $i \in \{1, 2\}$, collects a possibly random number of trajectories $N_{\mathcal{A}}$, and subsequently outputs a parameter estimate $\hat{\theta} \in \Theta$. The algorithm is not informed whether the underlying environment is $\mathcal{M}_1$ or $\mathcal{M}_2$, and its policies may depend on all previously observed trajectories. For $\mathcal{A}$ to be considered a successful ϵ -optimal algorithm on these instances, its expected suboptimality gap must satisfy

$$\mathbb{E}_{\mathcal{A}, \mathcal{M}_i} [J^{*,(i)} - J^{(i)}(\hat{\theta})] \leq \epsilon, \quad i \in \{1, 2\},$$

where $J^{*,(i)} = \max_{\theta \in \Theta} J^{(i)}(\theta)$. We define the worst-case expected trajectory budget as

$$n := \max_{i \in \{1, 2\}} \mathbb{E}_{\mathcal{A}, \mathcal{M}_i} [N_{\mathcal{A}}].$$

We may assume that $n < \infty$, since otherwise the claimed lower bound holds trivially. Using the suboptimality bound derived in Part III, any output $\hat{\theta} < 1/2$ under $\mathcal{M}_1$ incurs a gap of at least 4ϵ . Therefore,

$$\begin{aligned} \epsilon &\geq \mathbb{E}_{\mathcal{A}, \mathcal{M}_1} [J^{*,(1)} - J^{(1)}(\hat{\theta})] \\ &\geq 4\epsilon \mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \left(\hat{\theta} < \frac{1}{2} \right), \end{aligned}$$

which implies:

$$\mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \left(\hat{\theta} < \frac{1}{2} \right) \leq \frac{1}{4}.$$

By symmetric reasoning under $\mathcal{M}_2$, any output $\hat{\theta} \geq 1/2$ incurs a gap of at least 4ϵ , yielding,

$$\mathbb{P}_{\mathcal{A}, \mathcal{M}_2} \left(\hat{\theta} \geq \frac{1}{2} \right) \leq \frac{1}{4}.$$

Summing these error probabilities gives:

$$\mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \left(\hat{\theta} < \frac{1}{2} \right) + \mathbb{P}_{\mathcal{A}, \mathcal{M}_2} \left(\hat{\theta} \geq \frac{1}{2} \right) \leq \frac{1}{2}.$$

Let $\mathbb{P}_{\mathcal{A}, \mathcal{M}_1}$ and $\mathbb{P}_{\mathcal{A}, \mathcal{M}_2}$ denote the probability measures over the complete stopped interaction transcript under the two instances. This transcript includes the algorithm's internal randomization, the observed trajectories, and its final output $\hat{\theta}$. Interpreting $\hat{\theta} \geq 1/2$ as selecting $\mathcal{M}_1$, Le Cam's inequality gives:

$$\begin{aligned} \mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \left(\hat{\theta} < \frac{1}{2} \right) + \mathbb{P}_{\mathcal{A}, \mathcal{M}_2} \left(\hat{\theta} \geq \frac{1}{2} \right) \\ \geq 1 - d_{\text{TV}}(\mathbb{P}_{\mathcal{A}, \mathcal{M}_1}, \mathbb{P}_{\mathcal{A}, \mathcal{M}_2}). \end{aligned}$$

Consequently, the total variation distance must be at least $1/2$. Applying Pinsker's inequality, we obtain,

$$\begin{aligned} \frac{1}{2} &\leq d_{\text{TV}}(\mathbb{P}_{\mathcal{A}, \mathcal{M}_1}, \mathbb{P}_{\mathcal{A}, \mathcal{M}_2}) \\ &\leq \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \parallel \mathbb{P}_{\mathcal{A}, \mathcal{M}_2})}, \end{aligned}$$

and hence:

$$D_{\text{KL}}(\mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \parallel \mathbb{P}_{\mathcal{A}, \mathcal{M}_2}) \geq \frac{1}{2}.$$

To translate this information requirement into a trajectory complexity lower bound, we next upper-bound the divergence in terms of the expected number of collected trajectories.

Part V: Upper Bound on the KL Divergence. We upper-bound the KL divergence of the complete interaction transcript through the conditional chain rule for relative entropy. Conditional on the same realized history, the algorithm uses the same action-selection and stopping rules under both instances. Moreover, the initial state, reward function, and independent geometric termination mechanism are identical under $\mathcal{M}_1$ and $\mathcal{M}_2$. Consequently, only the environment transition kernels contribute to the conditional KL divergence. We first bound the divergence contributed by a single transition. Since $\bar{s}$ is absorbing and possesses identical transition dynamics in both instances,

$$D_{\text{KL}}\left(P^{(1)}(\cdot \mid \bar{s}, a) \parallel P^{(2)}(\cdot \mid \bar{s}, a)\right) = 0.$$

At the active state s , the Bernoulli parameters under $(\mathcal{M}_1, \mathcal{M}_2)$ are $(p + \delta, p)$ for a_0 and $(p, p + \delta)$ for a_1 . Using the standard inequality $d_{\text{KL}}(x \parallel y) \leq (x - y)^2 / [y(1 - y)]$, we obtain,

$$\begin{aligned} d_{\text{KL}}(p + \delta \parallel p) &\leq \frac{\delta^2}{p(1 - p)}, \\ d_{\text{KL}}(p \parallel p + \delta) &\leq \frac{\delta^2}{(p + \delta)(1 - p - \delta)}. \end{aligned}$$

Recall that $p = (2\gamma - 1)/\gamma$, $\gamma \geq 2/3$, and $\delta \leq (1 - \gamma)/(2\gamma)$. These conditions imply:

$$p \geq \frac{1}{2}, \quad 1 - p = \frac{1 - \gamma}{\gamma}, \quad 1 - p - \delta \geq \frac{1 - \gamma}{2\gamma}.$$

Thus, both variance terms are bounded below by $(1 - \gamma)/(4\gamma)$, yielding,

$$\max_{a \in \{a_0, a_1\}} D_{\text{KL}}\left(P^{(1)}(\cdot \mid s, a) \parallel P^{(2)}(\cdot \mid s, a)\right) \leq \frac{4\gamma\delta^2}{1 - \gamma} \leq \frac{4\delta^2}{1 - \gamma}.$$

This bound holds uniformly over the selected action. Next, we bound the number of informative transitions within each trajectory. Let H_j denote the number of observed transitions in trajectory j , and define:

$$T_j := \sum_{t=0}^{H_j-1} \mathbf{1}_{\{s_{j,t}=s\}},$$

where $s_{j,t}$ and $a_{j,t}$ denote the state and action at time t of trajectory j . Thus, T_j counts the observed transitions from the active state. Each trajectory starts at s and follows the independent geometric termination protocol. We set $T_j = 0$ when trajectory j is not collected. Let $\mathcal{F}_{j-1}$ denote the history immediately before trajectory j , including the algorithm's decision whether to collect it. In particular, $\{N_{\mathcal{A}} \geq j\}$ is $\mathcal{F}_{j-1}$ -measurable. On this event, the conditional probability of remaining in s after any active transition is at most $p + \delta$, regardless of the selected action. The geometric termination mechanism can only reduce the number of observed active transitions. Consequently,

$$\begin{aligned} \mathbb{P}_{\mathcal{A}, \mathcal{M}_1}(T_j > t \mid \mathcal{F}_{j-1}) &\leq (p + \delta)^t, \quad t \geq 0, \\ \mathbb{E}_{\mathcal{A}, \mathcal{M}_1}[T_j \mid \mathcal{F}_{j-1}] &\leq \sum_{t=0}^{\infty} (p + \delta)^t \\ &= \frac{1}{1 - p - \delta} \leq \frac{2\gamma}{1 - \gamma} \leq \frac{2}{1 - \gamma}. \end{aligned}$$

The conditional expectation bound also holds when the trajectory is not collected, since then $T_j = 0$. For a history at which trajectory j is collected, let $Q_{i,j}(\cdot \mid \mathcal{F}_{j-1})$ denote its conditional law under $\mathcal{M}_i$, and define:

$$K_j := D_{\text{KL}}(Q_{1,j}(\cdot \mid \mathcal{F}_{j-1}) \parallel Q_{2,j}(\cdot \mid \mathcal{F}_{j-1})).$$

We set $K_j = 0$ at histories where trajectory j is not collected. Applying the conditional chain rule within a collected trajectory gives:

$$\begin{aligned} K_j &= \mathbb{E}_{\mathcal{A}, \mathcal{M}_1} \left[\sum_{t=0}^{H_j-1} D_{\text{KL}}\left(P^{(1)}(\cdot \mid s_{j,t}, a_{j,t}) \parallel P^{(2)}(\cdot \mid s_{j,t}, a_{j,t})\right) \middle| \mathcal{F}_{j-1} \right] \\ &\leq \frac{4\delta^2}{1 - \gamma} \mathbb{E}_{\mathcal{A}, \mathcal{M}_1}[T_j \mid \mathcal{F}_{j-1}] \\ &\leq \frac{8\delta^2}{(1 - \gamma)^2}. \end{aligned}$$

Therefore, the same upper bound applies to every collected trajectory, uniformly over the preceding history and the algorithm's policy choices. Finally, applying the conditional chain rule across trajectories yields:

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \parallel \mathbb{P}_{\mathcal{A}, \mathcal{M}_2}) &= \sum_{j=1}^{\infty} \mathbb{E}_{\mathcal{A}, \mathcal{M}_1}[\mathbf{1}_{\{N_{\mathcal{A}} \geq j\}} K_j] \\ &\leq \frac{8\delta^2}{(1 - \gamma)^2} \sum_{j=1}^{\infty} \mathbb{P}_{\mathcal{A}, \mathcal{M}_1}(N_{\mathcal{A}} \geq j) \\ &= \frac{8\delta^2}{(1 - \gamma)^2} \mathbb{E}_{\mathcal{A}, \mathcal{M}_1}[N_{\mathcal{A}}] \\ &\leq \frac{8n\delta^2}{(1 - \gamma)^2}. \end{aligned}$$

The chain-rule identities for random trajectory lengths and the complete interaction history follow by padding observations with zeros after termination, applying the chain rule to finite prefixes, and passing to the limit. The algorithm's internal randomization has the same distribution under both instances, while its action-selection, continuation, and output rules coincide conditional on the same history. Hence, only the environment transition dynamics contribute to the KL divergence.

Part VI: Sample Complexity Lower Bound. Combining the results from Parts IV and V, any algorithm achieving an expected suboptimality gap bounded by ϵ on both instances must satisfy,

$$\frac{1}{2} \leq D_{\text{KL}}(\mathbb{P}_{\mathcal{A}, \mathcal{M}_1} \parallel \mathbb{P}_{\mathcal{A}, \mathcal{M}_2}) \leq \frac{8n\delta^2}{(1-\gamma)^2}.$$

This yields the trajectory budget requirement:

$$n \geq \frac{(1-\gamma)^2}{16\delta^2}.$$

Substituting the perturbation parameter selected in Part III,

$$\delta = \frac{64\epsilon(1-\gamma)^2}{R_{\max}(1-2\alpha)},$$

we obtain:

$$\begin{aligned} n &\geq \frac{(1-\gamma)^2}{16} \frac{R_{\max}^2(1-2\alpha)^2}{4096\epsilon^2(1-\gamma)^4} \\ &= \frac{R_{\max}^2(1-2\alpha)^2}{65536\epsilon^2(1-\gamma)^2}. \end{aligned}$$

In particular, fixing $\alpha = 1/4$ gives:

$$\max_{i \in \{1,2\}} \mathbb{E}_{\mathcal{A}, \mathcal{M}_i}[N_{\mathcal{A}}] \geq \frac{R_{\max}^2}{262144\epsilon^2(1-\gamma)^2}.$$

Thus, for $\gamma \in [2/3, 1)$ and

$$0 < \epsilon \leq \frac{R_{\max}}{256\gamma(1-\gamma)},$$

any algorithm guaranteeing ϵ -optimality on both possible unknown instances requires a worst-case expected number of trajectories satisfying

$$n = \Omega\left(\frac{R_{\max}^2}{(1-\gamma)^2\epsilon^2}\right).$$

This establishes the claimed trajectory complexity lower bound for adaptive algorithms, including algorithms whose number of collected trajectories is random. $\square$

H TECHNICAL LEMMAS

Lemma H.1. *Let $\mathbf{G} : [0, 1] \rightarrow \mathbb{R}^d$ be continuous on $[0, 1]$. If, for any $u \in [0, 1)$,*

$$\limsup_{\epsilon \downarrow 0} \frac{\|\mathbf{G}(u + \epsilon) - \mathbf{G}(u)\|}{\epsilon} \leq C,$$

then $\|\mathbf{G}(1) - \mathbf{G}(0)\| \leq C$.

Proof. If $\mathbf{G}(1) = \mathbf{G}(0)$ there is nothing to prove, so assume $\|\mathbf{G}(1) - \mathbf{G}(0)\| > 0$. By contradiction, suppose $\|\mathbf{G}(1) - \mathbf{G}(0)\| > C$ and choose $c \in \mathbb{R}$ such that $\|\mathbf{G}(1) - \mathbf{G}(0)\| > c > C$. Let

$$\mathbf{v} := \frac{\mathbf{G}(1) - \mathbf{G}(0)}{\|\mathbf{G}(1) - \mathbf{G}(0)\|},$$

and define $f : [0, 1] \rightarrow \mathbb{R}$ as $f(u) := \mathbf{v}^\top \mathbf{G}(u) - cu$. Then, $f(1) - f(0) = \|\mathbf{G}(1) - \mathbf{G}(0)\| - c > 0$. Since f is continuous on $[0, 1]$, it attains its minimum at some $u^* \in [0, 1]$. Since $f(1) > f(0)$, $u^* < 1$. For sufficiently small ϵ , $f(u^* + \epsilon) - f(u^*) \geq 0$, so:

$$\limsup_{\epsilon \downarrow 0} \frac{f(u^* + \epsilon) - f(u^*)}{\epsilon} \geq 0.$$

On the other hand, since $\|\mathbf{v}\| = 1$,

$$\begin{aligned} \limsup_{\epsilon \downarrow 0} \frac{f(u^* + \epsilon) - f(u^*)}{\epsilon} &= \limsup_{\epsilon \downarrow 0} \frac{\mathbf{v}^\top (\mathbf{G}(u^* + \epsilon) - \mathbf{G}(u^*))}{\epsilon} - c \\ &\leq \limsup_{\epsilon \downarrow 0} \frac{\|\mathbf{G}(u^* + \epsilon) - \mathbf{G}(u^*)\|}{\epsilon} - c \\ &\leq C - c < 0. \end{aligned}$$

□

Lemma H.2 (Recurrence). *Let $\{x_k\}_{k=0}^\infty$ be a non-negative sequence satisfying the following recurrence relation for all $k \geq 0$:*

$$x_{k+1} \leq x_k - \min\{ax_k^2, bx_k\} + e,$$

where $a > 0$, $b > 0$, and $e > 0$. For a given target accuracy $\epsilon > 0$, define the effective contraction rate $\rho_\epsilon := \min\{a\epsilon, b\}$. Assume the algorithmic parameters are configured such that $\rho_\epsilon \in (0, 1)$. If the noise term e satisfies $e \leq a\frac{\epsilon^2}{2}$ and $e \leq b\frac{\epsilon}{2}$, then, it suffices to choose a number of iterations K satisfying:

$$K \geq \max\left\{\frac{1}{a\epsilon}, \frac{1}{b}\right\} \log_+ \left(\frac{2x_0}{\epsilon}\right),$$

where $\log_+(a) := \log \max\{1, a\}$, to ensure that $x_k \leq \epsilon$ for every $k \geq K$.

Proof. We first establish that the sequence remains within the target neighborhood once it reaches it. By the stated noise conditions, we have:

$$e \leq \min\left\{\frac{a\epsilon^2}{2}, \frac{b\epsilon}{2}\right\} = \frac{\rho_\epsilon \epsilon}{2}.$$

Suppose that $x_k \leq \epsilon$. If $x_k \leq \epsilon/2$, dropping the non-negative descent term yields:

$$x_{k+1} \leq x_k + e \leq \frac{\epsilon}{2} + \frac{\rho_\epsilon \epsilon}{2} \leq \epsilon,$$

where the final inequality follows from $\rho_\epsilon \in (0, 1)$. Conversely, if $\epsilon/2 < x_k \leq \epsilon$, we lower-bound the descent term as follows:

$$\begin{aligned} \min\{ax_k^2, bx_k\} &= x_k \min\{ax_k, b\} \\ &\geq x_k \min\left\{\frac{a\epsilon}{2}, b\right\} \\ &\geq \frac{\rho_\epsilon}{2} x_k. \end{aligned}$$

Substituting this bound into the recurrence gives:

$$\begin{aligned} x_{k+1} &\leq \left(1 - \frac{\rho_\epsilon}{2}\right) x_k + e \\ &\leq \left(1 - \frac{\rho_\epsilon}{2}\right) \epsilon + \frac{\rho_\epsilon \epsilon}{2} = \epsilon. \end{aligned}$$

Consequently, the target neighborhood $[0, \epsilon]$ is invariant: if $x_k \leq \epsilon$ at any iteration, then $x_j \leq \epsilon$ for every $j \geq k$. We now evaluate the sequence using a hitting-time argument. Let $T_\epsilon := \inf\{k \geq 0 \mid x_k \leq \epsilon\}$ denote the first iteration at which the sequence reaches the target accuracy, with $\inf \emptyset := +\infty$. If $x_0 \leq \epsilon$, the invariance

established above directly guarantees the claim for every iteration. Assume therefore that $x_0 > \epsilon$. By the definition of the hitting time, $x_k > \epsilon$ for every $k < T_\epsilon$. This allows us to lower-bound the descent rate prior to reaching the target neighborhood. For any $k < T_\epsilon$, factoring out x_k yields:

$$\min \{ax_k^2, bx_k\} = x_k \min \{ax_k, b\} \geq x_k \min \{a\epsilon, b\}.$$

Substituting the definition of the effective contraction rate $\rho_\epsilon := \min\{a\epsilon, b\}$ into the original inequality yields a standard linear recurrence:

$$x_{k+1} \leq (1 - \rho_\epsilon)x_k + e.$$

To set the hyperparameters so that the target $x_K \leq \epsilon$ is reached, we partition the error budget equally, requiring the asymptotic bias to be bounded by $\frac{\epsilon}{2}$ and the transient term to be bounded by $\frac{\epsilon}{2}$. Bounding the asymptotic bias requires $\frac{e}{\rho_\epsilon} \leq \frac{\epsilon}{2}$, which rearranges to $e \leq \rho_\epsilon \frac{\epsilon}{2}$. Because ρ_ϵ evaluates to the minimum of two regimes, this condition must hold for both possible values. If $\rho_\epsilon = a\epsilon$, the constraint dictates $e \leq a\frac{\epsilon^2}{2}$. If $\rho_\epsilon = b$, the constraint dictates $e \leq b\frac{\epsilon}{2}$. Satisfying both of these stated hyperparameter conditions ensures the asymptotic bias remains strictly bounded by $\frac{\epsilon}{2}$. Consequently, as long as $x_k \geq \epsilon > 0$, we can rewrite the linear recurrence as:

$$x_{k+1} - \frac{\epsilon}{2} \leq (1 - \rho_\epsilon) \left(x_k - \frac{\epsilon}{2} \right).$$

For any integer $K < T_\epsilon$, unrolling this recursive inequality over K steps yields:

$$\begin{aligned} x_K &\leq \frac{\epsilon}{2} + (1 - \rho_\epsilon)^K \left(x_0 - \frac{\epsilon}{2} \right) \\ &\leq \frac{\epsilon}{2} + \exp(-\rho_\epsilon K) x_0, \end{aligned}$$

where the final bound applies the standard logarithmic inequality $1 - x \leq \exp(-x)$, valid for $\rho_\epsilon \in (0, 1)$. Bounding the remaining transient term by $\epsilon/2$ requires:

$$\begin{aligned} K &\geq \frac{1}{\rho_\epsilon} \log_+ \left(\frac{2x_0}{\epsilon} \right) \\ &= \max \left\{ \frac{1}{a\epsilon}, \frac{1}{b} \right\} \log_+ \left(\frac{2x_0}{\epsilon} \right). \end{aligned}$$

Fix an integer K satisfying this bound. If $T_\epsilon > K$, the preceding estimate gives $x_K \leq \epsilon$, contradicting the definition of T_ϵ . Hence, $T_\epsilon \leq K$. By the invariance of the target neighborhood, it follows that $x_k \leq \epsilon$ for every $k \geq K$, completing the proof. $\square$

I EXPERIMENTAL DETAILS

We evaluate the empirical sample efficiency of the **LVR-PG** framework on CartPole-v1, Swimmer-v4, HalfCheetah-v4, and Hopper-v4. The experiments have two objectives: (i) to compare the framework’s configurations with GPOMDP (Baxter and Bartlett, 2001) and N-PG-IGT (Fatkhullin et al., 2023), and (ii) to examine sensitivity to the recursive-update parameter α and refresh probability p . We first describe the experimental protocol, then present the environment comparisons, the gradient-variability diagnostic, and the hyperparameter studies.

Policies and Implementation

CartPole-v1 uses a softmax policy with two hidden layers of width 4. The continuous-control tasks use Gaussian policies with two hidden layers of width 32. All hidden layers use tanh activations, with Xavier-uniform weight initialization and zero biases. Gaussian policies learn a scalar log standard deviation shared across action coordinates, initialized to the values in Table 5.

The experiments use fixed trajectory horizons. Return comparisons and hyperparameter studies use Adam, whereas the gradient-variability diagnostic uses a constant scalar step size. Training and evaluation use separate random-number streams. Evaluation returns are undiscounted and obtained from stochastic-policy rollouts; evaluation trajectories are excluded from the training sample count. For the environment comparisons, each evaluation averages 10 episodes, with evaluations every 10 collected training trajectories and at the final updated policy.

Table 4: Mean normalized area under the learning curve (AUC) across the four environments; the $\pm$ indicate the estimated standard error of the mean. Bold entries indicate the largest mean AUC in each environment.

Environment	GPOMDP	DEF-STORM-PG	DEF-PAGE-PG	LVR-PG	N-PG-IGT
Cart-Pole	396.05 $\pm$ 34.56	481.90 $\pm$ 7.85	417.34 $\pm$ 37.96	482.39 $\pm$ 5.49	380.03 $\pm$ 39.46
Swimmer	24.19 $\pm$ 6.15	40.86 $\pm$ 7.94	26.39 $\pm$ 5.46	33.90 $\pm$ 13.28	26.22 $\pm$ 1.01
Half-Cheetah	13.49 $\pm$ 30.16	105.76 $\pm$ 51.57	24.08 $\pm$ 32.67	55.82 $\pm$ 41.76	87.67 $\pm$ 23.31
Hopper	201.90 $\pm$ 21.68	369.94 $\pm$ 89.62	233.65 $\pm$ 28.36	309.13 $\pm$ 44.92	368.81 $\pm$ 23.97

Table 5: Environment-specific hyperparameter choices for the experiments.

Environment	Horizon T	Batch N	Mini-batch B	Learning rate η	Initial std σ
CartPole-v1	500	50	5	10^{-2}	–
Swimmer-v4	500	100	10	10^{-3}	1
HalfCheetah-v4	100	256	16	10^{-4}	$\sqrt{0.3}$
Hopper-v4	100	256	16	10^{-4}	0.3

Algorithms

We consider DEF-STORM-PG ($\alpha > 0$, $p = 0$), DEF-PAGE-PG ($\alpha = 0$, $p > 0$), and a hybrid configuration ($\alpha > 0$, $p > 0$). The first two are the defensive-IS versions of STORM-PG (Yuan et al., 2020) and PAGE-PG (Gargiani et al., 2022), respectively. In the comparison tables and figures, **LVR-PG** denotes the best hybrid configuration, hence the configuration of α and p that allow **LVR-PG** to obtain the maximum performance (please refer to Section I.7 to an example of ablation study). The parameter p determines the probability of refreshing the gradient estimator with N new trajectories. The parameter α controls the recursive momentum update; increasing α therefore reduces the weight of this correction. All three configurations use defensive IS with $\xi = 0.5$, as suggested by the theory.

We compare these configurations with GPOMDP (Baxter and Bartlett, 2001) and N-PG-IGT (Fatkhullin et al. (2023)). GPOMDP computes a fresh on-policy gradient estimate at each update. N-PG-IGT combines normalized updates with implicit gradient transport, using neither importance sampling nor second-order information. The algorithms use an average baseline for gradient estimation. Table 6 summarizes the algorithmic settings used in the environment comparisons.

Performance Measures

Learning curves report mean evaluation returns over five independent runs. Shaded regions indicate the estimated standard error of the mean. We summarize performance over the training budget using normalized Area under the Curve (AUC). Higher AUC indicates that high returns are reached earlier and maintained over a larger part of the budget. Table 4 reports the mean AUC and its confidence interval, showing that the momentum-only configuration of **LVR-PG** performs consistently better than the competitors in almost all the environments.

I.1 CartPole-v1

CartPole-v1 has a four-dimensional state space ($\mathcal{S} \subset \mathbb{R}^4$) and two discrete actions ($\mathcal{A} = \{0, 1\}$). Figure 5 compares the five algorithms using the settings in Tables 5 and 6.

DEF-STORM-PG and **LVR-PG** reach returns close to the maximum episodic return of 500 after approximately 10^3 trajectories. Both improve more rapidly than GPOMDP and DEF-PAGE-PG. This shows that, empirically, momentum behaves better than probabilistic refreshes. N-PG-IGT attains high returns but exhibits larger training variance.

Table 6: Algorithmic settings for the return comparisons. (*The hybrid uses $p = 0.5$ on CartPole-v1.) For N-PG-IGT, α denotes its gradient-transport momentum parameter.

Hyperparameter	GPOMDP	DEF-STORM-PG	DEF-PAGE-PG	LVR-PG	N-PG-IGT
Refresh probability p	—	0	0.5	0.1*	—
Recursive parameter α	—	0.9	0	0.9	0.9
Defensive parameter ξ	—	0.5	0.5	0.5	—

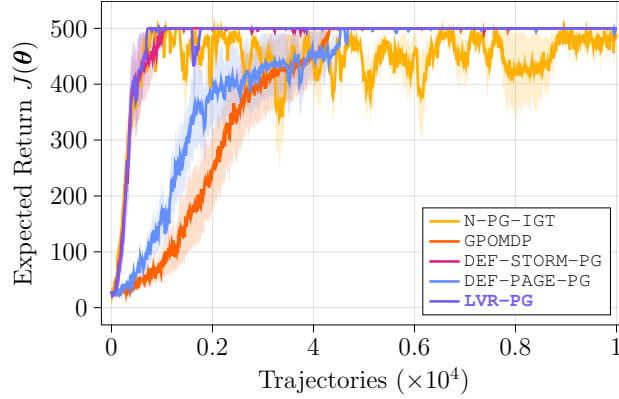


Figure 5: Learning curves on CartPole-v1. Curves show mean evaluation returns over five runs; shaded regions indicate the estimated standard error of the mean.

I.2 Swimmer-v4

Swimmer-v4 is a locomotion task with an eight-dimensional state space ($\mathcal{S} \subset \mathbb{R}^8$) and a two-dimensional action space ($\mathcal{A} \subset \mathbb{R}^2$). The policy controls an articulated swimmer moving through a fluid environment.

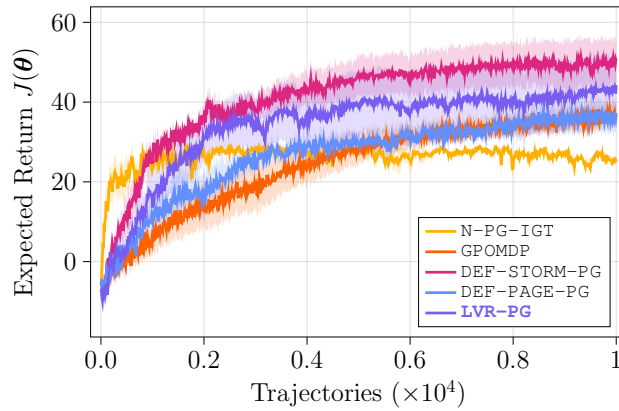


Figure 6: Learning curves on Swimmer-v4. Curves show mean evaluation returns over five runs; shaded regions indicate the estimated standard error of the mean.

Figure 6 shows that DEF-STORM-PG achieves the highest returns over the evaluated training budget. In this environment, the momentum-only configuration of **LVR-PG** provides the strongest empirical sample efficiency among the tested configurations. While learning faster in the initial part of the training, N-PG-IGT gets stuck in a suboptimal solution around a cumulative reward of 27.

I.3 HalfCheetah-v4

HalfCheetah-v4 has a 17-dimensional state space ($\mathcal{S} \subset \mathbb{R}^{17}$) and a six-dimensional action space ($\mathcal{A} \subset \mathbb{R}^6$). The policy controls the locomotion of an articulated agent.

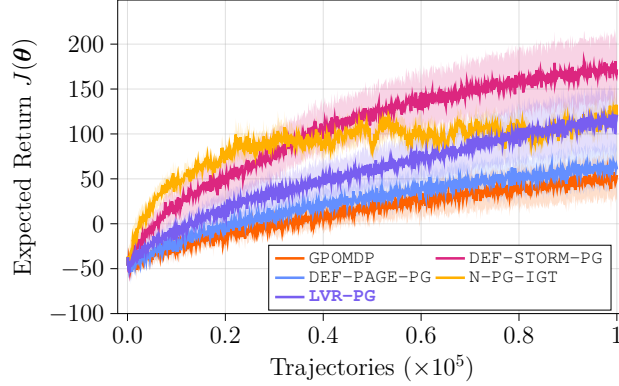


Figure 7: Learning curves on HalfCheetah-v4. Curves show mean evaluation returns over five runs; shaded regions indicate the estimated standard error of the mean.

In Figure 7, DEF-STORM-PG reaches the best return compared to the other algorithms. Again, N-PG-IGT learns faster in the first part of the learning, but early saturates to a lower cumulative reward compared to DEF-STORM-PG. The experiment also shows that DEF-PAGE-PG, and GPOMDP learn slower in the environment, and achieve a lower cumulative return after observing 10^5 trajectories.

I.4 Hopper-v4

Hopper-v4 is a locomotion task with an 11-dimensional state space ($\mathcal{S} \subset \mathbb{R}^{11}$) and a three-dimensional action space ($\mathcal{A} \subset \mathbb{R}^3$).

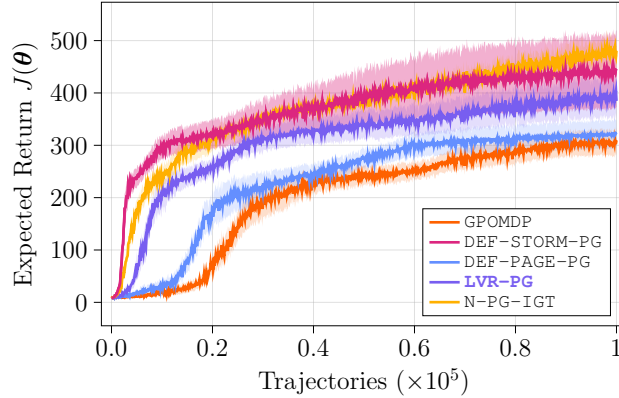


Figure 8: Learning curves on Hopper-v4. Curves show mean evaluation returns over five runs; shaded regions indicate the estimated standard error of the mean.

Figure 8 shows that DEF-STORM-PG and N-PG-IGT achieve better performance compared to the other algorithms. Again, this shows that momentum updates are empirically more efficient. Here, we see how DEF-PAGE-PG and GPOMDP learn consistently slower than the algorithms that implement momentum, as they start collecting reward after observing 2×10^4 trajectories.

I.5 Gradient-Estimator Variability on CartPole-v1

We complement the return comparisons with a diagnostic of per-trajectory gradient variability on CartPole-v1. For a batch of m_k trajectories, let $\mathbf{g}_{k,i}$ denote the unweighted, baseline-free GPOMDP contribution evaluated at

the current policy parameters. We compute the trace of their sample covariance,

$$\widehat{\mathbb{V}\text{ar}}\left(\widehat{\nabla}_{\theta}^{\dagger}J(\theta_k)\right) = \frac{1}{m_k - 1} \sum_{i=1}^{m_k} \|\mathbf{g}_{k,i} - \bar{\mathbf{g}}_k\|_2^2, \quad \bar{\mathbf{g}}_k = \frac{1}{m_k} \sum_{i=1}^{m_k} \mathbf{g}_{k,i}.$$

We plot the mean of $\log_{10} \widehat{\mathbb{V}\text{ar}}\left(\widehat{\nabla}_{\theta}^{\dagger}J(\theta_k)\right)$ across independent runs. The contributions are evaluated at the current policy parameters using the batches collected by each algorithm: on-policy batches for GPOMDP and N-PG-IGT, and defensive-mixture batches for the configurations of **LVR-PG**. Thus, the statistic summarizes the variability of the per-trajectory gradient contributions used in this diagnostic. The experiment uses the configurations of Tables 5 and 6.

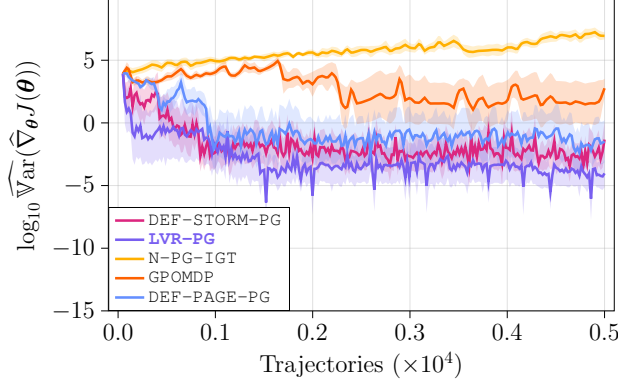


Figure 9: Per-trajectory gradient-variability diagnostic on CartPole-v1 over 5×10^3 training trajectories. Values are shown on a base-10 logarithmic scale and averaged over five independent runs; shaded regions indicate the estimated standard error of the mean.

Figure 9 shows lower per-trajectory gradient variability for DEF-STORM-PG and the hybrid **LVR-PG** configuration than for the baselines over much of training. DEF-PAGE-PG also achieves lower values than GPOMDP, while the values for N-PG-IGT increase over the observed budget.

I.6 Hyperparameter Ablations on CartPole-v1

We examine the sensitivity of DEF-PAGE-PG to $p \in \{0.1, 0.5, 0.9\}$ and of DEF-STORM-PG to $\alpha \in \{0.1, 0.5, 0.9\}$. Both studies use the CartPole-v1 environment with the configuration presented in Tables 5 and 6.

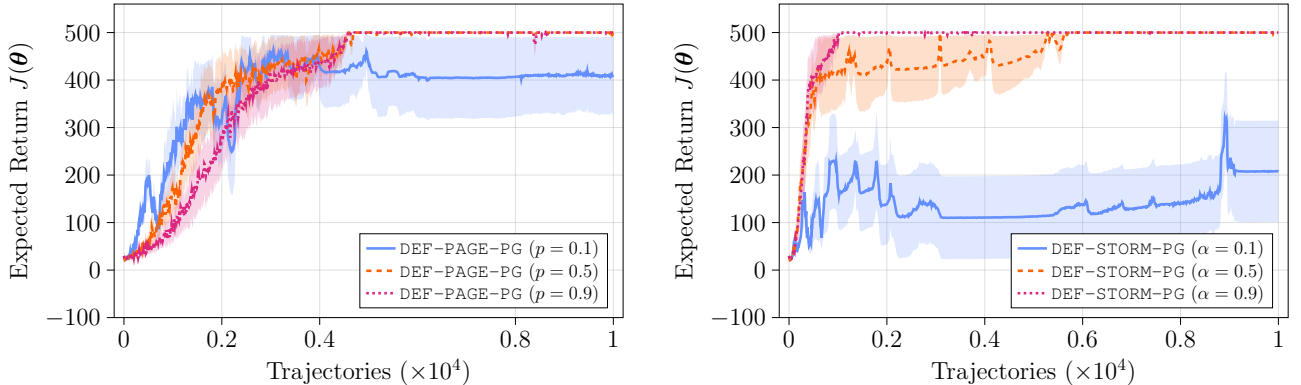


Figure 10: Hyperparameter ablations on CartPole-v1 over 10^4 training trajectories. **Left:** Refresh probability $p \in \{0.1, 0.5, 0.9\}$ for DEF-PAGE-PG. **Right:** Recursive-update parameter $\alpha \in \{0.1, 0.5, 0.9\}$ for DEF-STORM-PG. Curves show mean evaluation returns over five runs, shaded regions indicate the estimated standard error of the mean.

Figure 10 (left) shows that DEF-PAGE-PG is not much less sensitive to the choice of p , although frequent refreshes improve its final performance. Both $p = 0.5$ and $p = 0.9$ reach returns close to 500, whereas $p = 0.1$ settles

near 400. The configuration $p = 0.9$ maintains high returns while selecting recursive mini-batch updates with small probability. This suggests that, without momentum, relying heavily on recursive mini-batch updates is less effective in the tested setting. The improvement with more frequent full-batch refreshes is consistent with using larger on-policy batches to control gradient-estimation variance.

Figure 10 (right) shows that DEF-STORM-PG is very sensitive to the choice of α . The configuration $\alpha = 0.9$ rapidly reaches returns close to 500, while $\alpha = 0.5$ requires substantially more trajectories, and $\alpha = 0.1$ exhibits persistent fluctuations and lower returns, failing to solve the task in the environment. Thus, a large α substantially improves empirical sample efficiency in this setting. Since the historical correction is weighted by $1 - \alpha$, the best-performing configuration retains a smaller correction from the previous estimate.

I.7 Hyperparameter Grid Search for LVR-PG on CartPole-v1

We evaluate hybrid configurations of LVR-PG with $\alpha \in \{0.01, 0.5, 0.9\}$ and $p \in \{0.05, 0.3, 0.5\}$. The policy architecture and training settings match the ablation study in Section I.6. Each configuration uses five independent runs. Figure 11 summarizes normalized AUC and terminal return, with terminal return computed as the mean of the last three evaluations in each run.

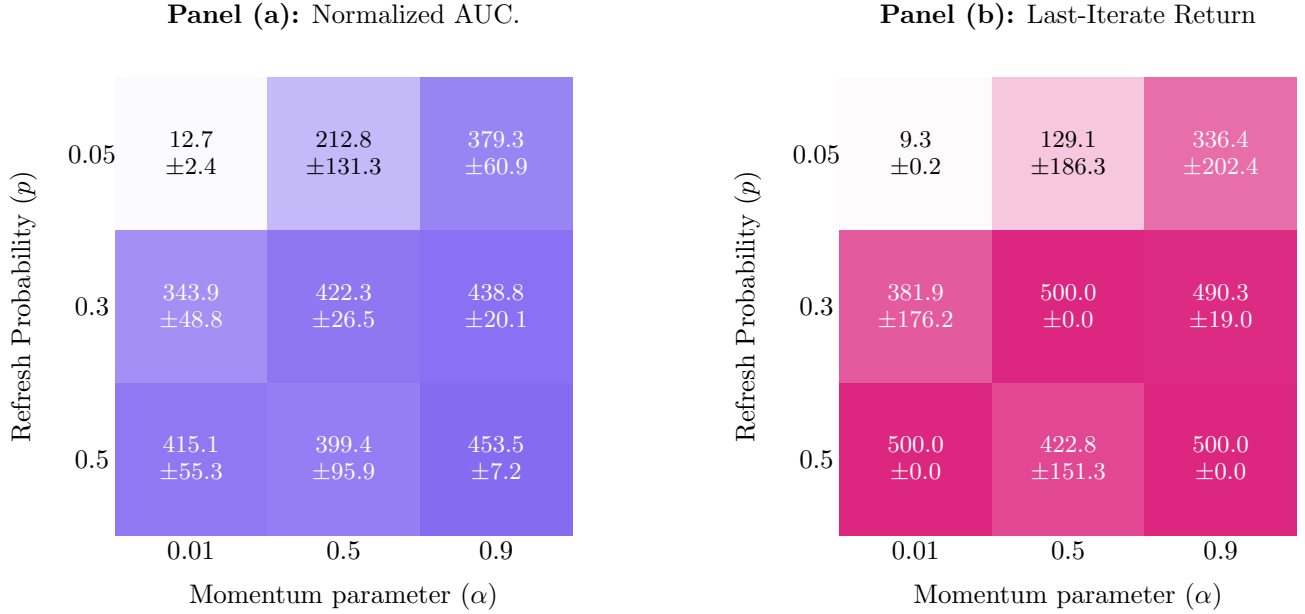


Figure 11: Hybrid hyperparameter grid on CartPole-v1. **Left:** Mean normalized AUC. **Right:** Mean terminal return, averaged over the last three evaluations in each run. Entries summarize five independent runs; shaded regions indicate the estimated standard error of the mean.

The grid shows sensitivity to both parameters. At $p = 0.05$, increasing α raises the mean AUC from 12.7 to 379.3 and the mean terminal return from 9.3 to 336.4. More frequent refreshes substantially improve the low- α configuration. Three configurations, $(0.5, 0.3)$, $(0.01, 0.5)$, and $(0.9, 0.5)$, attain a mean terminal return of 500 at the displayed precision, while their mean AUC values differ. The configuration $(\alpha, p) = (0.9, 0.5)$ achieves the largest mean AUC in the grid, 453.5 ± 7.2 , combining high terminal performance with the strongest performance over the training budget.

These results identify effective choices within the hybrid configuration and show that its performance depends on the interaction between α and p . Together with the experiments presented in Appendix I.6, they motivate selecting DEF-STORM-PG as the primary representative of LVR-PG, since it achieves consistently high mean AUC across the tested environments without periodic full-batch refreshes.

I.8 Computational Resources

All the experiments were run on a 2023 14-inch MacBook Pro equipped with an Apple M2 Pro chip and 16 GB of RAM.