PARAMETER STEP-BASED EXPLORATION FOR LEARNING DETERMINISTIC POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we investigate deterministic policy learning through an auxiliary objective, J_σ, induced by step-based Gaussian perturbations of the policy parameters, where $\sigma > 0$ denotes the exploration scale. First, we establish discrepancy bounds quantifying the gap between the auxiliary and deterministic objectives under both Lipschitz and smooth Markov Decision Processes (MDPs), of order $O(\sqrt{d}\sigma)$ and $O(d\sigma^2)$, respectively. Furthermore, analyzing a simultaneous actor-critic architecture under realistic Markovian sampling rather than idealized i.i.d. sampling, we prove that our approach reaches an ε-stationary point of J_σ, up to critic approximation error, with a sample complexity of $O(d^{1.5}\sigma^{-3}\varepsilon^{-2})$ in the Lipschitz case and $O(d\sigma^{-2}\varepsilon^{-2})$ in the smooth case. Under additional smoothness, we also bound the gradient discrepancy between the auxiliary and deterministic objectives by $O(d^{1.5}\sigma)$, thereby relating stationarity of the explored objective to stationarity of the deterministic deployment objective as the exploration scale decreases. Finally, we evaluate PSBE and a deep variant, Deep-PSBE, on standard MuJoCo continuous-control benchmarks against parameter- and action-space exploration baselines.

1 INTRODUCTION

Reinforcement learning in continuous action spaces raises a fundamental tension between exploration during training and precise control at deployment. In many continuous-control applications, including robotics and autonomous control, a deterministic policy is attractive because it maps each state to a reproducible action. This motivation underlies deterministic policy-gradient methods and their deep-RL extensions (Silver et al., 2014; Lillicrap et al., 2016; Fujimoto et al., 2018). At the same time, deterministic-policy methods typically rely on an explicit exploration mechanism during training to achieve sufficient state-action coverage.

A standard approach is **action-based exploration**, in which noise is added directly to the action produced by the deterministic policy (Lillicrap et al., 2016; Fujimoto et al., 2018; Montenegro et al., 2025). This mechanism is simple and compatible with transition-level actor-critic updates, but the perturbation acts locally in action space and is often sampled independently across decision steps. **Parameter-based exploration** instead perturbs the policy parameters, producing state-dependent changes in behavior (Sehnke et al., 2010; Salimans et al., 2017; Plappert et al., 2018; Fortunato et al., 2018; Montenegro et al., 2024). Classical parameter-search methods such as Parameter-Exploring Policy Gradients (PGPE; Sehnke et al., 2010) and Evolution Strategies (ES; Salimans et al., 2017) typically keep a perturbation fixed over an episode and evaluate it through trajectory returns, limiting their direct compatibility with transition-level temporal-difference (TD) learning.

Existing theory addresses complementary aspects of deterministic policy learning and exploration. Xiong et al. (2022) establish finite-sample convergence guarantees for deterministic actor-critic methods in a differentiable-policy setting, while Montenegro et al. (2024; 2025) study stochastic exploration with deterministic deployment under suitable structural assumptions. ZOAC (Lei et al., 2025) also considers step-wise parameter-space exploration within an actor-critic framework, but leaves finite-time convergence with value-function approximation error open. We instead analyze the finite-sample convergence of simultaneous actor-critic updates with a learned TD critic under dependent Markovian sampling, while explicitly controlling critic approximation error and the discrepancy between exploratory training and deterministic deployment.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study **Parameter Step-Based Exploration (PSBE)**, in which, at state s, the action is generated as $a = \mu_{\theta+\sigma u}(s)$ with $u \sim \mathcal{N}(0, I_d)$, where d is the policy-parameter dimension and $\sigma > 0$ is the exploration scale. A fresh perturbation is independently resampled at every decision step. This induces a stochastic policy $\pi_{\theta,\sigma}$ and an explored objective $J_\sigma(\theta)$, while deployment uses the deterministic policy μ_θ with objective $J_D(\theta)$. Because perturbations are resampled within trajectories, J_σ generally differs from the episode-level parameter-smoothed objective $\mathbb{E}_{u\sim\mathcal{N}(0,I_d)}[J_D(\theta+\sigma u)]$. Equivalently, the induced stochastic policy can be viewed as the Gaussian smoothing of the mapping from the policy parameter to the Dirac measure associated with its deterministic action.

Building on this construction, we analyze a simultaneous actor-critic scheme in which the actor is updated through a zeroth-order gradient estimator and a learned linear TD critic is updated from the same consecutive Markov mini-batch, without first completing policy evaluation. This setting introduces several coupled difficulties: the TD fixed point moves as the actor changes, the Gaussian multiplier u/σ is unbounded, the sampled states are dependent, and reducing σ improves agreement with deterministic deployment while increasing the variance of the zeroth-order estimator. Our analysis jointly controls actor progress and critic tracking and explicitly retains the effects of critic approximation and Markovian sampling.

Contributions. Throughout, sample complexity counts environment transitions, and fixed regularity, discount, projection, and conditioning constants are suppressed.

- **Exploration-to-deployment regularity and discrepancy.** We characterize the stochastic policy induced by step-wise Gaussian parameter exploration and derive its zeroth-order gradient representation. Under Lipschitz assumptions on the reward, transition kernel, and policy parameterization, we establish smoothness of J_σ without requiring differentiability of the deterministic controller. We further prove $|J_\sigma(\theta) - J_D(\theta)| = O(\sqrt{d}\,\sigma)$ in the Lipschitz regime, improving to $O(d\sigma^2)$ under additional smoothness.
- **Finite-sample convergence of simultaneous PSBE actor-critic.** We analyze the coupled evolution of the actor and a learned linear TD critic using consecutive Markovian mini-batches, explicitly accounting for critic approximation error. To achieve $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J_\sigma(\theta_t)\|^2 \le \varepsilon + O(\Xi_{\text{critic}})$, we establish sample-complexity bounds of $O(d^{1.5}\sigma^{-3}\varepsilon^{-2})$ in the Lipschitz regime and $O(d\sigma^{-2}\varepsilon^{-2})$ in the smooth regime. At fixed exploration scale, these bounds match the ε^{-2} accuracy dependence reported by Xiong et al. (2022), while extending the analysis to zeroth-order policy updates computed from consecutive Markovian mini-batches.
- **Deterministic-policy stationarity.** Under the smoothness assumptions, we establish $\|\nabla J_\sigma(\theta) - \nabla J_D(\theta)\| = O\Big(\frac{d^{3/2}\sigma}{1-\gamma}\Big)$. Combining this discrepancy bound with our finite-sample convergence result for J_σ, and choosing the exploration scale appropriately, we obtain
$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla J_D(\theta_t)\|^2\big] \le \varepsilon$$
with sample complexity $O\Big(\frac{d^4}{(1-\gamma)^5\varepsilon^3}\Big)$.

Finally, we introduce a deep variant of PSBE and evaluate both PSBE and Deep-PSBE on standard MuJoCo continuous-control benchmarks against parameter- and action-space exploration baselines, with particular attention to the exploration-to-deployment discrepancy and the performance of the deterministic policy after exploration is removed.

2 Preliminaries

2.1 Markov Decision Process

We consider the standard reinforcement learning (RL) setting where an agent interacts with a stochastic environment. Following foundational literature Puterman (1994); Sutton & Barto (2018), we model this system as a discrete-time discounted Markov Decision Process (MDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0)$. Here, $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ denote the continuous state and

action spaces, respectively. The transition kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ represents the probability density of transitioning to a next state $s' \in \mathcal{S}$ given the current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. The reward function $r : \mathcal{S} \times \mathcal{A} \to [0, R_{max}]$ maps state-action pairs to a bounded scalar feedback signal, $\gamma \in (0, 1)$ is the discount factor, and P_0 represents the initial state distribution. Throughout this paper, we assume that the Markov chains induced by the evaluated policies are ergodic and admit unique stationary distributions.

We define the deterministic policy as a function $\mu_\theta : \mathcal{S} \to \mathcal{A}$, the stochastic policy as a function $\pi^\theta : \mathcal{S} \to \Delta(\mathcal{A})$, such that given the current state s, the deterministic policy outputs a specific action $a = \mu_\theta(s)$, and the stochastic policy outputs a distribution $\pi^\theta(.|s)$ over the action space $\mathcal{A}$.

While traditional RL often relies on stochastic action selection, we focus on deterministic policies Silver et al. (2014), parameterized by a vector $\theta \in \mathbb{R}^d$. We define the deterministic policy as a function $\mu_\theta : \mathcal{S} \to \mathcal{A}$, such that given a current state s, the policy outputs a specific, continuous action $a = \mu_\theta(s)$.

To evaluate the performance of a given policy π (stochastic or deterministic), we define the state value function, which represents the expected sum of discounted rewards obtained by following π starting from state s:

$$\forall s \in \mathcal{S} \quad V^\pi(s) = \mathbb{E}_{\tau \sim p^\pi}[R(\tau)|s_0 = s], \tag{1}$$

where τ is a full trajectory $(s_0, a_0, s_1, a_1, ...)$ sampled wrt to the dsitribution $p^\pi(\tau) = \Pi_{i=0}^{\infty} P(s_{i+1}|s_i, a_i)\pi(a_i|s_i)$. Correspondingly, the action-value function $Q^\pi(s, a)$ evaluates the expected return of taking action a in state s, and following the policy μ_θ:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} \quad Q^\pi(s, a) = \mathbb{E}_{\tau \sim p^\pi}[R(\tau)|s_0 = s, a_0 = a]. \tag{2}$$

We recall that the action-value function satisfies the Bellman consistency equation:

$$Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]. \tag{3}$$

We denote by d^π the discounted visitation distribution given by:

$$d^\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s|\pi), \tag{4}$$

where $\mathbb{P}(s_t = s|\pi)$ is the probability of being in state s at time step t following the policy π, and it is given by the recursion relation

$$\mathbb{P}(s_t = s|\pi) = \int_{s',a'} P(s|s', a')\pi(a'|s')\mathbb{P}(s_{t-1} = s'|\pi)da'ds'. \tag{5}$$

It was shown in Konda & Tsitsiklis (2000), that for any policy π d^π is the stationnary distribution induced by the discounted transition kernel $\tilde{P^\pi}(s'|s) = \gamma P^\pi(s'|s) + (1-\gamma)P_0(s)$, where $P^\pi(s'|s) = \mathbb{E}_{a \sim \pi}[P(s'|s, a)]$.

Throughout this paper, we assume that the Markov chains induced by the evaluated policies and the discounted transition kernels are ergodic and admit unique stationary distributions.

2.2 Problem Formulation and Core Idea

Our primary goal is to maximize the deterministic performance objective:

$$J_D(\theta) = \mathbb{E}_{s \sim P_0}[V^{\mu_\theta}(s)]. \tag{6}$$

Under standard regularity assumptions on the MDP components, its gradient is given by the Deterministic Policy Gradient (DPG) theorem Silver et al. (2014):

$$\nabla J_D(\theta) = \mathbb{E}_{s \sim d^{\mu_\theta}}[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s, a)|_{a=\mu_\theta(s)}]. \tag{7}$$

To optimize this objective while ensuring robust exploration and circumventing the strict regularity assumptions required for direct DPG estimation, we introduce an auxiliary objective J_σ:

$$J_\sigma(\theta) = \mathbb{E}_{s \sim P_0}[V^{\pi^{\theta,\sigma}}(s)]. \tag{8}$$

Here, $\pi^{\theta,\sigma}$ is an induced stochastic policy operating via parameter-space perturbation: given a state s, we sample a standard Gaussian variable $u \sim \mathcal{N}(0, I_d)$ and execute the deterministic action $a = \mu_{\theta+\sigma u}(s)$. We formally establish the structural properties of this induced policy in the following proposition (proof deferred to the Appendix A).

Proposition 1. *For every parameter $\theta \in \mathbb{R}^d$ and state $s \in \mathcal{S}$, the induced policy $\pi^{\theta,\sigma}(\cdot|s)$ is a valid probability measure over the action space $\mathcal{A}$ satisfying the following two identities:*

- *The policy is exactly the Gaussian smoothing of the Dirac measure centered at the perturbed action:*
$$\pi^{\theta,\sigma}(\cdot|s) = \mathbb{E}_{u\sim\mathcal{N}(0,I_d)}[\delta_{\mu_{\theta+\sigma u}(s)}(\cdot)]. \tag{9}$$
- *For any positive measurable function $f : \mathcal{A} \to \mathbb{R}$, it holds that:*
$$\int_{\mathcal{A}} f(a)\pi^{\theta,\sigma}(da|s) = \mathbb{E}_{u\sim\mathcal{N}(0,I_d)}[f(\mu_{\theta+\sigma u}(s))]. \tag{10}$$

Key Insight: Bypassing Differentiability via Dirac Smoothing. In the classical analysis of vanilla policy gradient methods, the analytical differentiability of the stochastic policy with respect to its parameters is mandatory to apply the log-derivative trick. Conversely, standard DPG relies on the differentiability of both the deterministic policy μ_θ and the action-value function.

Leveraging the induced policy $\pi^{\theta,\sigma}$ completely bypasses these requirements. Because the induced stochastic policy is structurally equivalent to the Gaussian smoothing of a Dirac distribution, we are *not* smoothing the global deterministic objective directly—which is prone to high variance and fails to preserve Lipschitzness. Instead, we are smoothing only the deterministic policy itself. This transfers the mathematical smoothness of the Gaussian noise directly to the action distribution. Consequently, the auxiliary objective J_σ becomes differentiable without imposing *any* regularity assumptions on the underlying MDP components, as formalized below (proof deferred to the Appendix A).

Proposition 2. *The auxiliary objective J_σ is differentiable with respect to θ, and its gradient is given by:*

$$(1-\gamma)\nabla J_\sigma(\theta) = \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}\left[\frac{Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))}{\sigma}u\right] \tag{11}$$

$$= \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}\left[\frac{A^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))}{\sigma}u\right]. \tag{12}$$

where $A^{\pi^{\theta,\sigma}}(s,a) = Q^{\pi^{\theta,\sigma}}(s,a) - V^{\pi^{\theta,\sigma}}(s)$ is the advantage-function.

The second equality holds because the state-value baseline $V^{\pi^{\theta,\sigma}}(s)$ is independent of the Gaussian perturbation u, and the expectation of u is zero. Injecting a state-dependent baseline is a well-established technique in the policy gradient literature Williams (1992); Sutton et al. (1999) to mitigate the high variance inherent to Monte Carlo gradient estimators. In our framework, we theoretically demonstrate, in Appendix A, that subtracting the state-value baseline $V^{\pi^{\theta,\sigma}}(s)$ achieves the asymptotic order of variance reduction for our zero-order estimator, as subtracting the analytically optimal baseline.

2.3 Actor-critic architecture and the mini-batch Markovian sampling

Theoretical analyses of actor-critic architectures typically evaluate gradient estimators under either independent and identically distributed (i.i.d.) sampling or Markovian sampling (Xu et al., 2020; Wu et al., 2020). While i.i.d. sampling yields unbiased gradients, it relies on the unrealistic assumption of a generative oracle to draw states directly from the stationary distribution (Xu et al., 2020; Wu et al., 2020). Conversely, Markovian sampling collects states sequentially via the environment's natural transition dynamics, introducing a Markovian bias that must be controlled (Xu et al., 2020; Wu et al., 2020).

To theoretically bound this bias and guarantee that trajectory averages converge to the true spatial expectation, we rely on the standard uniform ergodicity assumption (Xu et al., 2020; Wu et al., 2020).

Assumption 1 (Uniform Mixing Time of the Discounted Chain)**.** For any policy π, we define the discounted transition kernel as:
$$\tilde{P}^\pi(\cdot \mid s) = \gamma P^\pi(\cdot \mid s) + (1-\gamma)P_0(\cdot). \tag{13}$$

Algorithm 1 PSBE

Initialize:
Actor parameter θ_0, critic parameter ω_0, draw $s_{0,0}$ from some initial distribution
for $t = 0, \dots, T-1$ **do**
  Sample M **Markovian transitions:**
  $s_{t,0} = s_{t-1,M}$
  for $i = 0, \dots, M-1$ **do**
    $u_i \sim \mathcal{N}(0, I)$, $a_{t,i} = \mu_{\theta_t + \sigma u_i}(s_{t,i})$, $s_{t,i+1} \sim \tilde{P}(\cdot|s_{t,i}, a_{t,i})$
  end for
  Temporal difference error:
    $\forall i \in \{0, \dots, M-1\} : \delta_{t,i} = r_{t,i} + \gamma \varphi(s_{t,i+1})^T \omega_t - \varphi(s_{t,i})^T \omega_t$
  Update the actor:
    $\theta_{t+1} = \theta_t + \xi^\theta \frac{1}{M} \sum_{i=0}^{M-1} \frac{\delta_{t,i}}{\sigma} u_i$
  Update the critic:
    $\omega_{t+1} = \Pi_{R_\omega} \left(\omega_t + \xi^\omega \frac{1}{M} \sum_{i=0}^{M-1} \delta_{t,i} \varphi(s_{t,i}) \right)$
end for

We assume the Markov chain induced by $\tilde{P}^\pi$ is uniformly geometrically ergodic. Specifically, there exist constants $\kappa > 0$ and $\rho \in (0, 1)$ such that for any initial state $s_0 \in \mathcal{S}$ and time step $t \geq 0$, the total variation distance to its unique stationary distribution d^π satisfies:

$$d_{TV} \left(\tilde{P}_t^\pi(\cdot \mid s_0), d^\pi \right) \leq \kappa \rho^t, \tag{14}$$

where $\tilde{P}_t^\pi(\cdot \mid s_0)$ is the t-step transition probability starting from s_0.

Operating strictly on a single, continuous sample path without resets, we propose the Parameter Step-Based Exploration (PSBE) method, detailed in Algorithm 1. At iteration t, we sample a mini-batch of size M. To maintain the Markovian property across the entire learning process without a generative model, the final state of the current iteration initializes the next iteration ($s_{t+1,0} = s_{t,M}$).

Critic Evaluation via TD(0). The critic evaluates the current policy by estimating the state-value function via Temporal Difference learning, TD(0), utilizing linear function approximation $V(s) \approx \omega^T \varphi(s)$ Bhandari et al. (2018). It computes the TD error $\delta_{t,i}$ to serve as the advantage estimator for the actor. The critic parameter ω is updated via semi-gradient descent and bounded by a projection operator Π_{R_ω} to ensure stability (Xu et al., 2019). For this linear TD(0) update to successfully converge to the optimal projected value function, the underlying feature matrix must be well-conditioned.

To guaranty the existence and uniqueness of ω_θ^*, the following assumptions are made in the analysis of TD(0) Bhandari et al. (2018); Zou et al. (2019); Xu et al. (2019; 2020).

Assumption 2. The feature mapping has a bounded norm $\|\varphi(.)\| \leq \Phi_{\max}$. For all policy parameters $\theta \in \mathbb{R}^d$, the matrix $A(\theta)$ is negative definite, and its maximum eigenvalue is bounded above by a constant $-\lambda < 0$.

A crucial assumption, made also in the previous analysis of the actor critic architecture Bhandari et al. (2018); Zou et al. (2019); Qiu et al. (2019); Xu et al. (2020), is the approximation error defined as $\mathbb{E}_{s \sim d^{\pi^{\theta,\sigma}}} [|\varphi(s)^T \omega_\theta^* - V^{\pi^{\theta,\sigma}}(s)|^2]$ is uniformly bounde wrt θ.

Assumption 3. There exists $\Xi_{critic} > 0$ such that: $\sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{s \sim d^{\pi^{\theta,\sigma}}} [|\varphi(s)^T \omega_\theta^* - V^{\pi^{\theta,\sigma}}(s)|^2] \leq \Xi_{critic}$.

Two-Time-Scale Architecture. Algorithm 1 integrates these simultaneous updates into a two-time-scale stochastic approximation framework (Wu et al., 2020). Rather than fully decoupling the actor and critic—which inefficiently discards historical parameters to generate fresh episodes for every critic evaluation—both parameters are updated concurrently from the same continuous Markovian stream (Wu et al., 2020). To guarantee convergence, the critic utilizes a significantly larger step size than the actor ($\xi^\omega \gg \xi^\theta$) (Wu et al., 2020). This timescale separation forces the actor to change slowly, allowing the faster-moving critic to continuously and accurately track the target value function as if the policy were virtually stationary (Wu et al., 2020).

3 FUNDAMENTAL ASSUMPTIONS

Our theoretical analysis operates under two distinct regularity regimes. The first is a foundational Lipschitz regime, requiring only that the MDP's transition kernel, reward function, and the deterministic policy μ_θ are Lipschitz continuous. The second is a Smooth regime, which extends these conditions by requiring the gradients of both the MDP components and the policy to also be Lipschitz continuous. These assumptions align with standard practices in the recent reinforcement learning literature (Papini et al., 2018; Agarwal et al., 2020; Zhang et al., 2020).

3.1 LIPSCHITZ REGIME

Assumption 4 (Lipschitz Continuous MDP)**.** The MDP satisfies the following Lipschitz conditions with respect to the action variable: The log-transition probability $\log P(s' \mid s, a)$ is L_P-Lipschitz:

$$|\log P(s' \mid s, a_1) - \log P(s' \mid s, a_2)| \leq L_P \|a_1 - a_2\|,$$

for all $s, s' \in \mathcal{S}$ and $a_1, a_2 \in \mathcal{A}$. The reward function $r(s, a)$ is L_r-Lipschitz:

$$|r(s, a_1) - r(s, a_2)| \leq L_r \|a_1 - a_2\|,$$

for all $s \in \mathcal{S}$ and $a_1, a_2 \in \mathcal{A}$.

Assumption 5 (Lipschitz Continuous Policy)**.** The deterministic parametric policy μ_θ is L_μ-Lipschitz with respect to its parameter vector θ:

$$\|\mu_{\theta_1}(s) - \mu_{\theta_2}(s)\| \leq L_\mu \|\theta_1 - \theta_2\|,$$

for all $s \in \mathcal{S}$ and $\theta_1, \theta_2 \in \mathbb{R}^d$.

3.2 SMOOTH REGIME

Assumption 6 (Smooth MDP)**.** The MDP satisfies the following smoothness conditions: The log-transition probability $\log P(s'|s, a)$ is $L_{2,P}$-smooth:

$$\|\nabla_a \log P(s'|s, a)|_{a=a_1} - \nabla_a \log P(s'|s, a)|_{a=a_2}\| \leq L_{2,P} \|a_1 - a_2\|,$$

for all $s, s' \in \mathcal{S}$ and $a_1, a_2 \in \mathcal{A}$. The reward function $r(s, a)$ is $L_{2,r}$-smooth:

$$\|\nabla_a r(s, a)|_{a=a_1} - \nabla_a r(s, a)|_{a=a_2}\| \leq L_{2,r} \|a_1 - a_2\|,$$

for all $s \in \mathcal{S}$ and $a_1, a_2 \in \mathcal{A}$.

Assumption 7 (Smooth Policy)**.** The deterministic parametric policy μ_θ is $L_{2,\mu}$-smooth with respect to its parameter vector θ:

$$\|\nabla_\theta \mu_{\theta_1}(s) - \nabla_\theta \mu_{\theta_2}(s)\| \leq L_{2,\mu} \|\theta_1 - \theta_2\|,$$

for all $s \in \mathcal{S}$ and $\theta_1, \theta_2 \in \mathbb{R}^d$.

4 MAIN THEORY

4.1 PERFORMANCE DISCREPANCY BOUNDS

In the PSBE approach, $J_\sigma(\theta)$ serves as an auxiliary surrogate objective for learning a stationary policy. To relate optimization of J_σ to deterministic deployment, we establish discrepancy bounds between the auxiliary objective J_σ and the deterministic objective J_D.

Proposition 3 (Discrepancy Bound 1)**.** *For any policy parameter $\theta \in \mathbb{R}^d$, the discrepancy between the smoothed auxiliary objective $J_\sigma(\theta)$ and the true deterministic objective $J_D(\theta)$ is bounded as follows:*

Lipschitz Regime: *Under Assumptions 4 and 5, we have:*

$$|J_\sigma(\theta) - J_D(\theta)| \leq \mathcal{O}\left(\frac{\sqrt{d}\sigma}{1-\gamma}\right). \tag{15}$$

Smooth Regime: *Under the additionnal assumptions 6 and 7, we have:*

$$|J_\sigma(\theta) - J_D(\theta)| \leq= \mathcal{O}\left(\frac{d\sigma^2}{1-\gamma}\right). \tag{16}$$

$\mathcal{O}()$ contains MDP-dependant constants.

The next proposition shows that, under the smoothness assumptions of the MDP, the difference in norme between the gradients of the two objectives can be controlled as well by the width σ. In fact:

Proposition 4 (Discrepancy bound 2)**.** *For any policy parameter $\theta \in \mathbb{R}^d$ and under assumptions 4, 5, 6 and 7, it holds:*

$$\|\nabla J_\sigma(\theta) - \nabla J_D(\theta)\| \leq \mathcal{O}(\frac{d^{\frac{3}{2}}\sigma}{1-\gamma}). \tag{17}$$

$\mathcal{O}()$ contains MDP-dependant constants.

4.2 Smoothness of the auxiliary objective

To guarantee an ϵ-stationary point for the auxiliary function—and consequently for the deterministic objective via discrepancy bound 2—we first establish smoothness inequalities for the two aforementioned regimes (proof deferred to the Appendix B).

Theorem 5. *Under assumptions 4 and 5, the auxiliary objective J_σ is $L^L_{2,\sigma}$-smooth:*

$$\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| \leq L^L_{2,\sigma}\|\theta_1 - \theta_2\|, \tag{18}$$

for all $\theta_1, \theta_2 \in \mathbb{R}^d$, with $L^L_{2,\sigma} = \mathcal{O}(\frac{\sqrt{d}}{\sigma(1-\gamma)^2})$. $\mathcal{O}()$ contains MDP-dependant constants.

As in the Gaussian smoothing theory of Nesterov & Spokoiny (2017), the smoothness constante is of order d/σ, when the objective is only supposed lipschitz. On the other hand and in the smooth regime the smoothness constante is excatly the smoothness constant of the deterministic objective.

Theorem 6. *Under assumptions 4, 5, 6 and 7, the auxilary objective J_σ is $L^S_{2,\sigma}$-smooth:*

$$\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| \leq L^S_{2,\sigma}\|\theta_1 - \theta_2\|, \tag{19}$$

for all $\theta_1, \theta_2 \in \mathbb{R}^d$, with $L^S_{2,\sigma} = \mathcal{O}(\frac{1}{(1-\gamma)^2})$. $\mathcal{O}()$ contains MDP-dependant constants.

4.3 Convergence analysis of PSBE

In algorithm 1, the critic uses TD with linear function approximation $\tilde{V}(s,\omega) = \varphi(s)^T\omega$ where $s \in \mathcal{S} \mapsto \varphi(s)$ is a feature mapping. We denote by ω^*_θ the limiting point of TD(0) under the stochastic policy $\pi^{\theta,\sigma}$. More precisely, we define the following quantities:

$$A(\theta) := \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}}, u\sim\mathcal{N}(0,I), s'\sim P(\cdot|s,\mu_{\theta+\sigma u}(s))}\Big[\varphi(s)\big(\gamma\varphi(s') - \varphi(s)\big)^\top\Big], \tag{20}$$

$$b(\theta) := \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}}}\Big[\varphi(s)r(s)\Big]. \tag{21}$$

Then the limiting point ω^*_θ satisfies:

$$A(\theta)\omega^*_\theta + b = 0. \tag{22}$$

First, we establish an ascent lemma derived directly from the smoothness of J_σ. For brevity, let $L_{2,\sigma}$ denote the smoothness constant in either the Lipschitz or the Smooth regime. We denote by $\mathcal{F}_t$ the filtration containing all randomness up to iteration t of Algorithm 1 (Proof of all the results of the current subsection are deferred in Appendix D)

Lemma 7 (Ascent Lemma)**.** *Let $\theta_t \in \mathbb{R}^d$, and θ_{t+1} be the actor update in Algorithm 1. Then,*

$$\begin{aligned}\mathbb{E}\left[J_\sigma(\theta_{t+1}) \mid \mathcal{F}_t\right] \geq J_\sigma(\theta_t) + \frac{\xi^\theta}{4}\mathbb{E}[\|\nabla J_\sigma(\theta_t)\|^2|\mathcal{F}_t] - \frac{5\xi^\theta}{4}\mathbb{E}\left[\|\bar{h}(\omega_t) - H(\omega_t)\|^2 \mid \mathcal{F}_t\right] \\ - \frac{5\xi^\theta}{4}\mathbb{E}\left[\|H(\omega_t) - \nabla J_\sigma(\theta_t)\|^2 \mid \mathcal{F}_t\right] - \frac{L_{2,\sigma}(\xi^\theta)^2}{2}\sigma^2_M.\end{aligned} \tag{23}$$

Remark 1. This inequality breaks down the expected policy update into the true gradient improvement, offset by two distinct sources of error: estimation biases (Markovian, TD, and function approximation) and the exploration variance σ^2_M.

It is worth highlighting why the variance inevitably scales as $\mathcal{O}(d/\sigma^2)$ while the biases do not. By design, the expected value of our zero-order estimator bypasses dimensional dependence, satisfying

$\mathbb{E}_u\left[f(\theta+\sigma u)\frac{u}{\sigma}\right] = \mathbb{E}[\nabla f(\theta+\sigma u)]$. However, bounding the *variance* relies on the second moment, which brings the d/σ^2 factor back into play. This is the fundamental mathematical cost of exploring a d-dimensional parameter space using only scalar rewards. Fortunately, because this variance penalty is scaled quadratically by $(\xi^\theta)^2$, we can neutralize it by choosing a sufficiently small actor learning rate.

In contrast, the bias terms are completely independent of d and σ because they govern the state-space rather than the parameter-space. They are driven entirely by the environment's dynamics—specifically, the MDP's mixing time and the critic's structural limits. Since these biases arise from temporal distribution shifts rather than parameter perturbations, they sidestep the d/σ^2 penalty entirely. Instead, they must be controlled by increasing the batch size M, ensuring our empirical samples adequately capture the true stationary distribution.

Remark 2. It is worth noting that the identity $\mathbb{E}_u\left[f(\theta+\sigma u)\frac{u}{\sigma}\right] = \mathbb{E}[\nabla f(\theta+\sigma u)]$ holds even if we only assume f is Lipschitz continuous. By Rademacher's theorem, a Lipschitz continuous function is differentiable almost everywhere (a.e.). Consequently, applying integration by parts with respect to the Gaussian measure remains mathematically valid Leoni (2017). This identity is particularly useful because our analysis never evaluates the gradient of f pointwise; instead, we manipulate its expectation over the Gaussian distribution, meaning the points where the function is not differentiable do not affect the expectation.

Theorem 8 (Coupled actor-critic inequalities)**.** *Let $\xi^\omega \le \min\{\frac{\lambda}{2L_g^2}, \frac{1}{\lambda}\}$ and let $C > 0$ be an MDP-dependent constant (defined in Appendix D). Assume the actor step size satisfies $\xi^\theta \ll \xi^\omega$ such that the coupling factor $\beta(\xi^\theta,\xi^\omega) := \frac{4}{\lambda^2(1-\gamma)^2}\left(\frac{\xi^\theta}{\xi^\omega}\right)^2 + \frac{4}{\lambda(1-\gamma)}\left(\frac{\xi^\theta}{\xi^\omega}\right) \le \frac{1}{C}$. With $\sigma_M^2 = \mathcal{O}\left(\frac{d}{\sigma^2 M}\right)$, there exist MDP-dependent constants $c_1, c_2, c_3, c_1', c_2', c_3' > 0$ such that:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J_\sigma(\theta_t)\|^2\right] \le \frac{c_1}{\xi^\theta T} + \frac{c_2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\omega_t - \omega_{\theta_t}^*\|^2\right] + \frac{c_3}{M} + L_{2,\sigma}\frac{\xi^\theta}{2}\sigma_M^2 + \mathcal{O}(\Xi_{\textit{critic}}),$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\omega_t - \omega_{\theta_t}^*\|^2\right] \le \frac{c_1'}{T} + \beta(\xi^\theta,\xi^\omega)\frac{c_2'}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J_\sigma(\theta_t)\|^2\right] + \frac{c_3'}{M} + \beta(\xi^\theta,\xi^\omega)\sigma_M^2 + \mathcal{O}(\Xi_{\textit{critic}}).$$

Proof is deferred to Appendix D.

Remark 3. The first inequality follows from the analysis of the actor, while the second is derived from the analysis of the critic. As mentioned earlier, the exploration variance σ_M^2 is controlled not only directly by the actor step size ξ^θ, but also by the coupling factor $\beta(\xi^\theta,\xi^\omega)$, which captures the relative update speed of the critic with respect to the actor. This mechanism aligns with the standard two-time-scale assumption $\xi^\theta \ll \xi^\omega$. Therefore, choosing a sufficiently small actor learning rate ξ^θ serves a dual purpose: it ensures the required time-scale separation for the critic's convergence while simultaneously suppressing the high-dimensional variance σ_M^2.

The next theorem establishes the sample complexity to find an ϵ-stationary policy for the auxiliary objective J_σ. As a result, when we actually deploy the policy, we can simply tune the exploration scale σ to obtain an ϵ-stationary policy for the deterministic objective J_D.

Theorem 9. *Let $\xi^\omega \le \min\{\frac{\lambda}{2L_g^2}, \frac{1}{\lambda}\}$. There exist MDP-dependent constants $a_1, a_2, a_3 > 0$ such that if the actor step size satisfies $\xi^\theta \ll \xi^\omega$ with $\beta(\xi^\theta,\xi^\omega) \le \frac{1}{C}$, then:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J_\sigma(\theta_t)\|^2\right] \le \frac{a_1}{\xi^\theta T} + \frac{a_2}{M} + \left(L_{2,\sigma}\xi^\theta + a_3\beta(\xi^\theta,\xi^\omega)\right)\sigma_M^2 + \mathcal{O}(\Xi_{\textit{critic}}), \tag{24}$$

where $\sigma_M^2 = \mathcal{O}(d/(\sigma^2 M))$. Consequently, setting $\xi^\theta \le \mathcal{O}\left(\frac{(1-\gamma)\sigma^2}{L_{2,\sigma}d}\xi^\omega\right)$ guarantees $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J_\sigma(\theta_t)\|^2\right] \le \epsilon$ with a total sample complexity :

$$TM \ge \begin{cases} \mathcal{O}\left(\frac{d^{1.5}}{(1-\gamma)^3\sigma^3\epsilon^2}\right) & \textit{(Lipschitz regime)}, \\ \mathcal{O}\left(\frac{d}{(1-\gamma)^3\sigma^2\epsilon^2}\right) & \textit{(Smooth regime)}. \end{cases} \tag{25}$$

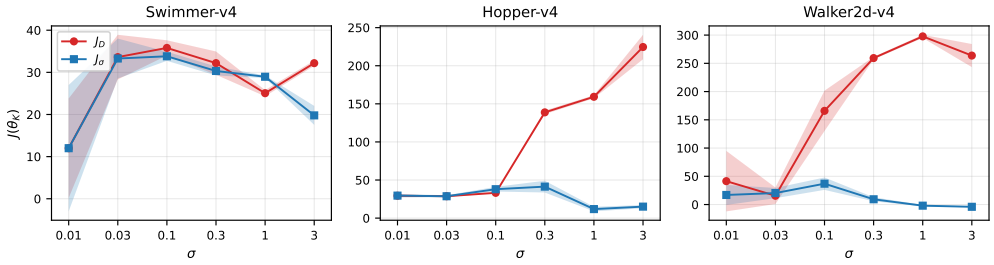

Figure 1: J_σ and J_D against the exploration parameter σ on MuJoCo (5 runs, mean $\pm$ 95% C.I.).

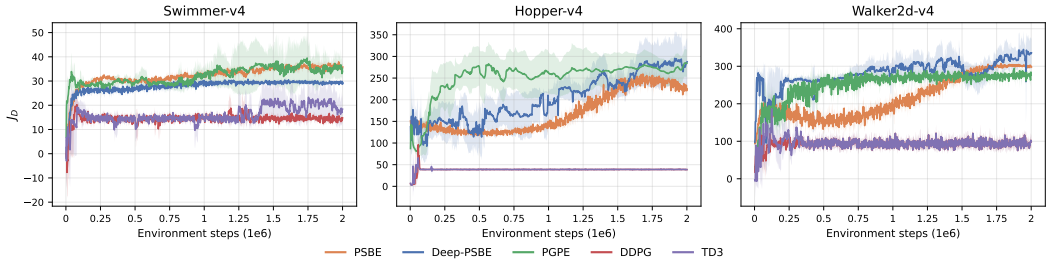

Figure 2: Learning curves on three MuJoCo environments (5 runs, mean $\pm$ 95% C.I.).

Corrolary 10. *In the Smooth regime, and under assumptions 2, 3, the actor and the critic step sizes given in theorem 9 guarantees* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla J_D(\theta_t)\|^2\right] \leq \epsilon$ *with a sample complexity* $TM \geq \mathcal{O}\left(\frac{d^4}{(1-\gamma)^5\epsilon^3}\right)$.

5 Experiments

We first study the performance gap between the deterministic objective J_D and the auxiliary objective J_σ optimized by PSBE as the exploration scale σ varies. Figure 1 shows that larger values of σ are generally associated with a larger exploration-to-deployment gap, qualitatively consistent with Proposition 3.

We then compare PSBE and Deep-PSBE with PGPE (Sehnke et al., 2010; Montenegro et al., 2024), a parameter-space baseline, and DDPG (Lillicrap et al., 2016) and TD3 (Fujimoto et al., 2018), two action-space baselines, under the same environment-interaction budget. Deep-PSBE is a practical extension of PSBE using a nonlinear critic; its complete description is given in Appendix G. For DDPG and TD3, we use linear-critic variants to match the critic class used in PSBE. Figure 2 reports the resulting learning curves.

All methods are evaluated on three MuJoCo tasks (Todorov et al., 2012): Swimmer-v4, Hopper-v4, and Walker2d-v4, using 2×10^6 environment transitions. PSBE and Deep-PSBE use $T = 1000$ iterations with mini-batches of $M = 2000$ transitions, linear policies, and $\gamma = 0.99$. Results are averaged over five seeds and reported with 95% confidence intervals. Hyperparameter grids and additional learning curves are provided in Appendix H.

AI use statement

We used generative AI tools to assist with writing and polishing the manuscript, including improvements to wording, grammar, and clarity. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of the work.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of Temporal Difference learning with linear function approximation. In *Conference on Learning Theory*, pp. 1691–1692. PMLR, 2018.

Erhan Çınlar. *Probability and Stochastics*, volume 261. Springer Science & Business Media, 2011.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, et al. Noisy networks for exploration. In *International Conference on Learning Representations*, 2018.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in Actor-Critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1587–1596. PMLR, 2018.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1008–1014, 2000.

Yuheng Lei, Yao Lyu, Guojian Zhan, Tao Zhang, Jiangtao Li, Jianyu Chen, Shengbo Eben Li, and Sifa Zheng. Zeroth-order actor–critic: An evolutionary framework for sequential decision problems. *IEEE Transactions on Evolutionary Computation*, 29(2):555–569, 2025.

Giovanni Leoni. *A First Course in Sobolev Spaces*. American Mathematical Society, second edition, 2017.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Alexander Yu Mitrophanov. Sensitivity and robustness of finite-state Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Alessandro Montenegro, Marco Mussi, Alberto Maria Metelli, and Matteo Papini. Learning optimal deterministic policies with stochastic policy gradients. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Alessandro Montenegro, Marco Mussi, Matteo Papini, and Alberto Maria Metelli. Convergence analysis of policy gradient methods with dynamic stochasticity. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pp. 4026–4035. PMLR, 2018.

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. In *International Conference on Learning Representations*, 2018.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

Shicong Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On the finite-time convergence of actor-critic algorithm. In *Optimization Foundations for Reinforcement Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 387–395, 2014.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063, 1999.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

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

Huaqing Xiong, Tengyu Xu, Lin Zhao, Yingbin Liang, and Wei Zhang. Deterministic policy gradient: Convergence analysis. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180, pp. 2159–2169. PMLR, 2022.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. In *Advances in Neural Information Processing Systems*, pp. 10633–10643, 2019.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) Actor-Critic algorithms. In *Advances in Neural Information Processing Systems*, 2020.

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

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis of two-time-scale actor-critic algorithms. *IEEE Transactions on Automatic Control*, 64(12):4976–4991, 2019.

A Appendix A

A.1 Proof of proposition 1

To rigorously define the stochastic policy $\pi^{\theta,\sigma}(\cdot \mid s)$, we introduce the random variable:

$$\begin{aligned} X_s^\theta \colon \mathbb{R}^d &\longrightarrow \mathcal{A} \\ u &\longmapsto \mu_{\theta+\sigma u}(s) \end{aligned} \tag{26}$$

where $\mathbb{R}^d$ is equipped with the standard Gaussian probability measure P, whose density is denoted by p. Thus, the probability law of X_s^θ, which defines our stochastic policy, is given for any measurable set $B \subset \mathcal{A}$ by:

$$\pi^{\theta,\sigma}(B \mid s) = P\big(\{u \in \mathbb{R}^d \mid \mu_{\theta+\sigma u}(s) \in B\}\big) \tag{27}$$

Therefore, applying the Transfer Lemma, for any given positive measurable function $f \colon \mathcal{A} \to \mathbb{R}^+$, it holds that:

$$\int_{\mathcal{A}} f(a)\, \pi^{\theta,\sigma}(\mathrm{d}a \mid s) = \int_{\mathbb{R}^d} f(\mu_{\theta+\sigma u}(s)) p(u)\, \mathrm{d}u \tag{28}$$

An alternative and rigorous perspective on the stochastic policy $\pi^{\theta,\sigma}(.|s)$ is obtained by defining a mapping F^s that takes a policy parameter θ as input and outputs a Dirac probability measure $\delta_{\mu_\theta(s)}(\cdot) \in \Delta(\mathcal{A})$ on the action space. Under this view, $\pi^{\theta,\sigma}(\cdot \mid s)$ is precisely the Gaussian smoothing of the mapping F^s at the parameter θ, and it is expressed as:

$$\pi^{\theta,\sigma}(\cdot \mid s) = \int_{\mathbb{R}^d} F^s(\theta + \sigma u) p(u)\, \mathrm{d}u = \int_{\mathbb{R}^d} \delta_{\mu_{\theta+\sigma u}(s)}(\cdot) p(u)\, \mathrm{d}u \tag{29}$$

To formally establish that this integral constitutes a valid probability measure on the action space $\mathcal{A}$, we define the transition kernel K_s^θ over the Borel σ-algebra $\mathcal{B}(\mathcal{A})$:

$$\begin{aligned} K_s^\theta \colon \mathbb{R}^d \times \mathcal{B}(\mathcal{A}) &\longrightarrow [0,1] \\ (u, B) &\longmapsto \delta_{\mu_{\theta+\sigma u}(s)}(B) \end{aligned} \tag{30}$$

By Theorem 6.3 in Çınlar (2011), the composition of this kernel with the standard Gaussian probability measure P (possessing density p) yields a well-defined probability measure on $\mathcal{A}$, denoted as PK:

$$(PK)(\cdot) := \int_{\mathbb{R}^d} K_s^\theta(u, \cdot) P(\mathrm{d}u) = \int_{\mathbb{R}^d} \delta_{\mu_{\theta+\sigma u}(s)}(\cdot) p(u)\, \mathrm{d}u \tag{31}$$

Furthermore, applying Theorem 6.3 in Çınlar (2011) to any positive measurable function f on $\mathcal{A}$, it holds that:

$$\int_{\mathcal{A}} f(a)(PK)(\mathrm{d}a) = \int_{\mathbb{R}^d} \left(\int_{\mathcal{A}} f(a) K_s^\theta(u, \mathrm{d}a) \right) P(\mathrm{d}u) = \int_{\mathbb{R}^d} f(\mu_{\theta+\sigma u}(s)) p(u)\, \mathrm{d}u \tag{32}$$

This strictly verifies the definition in Equation 29 and formalizes the transformation of expectations from the action space to the parameter space.

A.2 Proof of proposition 2

We can write

$$\begin{aligned} V^{\pi^{\theta,\sigma}}(s) &= \int_a \pi^{\theta,\sigma}(a|s) Q^{\pi^{\theta,\sigma}}(s,a) da \\ &= \int_a \int_u \delta(a - \mu_{\theta+\sigma u}(s)) Q^{\pi^{\theta,\sigma}}(s,a) p(u) du da \\ &= \int_u \left(\int_a \delta(a - \mu_{\theta+\sigma u}(s)) Q^{\pi^{\theta,\sigma}}(s,a) p(u) da \right) du \\ &= \mathbb{E}_u[Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s))] \end{aligned} \tag{33}$$

The Bellman consistency equation gives:

$$Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s)) = r(s, \mu_{\theta+\sigma u}) + \gamma \int_{s'} P(s'|s, \mu_{\theta+\sigma u}(s)) V^{\pi^{\theta,\sigma}}(s') ds'$$

taking the gradient of $V^{\pi^{\theta,\sigma}}(s)$, we get:

$$\begin{aligned}
\nabla_\theta V^{\pi^{\theta,\sigma}}(s) &= \nabla_\theta[r(s,\mu_{\theta+\sigma u}(s))] + \gamma\nabla_\theta \int_{s'} \mathbb{E}_u[P(s'|s,\mu_{\theta+\sigma u}(s))]V^{\pi^{\theta,\sigma}}(s')ds' \\
&= \mathbb{E}_u\left[\frac{r(s,\mu_{\theta+\sigma u}(s))}{\sigma}u\right] + \gamma\int_{s'}\nabla_\theta\mathbb{E}_u[P(s'|s,\mu_{\theta+\sigma u}(s))]V^{\pi^{\theta,\sigma}}(s')ds' \\
&\quad + \gamma\int_{s'}\mathbb{E}_u[P(s'|s,\mu_{\theta+\sigma u}(s))]\nabla V^{\pi^{\theta,\sigma}}(s')ds' \qquad (34) \\
&= \mathbb{E}_u\left[\frac{r(s,\mu_{\theta+\sigma u}(s)) + \gamma\int_{s'}P(s'|s,\mu_{\theta+\sigma u}(s))V^{\pi^{\theta,\sigma}}(s')ds'}{\sigma}u\right] \\
&\quad + \gamma\int_{s'}\mathbb{E}_u[P(s'|s,\mu_{\theta+\sigma u}(s))]\nabla V^{\pi^{\theta,\sigma}}(s')ds' \qquad (35) \\
&= \mathbb{E}_u\left[\frac{Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))}{\sigma}u\right] + \gamma\int_{s'}\mathbb{E}_u[P(s'|s,\mu_{\theta+\sigma u}(s))]\nabla V^{\pi^{\theta,\sigma}}(s')ds' \\
&= \mathbb{E}_u\left[\frac{Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s)) - V^{\pi^{\theta,\sigma}}(s)}{\sigma}u\right] \\
&\quad + \gamma\int_{s'}\mathbb{E}_u[P(s'|s,\mu_{\theta+\sigma u}(s))]\nabla V^{\pi^{\theta,\sigma}}(s')ds' \qquad (36) \\
&= \mathbb{E}_u\left[\frac{Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s)) - V^{\pi^{\theta,\sigma}}(s)}{\sigma}u\right] \\
&\quad + \gamma\int_a\int_{s'}\pi^{\theta,\sigma}(a|s)P(s'|s,a)\nabla_\theta V^{\pi^{\theta,\sigma}}(s')ds'da
\end{aligned}$$

Where the lines 34 and 35 follow the property of the derivative of the guassian smoothing, and the line 36 uses the fact that the mean of u is zero. With a recursion as in Silver et al. (2014), we conclude the desired results.

A.3 Value-function as a baseline for variance reduction

Proposition 11. *Let $\theta \in \mathbb{R}^d$ be a policy parameter. For every baseline $b \in \mathcal{F}(\mathcal{S})$ and $u \in \mathbb{R}^d$, we define:*

$$F(b,u) = \frac{Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s)) - b^*(s)}{\sigma}u.$$

Let $b^* \in \mathcal{F}(\mathcal{S})$ such that

$$\mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}[\|F(b^*,u)\|^2] = \min_{b\in\mathcal{F}(\mathcal{S})}\mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}[\|F(b,u)\|^2].$$

Then, under assumptions 4,5 it holds:

$$\mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}[\|F(V^{\pi^{\theta,\sigma}},u)\|^2] = \min_{b\in\mathcal{F}(\mathcal{S})}\mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}[\|F(b,u)\|^2] + \mathcal{O}_{\sigma,d}(1). \qquad (37)$$

And under the additional assumptions 6,7 it holds:

$$\mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}[\|F(V^{\pi^{\theta,\sigma}},u)\|^2] = \min_{b\in\mathcal{F}(\mathcal{S})}\mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}},u\sim N(0,I)}[\|F(b,u)\|^2] + \mathcal{O}_{\sigma,d}(\sigma^2). \qquad (38)$$

We begin by establishing the expression of $b^*(s)$:

$$b^*(s) = \int_u Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\frac{\|u\|^2}{d}p(u)du \quad \text{a.e} \qquad (39)$$

In fact, we write:

$$\begin{aligned}
&\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))-b(s)\right)^2\|u\|^2\right]\\
&\quad=d\cdot\mathbb{E}_s\left[b^2(s)\right]-2\mathbb{E}_s\left[b(s)\mathbb{E}_s\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\|u\|^2\right]\right]\\
&\qquad+\mathbb{E}_{s,u}\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))^2\|u\|^2\right]\\
&\quad=\mathbb{E}_s\left[\left(\sqrt{d}\cdot b(s)-\frac{1}{\sqrt{d}}\mathbb{E}_u\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\|u\|^2\right]\right)^2\right]\\
&\qquad+\left(\mathbb{E}_{s,u}\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))^2\|u\|^2\right]-\frac{1}{d}\mathbb{E}_s\left[\left(\mathbb{E}_u\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\|u\|^2\right]\right)^2\right]\right) \quad (40)
\end{aligned}$$

The last expression attains its minimum when the first term is null, which implies that:

$$b^*(s)=\mathbb{E}_u\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\frac{\|u\|^2}{d}\right]\quad\text{a.e} \quad (41)$$

Now we prove that the two baselines b^* and $V^{\pi^{\theta,\sigma}}$ give the same order of variance reduction, we write:

$$\begin{aligned}
&\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s)-b^*(s))\right)^2\|u\|^2\right]\\
&\quad=\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))-V^{\pi^{\theta,\sigma}}(s)-\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)\right)^2\|u\|^2\right]\\
&\quad=\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))-V^{\pi^{\theta,\sigma}}(s)\right)^2\|u\|^2\right]+\mathbb{E}_{s,u}\left[\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)^2\|u\|^2\right]\\
&\qquad-2\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))-V^{\pi^{\theta,\sigma}}(s)\right)\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)\|u\|^2\right] \quad (42)
\end{aligned}$$

The last term can be written as:

$$\begin{aligned}
&\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))-V^{\pi^{\theta,\sigma}}(s)\right)\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)\|u\|^2\right]\\
&\quad=\mathbb{E}_s\left[\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)\left(\mathbb{E}_u\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\|u\|^2\right]-d\cdot V^{\pi^{\theta,\sigma}}(s)\right)\right]\\
&\quad=d\cdot\mathbb{E}_{s,u}\left[\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)^2\|u\|^2\right] \quad (43)
\end{aligned}$$

We then get:

$$\begin{aligned}
&\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s)-b^*(s))\right)^2\|u\|^2\right]\\
&\quad=\mathbb{E}_{s,u}\left[\left(Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))-V^{\pi^{\theta,\sigma}}(s)\right)^2\|u\|^2\right]-d\cdot\mathbb{E}_s\left[\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)^2\right] \quad (44)
\end{aligned}$$

Now it remains to upper bound the last term. In fact, by the definition of $b^*(s)$ and using the Jenson inequality, we get:

$$\begin{aligned}
\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)^2&=\left(\mathbb{E}_u\left[Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\left(\frac{\|u\|^2}{d}-1\right)\right]\right)^2\\
&=\frac{1}{d^2}\left(\mathbb{E}_u\left[\left\langle\nabla_u Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s)),u\right\rangle\right]\right)^2 \quad (45)
\end{aligned}$$

where the second equality holds true according to Remark 2, since the function $u\mapsto Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))$ is Lipschitz, which implies that $\|\nabla_u Q^{\pi^{\theta,\sigma}}(s,\mu_{\theta+\sigma u}(s))\|\leq L_{1,Q}L_\mu\sigma$ a.e. Thus, in the Lipschitz Regime, we have:

$$d\cdot\mathbb{E}_s\left[\left(b^*(s)-V^{\pi^{\theta,\sigma}}(s)\right)^2\right]\leq L_{1,Q}^2L_\mu^2 \quad (46)$$

In the smooth regime, the function $u \mapsto Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s))$ is $\sigma^2(L_{2,\mu}L_{1,Q} + L_{1,\mu}^2 L_{2,Q})$-smooth, thus $\|\nabla_u^2 Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s))\|$ is $\sigma^2(L_{2,\mu}L_{1,Q} + L_{1,\mu}^2 L_{2,Q})$-smooth. Therefore:

$$\begin{aligned}
\left(b^*(s) - V^{\pi^{\theta,\sigma}}(s)\right)^2 &= \left(\mathbb{E}_u\left[Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s))\left(\frac{\|u\|^2}{d} - 1\right)\right]\right)^2 \\
&= \frac{1}{d^2}\left(\mathbb{E}_u\left[\left\langle \nabla_u Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s)), u\right\rangle\right]\right)^2 \\
&= \frac{1}{d^2}\left(\mathrm{Tr}\left[\mathbb{E}_u\left[\nabla_u^2 Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s))\right]\right]\right)^2 \\
&\leq \frac{1}{d}\sigma^4(L_{2,\mu}L_{1,Q} + L_{1,\mu}^2 L_{2,Q}) \qquad (47)
\end{aligned}$$

B APPENDIX B

B.1 PROOF OF PROPOSITION 3

In order to establish the dicrepancy bounds, we use the performane difference lemma Kakade & Langford (2002):

$$\begin{aligned} J_\sigma(\theta) - J_D(\theta) &= \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}}} \mathbb{E}_{a\sim\pi^{\theta,\sigma}(\cdot|s)} \left[A^{\mu_\theta}(s,a)\right] \\ &= \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}}} \left[\int_a Q^{\mu_\theta}(s,a)\pi^{\theta,\sigma}(a|s)da - V^{\mu_\theta}(s)\right] \\ &= \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}}} \left[\int_u Q^{\mu_\theta}(s,\mu_{\theta+\sigma u}(s))p(u)du - Q^{\mu_\theta}(s,\mu_\theta(s))\right] \\ &= \mathbb{E}_{s\sim d^{\pi^{\theta,\sigma}}} \left[\int_u \left(Q^{\mu_\theta}(s,\mu_{\theta+\sigma u}(s)) - Q^{\mu_\theta}(s,\mu_\theta(s))\right)p(u)du\right] \end{aligned} \tag{48}$$

In the case of a non-smooth MDP we use the lipschitzness of the action-value function 12 and the lipschitzness of the policy $\mu_\theta(s)$ 5 which gives:

$$\begin{aligned} |J_\sigma(\theta) - J_D(\theta)| &\leq L_Q L_\mu \sqrt{d}\sigma \\ &= \left(L_r + \frac{\gamma R_{\max}}{1-\gamma} L_P\right) L_\mu \sqrt{d}\sigma \\ &= \mathcal{O}\left(\frac{\sqrt{d}\sigma}{1-\gamma}\right) \end{aligned} \tag{49}$$

In the case of a smooth MDP, we use the smoothness of the action-value function 14 and the smoothness of the policy $\mu_\theta(s)$ 7. First we write:

$$\begin{aligned} &Q^{\mu_\theta}(s,\mu_{\theta+\sigma u}(s)) - Q^{\mu_\theta}(s,\mu_\theta(s)) \\ &\quad= \Big(Q^{\mu_\theta}(s,\mu_{\theta+\sigma u}(s)) - Q^{\mu_\theta}(s,\mu_\theta(s)) \\ &\qquad - \left\langle \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}, \mu_{\theta+\sigma u}(s) - \mu_\theta(s)\right\rangle\Big) \\ &\qquad + \left\langle \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}, \mu_{\theta+\sigma u}(s) - \mu_\theta(s) - \sigma\nabla_\theta\mu_\theta(s)u\right\rangle \\ &\qquad + \sigma\left\langle \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}, \nabla_\theta\mu_\theta(s)u\right\rangle \end{aligned} \tag{50}$$

By taking the expectation over u first, and then usign the triangle inequality we get:

$$\begin{aligned} |J_\sigma(\theta) - J_D(\theta)| &\leq \frac{L_{2,Q}L_\mu^2 + L_Q L_{2,\mu}}{2} d\sigma^2 \\ &= \frac{1}{2}\left(L_{2,r}L_\mu^2 + L_r L_{2,\mu} + \frac{\gamma R_{\max}}{1-\gamma}(L_{2,P}L_\mu^2 + L_P L_{2,\mu})\right) d\sigma^2 \\ &= \mathcal{O}\left(\frac{d\sigma^2}{1-\gamma}\right) \end{aligned} \tag{51}$$

B.2 PROOF OF PROPOSITION 4

We recall that the deterministic policy gradient Silver et al. (2014) gives

$$\nabla_\theta J_D(\theta) = \mathbb{E}_{s\sim d^{\mu_\theta}}\left[\nabla\mu_\theta(s)\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\right] \tag{52}$$

We can write

$$\begin{aligned}
&\|\nabla J_\sigma(\theta) - \nabla J_D(\theta)\| \\
&= \left\| \int_s \left[\left(\int_u \frac{Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s))}{\sigma} up(u)du \right) d^{\pi^{\theta,\sigma}}(s) \right.\right. \\
&\quad \left.\left. - \nabla\mu_\theta(s)\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} d^{\mu_\theta}(s) \right] ds \right\| \\
&\leq \underbrace{\left\| \int_s \left[\left(\int_u \frac{Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s))}{\sigma} up(u)du \right) - \nabla\mu_\theta(s)\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \right] d^{\pi^{\theta,\sigma}}(s) ds \right\|}_{\text{Term 1}} \\
&\quad + \underbrace{\int_s \|\nabla\mu_\theta(s)\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\| \cdot \left| d^{\pi^{\theta,\sigma}}(s) - d^{\mu_\theta}(s) \right| ds}_{\text{Term 2}}
\end{aligned} \tag{53}$$

We first upper bound Term 1. Using the triangle inequality and the fact that $\mathbb{E}[u] = 0$, we get:

$$\begin{aligned}
\text{Term 1} \leq & \underbrace{\left\| \int_{s,u} \frac{Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s)) - Q^{\mu_\theta}(s, \mu_{\theta+\sigma u}(s))}{\sigma} up(u) d^{\pi^{\theta,\sigma}}(s) du ds \right\|}_{\text{Term 1.1}} \\
& + \underbrace{\left\| \int_s \left[\left(\int_u \frac{Q^{\mu_\theta}(s, \mu_{\theta+\sigma u}(s)) - Q^{\mu_\theta}(s, \mu_\theta(s))}{\sigma} up(u)du \right) - \nabla\mu_\theta(s)\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \right] d^{\pi^{\theta,\sigma}}(s) ds \right\|}_{\text{Term 1.2}}
\end{aligned} \tag{54}$$

For the Term 1.2, we observe that

$$\nabla\mu_\theta(s)\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} = \nabla_\varphi Q^{\mu_\theta}(s, \mu_\varphi(s))|_{\varphi=\theta} \tag{55}$$

We use the $L_{2,\varphi}$ smoothness of the mapping $\varphi \in \mathbb{R}^d \mapsto Q^{\mu_\theta}(s, \mu_\varphi(s))$, where $L_{2,\varphi} = L_{2,\mu}L_Q + L_\mu^2 L_{2,Q}$. On the other hand, it can proved that

$$\nabla_\varphi Q^{\mu_\theta}(s, \mu_\varphi(s))|_{\varphi=\theta} = \int_u \langle \nabla_\varphi Q^{\mu_\theta}(s, \mu_\varphi(s))|_{\varphi=\theta}, u \rangle \, up(u)du \tag{56}$$

Using this identity in Term 1.2, we get:

$$\begin{aligned}
\text{Term 1.2} &\leq \frac{1}{\sigma} \int_{s,u} \left| Q^{\mu_\theta}(s, \mu_{\theta+\sigma u}(s)) - Q^{\mu_\theta}(s, \mu_\theta(s)) \right. \\
&\qquad \left. - \sigma \langle \nabla_\varphi Q^{\mu_\theta}(s, \mu_\varphi(s))|_{\varphi=\theta}, u \rangle \right| \cdot \|u\| p(u) d^{\pi^{\theta,\sigma}}(s) du ds \\
&\leq \frac{L_{2,\mu}L_Q + L_\mu^2 L_{2,Q}}{2} \mathbb{E}_u[\|u\|^3]\sigma \\
&\leq \frac{L_{2,\mu}L_Q + L_\mu^2 L_{2,Q}}{2} (d+3)^{\frac{3}{2}}\sigma
\end{aligned} \tag{57}$$

Now we upper bound Term 1.1. With the Bellman equation, we can write:

$$\begin{aligned}
&Q^{\pi^{\theta,\sigma}}(s, \mu_{\theta+\sigma u}(s)) - Q^{\mu_\theta}(s, \mu_{\theta+\sigma u}(s)) \\
&\qquad = \mathbb{E}_{s' \sim P(\cdot|s, \mu_{\theta+\sigma u}(s))} \left[V^{\pi^{\theta,\sigma}}(s') \right] - V^{\mu_\theta}(s) \\
&\qquad = \int_{s'} \left[V^{\pi^{\theta,\sigma}}(s') - V^{\mu_\theta}(s') \right] P(s'|s, \mu_{\theta+\sigma u}(s)) ds'
\end{aligned} \tag{58}$$

Thus, the Term 1.1 measures exactly the discrepancy between the two objective where the initial distribution is $P(.|s, \mu_{\theta+\sigma u})$. Using the discrepancy bound in proposition **??**, we get:

$$\text{Term 1.1} \leq \frac{L_{2,\mu}L_Q + L_\mu L_{2,Q}}{2} d(d+1)^{\frac{1}{2}}\sigma \leq \frac{L_{2,\mu}L_Q + L_\mu^2 L_{2,Q}}{2} (d+3)^{\frac{3}{2}}\sigma \tag{59}$$

which gives:

$$\text{Term 1} \leq (L_{2,\mu} L_Q + L_\mu^2 L_{2,Q})(d+3)^{\frac{3}{2}} \sigma \tag{60}$$

For the upper bound of Term 2, we use the Lipschitzness of μ_θ, the action-value function and we apply lemma 20 to bound the total variation between $d^{\pi^{\theta,\sigma}}$ adn d^{μ_θ}, we get:

$$\text{Term 2} \leq L_\mu L_Q \gamma \frac{L_{2,P} L_\mu^2 + L_P L_{2,\mu} + L_P^2 L_\mu^2}{2(1-\gamma)} d\sigma^2 \tag{61}$$

C Appendix C

C.1 Proof of theorem 5

For $\theta_1, \theta_2 \in \mathbb{R}^d$, we have:

$$\begin{aligned}
&\sigma\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| \\
&= \left\| \int_{s,u} \left(A^{\pi^{\theta_1,\sigma}}(s, \mu_{\theta_1+\sigma u}(s)) d^{\pi^{\theta_1,\sigma}}(s) - A^{\pi^{\theta_2,\sigma}}(s, \mu_{\theta_2+\sigma u}(s)) d^{\pi^{\theta_2,\sigma}}(s) \right) up(u) du ds \right\| \\
&\leq \int_s \left\| \mathbb{E}_u \left[A^{\pi^{\theta_1,\sigma}}(s, \mu_{\theta_1+\sigma u}(s)) u \right] \right\| . \left| d^{\pi^{\theta_1,\sigma}}(s) - d^{\pi^{\theta_2,\sigma}}(s) \right| ds \\
&\quad + \int_s \left\| \mathbb{E}_u \left[\left(A^{\pi^{\theta_1,\sigma}}(s, \mu_{\theta_1+\sigma u}(s)) - A^{\pi^{\theta_1,\sigma}}(s, \mu_{\theta_2+\sigma u}(s)) \right) u \right] \right\| d^{\pi^{\theta_2,\sigma}}(s) ds \\
&\quad + \int_s \left\| \mathbb{E}_u \left[\left(A^{\pi^{\theta_1,\sigma}}(s, \mu_{\theta_2+\sigma u}(s)) - A^{\pi^{\theta_2,\sigma}}(s, \mu_{\theta_2+\sigma u}(s)) \right) u \right] \right\| d^{\pi^{\theta_2,\sigma}}(s) ds du \\
&\overset{(i)}{\leq} \int_s \left\| \mathbb{E}_u \left[A^{\pi^{\theta_1,\sigma}}(s, \mu_{\theta_1+\sigma u}(s)) u \right] \right\| . \left| d^{\pi^{\theta_1,\sigma}}(s) - d^{\pi^{\theta_2,\sigma}}(s) \right| ds \\
&\quad + \left(L_Q L_\mu + \gamma L_V \right) \|\theta_1 - \theta_2\| \qquad (62)
\end{aligned}$$

Where (i) follows from propositions 12 and 13 (see appendix E). For the first term,

$$\left\| \mathbb{E}_u \left[A^{\pi^{\theta_1,\sigma}}(s, \mu_{\theta+\sigma u}(s)) u \right] \right\| \leq \frac{R_{\max}}{1-\gamma} \sqrt{d} \qquad (63)$$

$$(64)$$

where we just upper bounded the action-value function by $\frac{R_{\max}}{1-\gamma}$, and we used $\mathbb{E}_u[V^{\pi^{\theta,\sigma}}(s)] = 0$. It remains to bound the total variation between the visitation distributions $d^{\pi^{\theta_1,\sigma}}$ and $d^{\pi^{\theta_2,\sigma}}$ given by $\int_s \left| d^{\pi^{\theta_1,\sigma}}(s) - d^{\pi^{\theta_2,\sigma}}(s) \right| ds$. In the literature, authors such as Xu et al. (2020) have bounded this quantity by invoking Theorem 3.1 from Mitrophanov (2005), an approach that strictly relies on the uniform ergodicity of the Markov chain 1. In contrast, we employ an alternative technique that circumvents the need for the ergodicity assumption, and gives a bound that does not depend on $\frac{1}{1-\rho}$, thereby bypassing the problem of slow-mixing Markov chains (where $\rho \to 1$). In fact, we directly use the infinite-series definition of the discounted visitation distribution $d^{\pi^{\theta,\sigma}}$, combined with the regularity of the environment and the policy. Lemma 18 gives:

$$\int_s |d^{\pi^{\theta_1,\sigma}}(s) - d^{\pi^{\theta_2,\sigma}}(s)| ds \leq \frac{\gamma L_P L_\mu}{1-\gamma} \|\theta_1 - \theta_2\| \qquad (65)$$

Using all the previous bounds, we get:

$$\begin{aligned}
\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| &\leq \frac{1}{\sigma} \left(\frac{\gamma L_P L_\mu \sqrt{d} \sigma}{(1-\gamma)^2} + L_Q L_\mu + \gamma L_V \right) \|\theta_1 - \theta_2\| \\
&= \mathcal{O}\left(\frac{d}{\sigma(1-\gamma)^2} \right) . \|\theta_1 - \theta_2\| \qquad (66)
\end{aligned}$$

C.2 Proof of theorem 6

Let $\theta_1, \theta_2 \in \mathbb{R}^d$. By the Deterministic Policy Gradient theorem and the property of Gaussian smoothing, the gradient of the objective is given by an expectation over the policy's specific state visitation distribution d^{π_θ}:

$$\nabla J_\sigma(\theta) = \int_{\mathcal{S}} d^{\pi_\theta}(s) g(\theta, s) ds \qquad (67)$$

where $g(\theta, s) = \mathbb{E}_u\Big[\nabla_\theta \mu_{\theta+\sigma u}(s)^\top \nabla_a Q^{\pi^{\theta,\sigma}}(s,a)\Big|_{a=\mu_{\theta+\sigma u}(s)}\Big]$.

The difference in gradients for two parameters θ_1, θ_2 is:

$$\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| = \|\int_{\mathcal{S}} d^{\pi_1}(s) g(\theta_1, s) ds - \int_{\mathcal{S}} d^{\pi_2}(s) g(\theta_2, s) ds\| \tag{68}$$

Adding and subtracting the cross-term $\int_{\mathcal{S}} d^{\pi_1}(s) g(\theta_2, s) ds$ inside the norm and applying the triangle inequality gives:

$$\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| \leq \int_{\mathcal{S}} d^{\pi_1}(s) \|g(\theta_1, s) - g(\theta_2, s)\| ds \tag{69}$$

$$+ \int_{\mathcal{S}} \big|d^{\pi_1}(s) - d^{\pi_2}(s)\big| \|g(\theta_2, s)\| ds \tag{70}$$

We bound the two terms separately. First, for term (69), we bound $\|g(\theta_1, s) - g(\theta_2, s)\|$ for any fixed state s. Using the triangle inequality, we add and subtract $\mathbb{E}_u\big[\nabla_\theta \mu_{\theta_1+\sigma u}(s)^\top \nabla_a Q^{\pi_2}(s, a_2)\big]$:

$$\begin{aligned}
\|g(\theta_1, s) - g(\theta_2, s)\| &\leq \mathbb{E}_u\Bigg[\|\nabla_\theta \mu_{\theta_1+\sigma u}(s)\| \|\nabla_a Q^{\pi_1}(s, a_1) - \nabla_a Q^{\pi_2}(s, a_2)\| \\
&\qquad + \|\nabla_a Q^{\pi_2}(s, a_2)\| \|\nabla_\theta \mu_{\theta_1+\sigma u}(s) - \nabla_\theta \mu_{\theta_2+\sigma u}(s)\|\Bigg] \\
&\leq \mathbb{E}_u\Big[\|\nabla_\theta \mu_{\theta_1+\sigma u}(s)\| \|\nabla_a Q^{\pi_1}(s, a_1) - \nabla_a Q^{\pi_1}(s, a_2)\|\Big] \\
&\qquad + \mathbb{E}_u\Bigg[\|\nabla_\theta \mu_{\theta_1+\sigma u}(s)\| \|\nabla_a Q^{\pi_1}(s, a_2) - \nabla_a Q^{\pi_2}(s, a_2)\| \\
&\qquad + \|\nabla_a Q^{\pi_2}(s, a_2)\| \|\nabla_\theta \mu_{\theta_1+\sigma u}(s) - \nabla_\theta \mu_{\theta_2+\sigma u}(s)\|\Bigg]
\end{aligned} \tag{71}$$

Thus, using the lipschitzness and smoothnes of the policy wrt its parameter and proposition **??** we get:

$$\|g(\theta_1, s) - g(\theta_2, s)\| \leq \big(L_\mu^2 L_{2,Q} + L_\mu L'_{2,Q} + L_Q L_{2,\mu}\big)\|\theta_1 - \theta_2\| \tag{72}$$

Integrating over $d^{\pi_1}(s)$ preserves the bound for term (69).

Second, for term (70), we bound the magnitude of the gradient $\|g(\theta_2, s)\| \leq L_\mu L_Q$ and we use lemma 18 to get

$$\int_{\mathcal{S}} \big|d^{\pi_1}(s) - d^{\pi_2}(s)\big| \|g(\theta_2, s)\| ds \leq \gamma \frac{L_P L_\mu^2 L_Q}{1-\gamma} \|\theta_1 - \theta_2\| \tag{73}$$

Finally, we get:

$$\|\nabla J_\sigma(\theta_1) - \nabla J_\sigma(\theta_2)\| \leq \Big(L_\mu^2 L_{2,Q} + L_\mu L'_{2,Q} + L_Q L_{2,\mu} + \gamma \frac{L_P L_\mu^2 L_Q}{1-\gamma}\Big)\|\theta_1 - \theta_2\| \tag{74}$$

Therefore, the objective $J_\sigma(\theta)$ is L_J-smooth with $L_J = G_\mu^2 L_{2,Q} + G_\mu L_Q + G_Q L_{2,\mu} + G_\mu G_Q L_d$.

D APPENDIX D

First, we make the following notations:

$$\delta_{t,i}(\omega_t, s_{t,i}, u_i, s_{t,i+1}) = r(s_{t,i}, \mu_{\theta_t+\sigma u_i}(s_{t,i})) + \gamma\varphi(s_{t,i+1})^T\omega_t - \varphi(s_{t,i})^T\omega_t \tag{75}$$

$$h(\omega_t) = \frac{1}{M}\sum_{i=0}^{M-1}\frac{\delta_{t,i}(\omega_t, s_{t,i}, u_i, s_{t,i+1})}{\sigma}u_i \tag{76}$$

$$\bar{h}(\omega_t) = \frac{1}{M}\sum_{i=0}^{M-1}\mathbb{E}\Big[\frac{\delta_{t,i}(\omega_t, s_{t,i}, u_i, s_{t,i+1})}{\sigma}u_i|\mathcal{F}_t\Big] \tag{77}$$

$$H(\omega_t) = \mathbb{E}_{s\sim d^{\pi^{\theta_t,\sigma}}, u\sim N(0,I), s'\sim P(.|s,\mu_{\theta_t+\sigma u}(s))}[\delta(\omega_t, s, u, s')] \tag{78}$$

D.1 PROOF OF LEMMA 7

For the moment, we assume that the variance $\mathbb{E}[\|h(\omega_t) - \bar{h}(\omega_t)\|^2|\mathcal{F}_t]$ is bounded by some constant $\sigma_M < \infty$. We prove this is actually true in Lemma 21, where we show that $\sigma_M = \mathcal{O}\Big(\frac{d}{\sigma^2 M}\Big)$.

Applying the smoothness inequality to J_σ, we have:

$$J_\sigma(\theta_{t+1}) \geq J_\sigma(\theta_t) + \xi^\theta\langle\nabla J_\sigma(\theta_t), h(\omega_t)\rangle - \frac{L_{2,\sigma}(\xi^\theta)^2}{2}\|h(\omega_t)\|^2 \tag{79}$$

Taking the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_t]$ on both sides:

$$\mathbb{E}[J_\sigma(\theta_{t+1})|\mathcal{F}_t] \geq J_\sigma(\theta_t) + \xi^\theta\langle\nabla J_\sigma(\theta_t), \bar{h}(\omega_t)\rangle - \frac{L_{2,\sigma}(\xi^\theta)^2}{2}\mathbb{E}\big[\|h(\omega_t)\|^2\big|\mathcal{F}_t\big] \tag{80}$$

We now apply the identity $\langle a, b\rangle = \|a\|^2 + \langle a, b-a\rangle \geq \frac{1}{2}\|a\|^2 - \frac{1}{2}\|b-a\|^2$ exclusively to the bias term:

$$\langle\nabla J_\sigma(\theta_t), \bar{h}(\omega_t)\rangle \geq \frac{1}{2}\|\nabla J_\sigma(\theta_t)\|^2 - \frac{1}{2}\|\bar{h}(\omega_t) - \nabla J_\sigma(\theta_t)\|^2 \tag{81}$$

$$\geq \frac{1}{2}\|\nabla J_\sigma(\theta_t)\|^2 - \|\bar{h}(\omega_t) - H(\omega_t)\|^2 - \|H(\omega_t) - \nabla J_\sigma(\theta_t)\|^2 \tag{82}$$

For the squared norm term, we decompose it into its expected value and variance, and apply the inequality $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$:

$$\begin{aligned}\mathbb{E}\big[\|h(\omega_t)\|^2\big|\mathcal{F}_t\big] &= \|\bar{h}(\omega_t)\|^2 + \sigma_M^2 \\ &\leq 3\|\nabla J_\sigma(\theta_t)\|^2 + 3\|\bar{h}(\omega_t) - H(\omega_t)\|^2 + 3\|H(\omega_t) - \nabla J_\sigma(\theta_t)\|^2 + \sigma_M^2\end{aligned} \tag{83}$$

Substituting equations (82) and (83) back into (80) and grouping the terms yields:

$$\begin{aligned}\mathbb{E}[J_\sigma(\theta_{t+1})|\mathcal{F}_t] \geq J_\sigma(\theta_t) &+ \left(\frac{\xi^\theta}{2} - \frac{3L_{2,\sigma}(\xi^\theta)^2}{2}\right)\|\nabla J_\sigma(\theta_t)\|^2 \\ &- \left(\xi^\theta + \frac{3L_{2,\sigma}(\xi^\theta)^2}{2}\right)\|\bar{h}(\omega_t) - H(\omega_t)\|^2 \\ &- \left(\xi^\theta + \frac{3L_{2,\sigma}(\xi^\theta)^2}{2}\right)\|H(\omega_t) - \nabla J_\sigma(\theta_t)\|^2 - \frac{L_{2,\sigma}(\xi^\theta)^2}{2}\sigma_M^2\end{aligned} \tag{84}$$

By setting $\xi^\theta \leq \frac{1}{6L_{2,\sigma}}$, we get:

$$\begin{aligned}\mathbb{E}[J_\sigma(\theta_{t+1})|\mathcal{F}_t] \geq J_\sigma(\theta_t) &+ \frac{\xi^\theta}{4}\|\nabla J_\sigma(\theta_t)\|^2 - \frac{5\xi^\theta}{4}\underbrace{\|\bar{h}(\omega_t) - H(\omega_t)\|^2}_{\text{Term 1}} \\ &- \frac{5\xi^\theta}{4}\underbrace{\|H(\omega_t) - \nabla J_\sigma(\theta_t)\|^2}_{\text{Term 2}} - \frac{L_{2,\sigma}(\xi^\theta)^2}{2}\sigma_M^2\end{aligned} \tag{85}$$

D.2 PROOF OF THEOREM 8

D.2.1 ANALYSIS OF THE ACTOR: FIRST INEQUALITY

Following the proof of the previous lemma, we will bound each term of the RHS of inequality 85 separetly.

Upper bound of Term 1: We define

$$g_{t,i}(\omega_t) = \frac{\delta(\omega_t, s_{t,i}, u_i, s_{t,i+1})}{\sigma} u_i \tag{86}$$

which gives $\bar{h}(\omega_t) = \frac{1}{M}\sum_{i=0}^{M-1} \mathbb{E}[g_{t,i}|\mathcal{F}_t]$ so we can write:

$$g_{t,i}(\omega_t) = g_{t,i}(\omega^*_{\theta_t}) + (\gamma\varphi(s_{t,i+1}) - \varphi(s_{t,i}))^T(\omega_t - \omega^*_{\theta_t})\frac{u_i}{\sigma} \tag{87}$$

In order to apply lemma 22, we will upper bound each term in the RHS of the inequality:

$$\|\mathbb{E}[g_{t,i}(\omega_t)|s_{t,i}]\|^2 \leq 2\|\mathbb{E}[g_{t,i}(\omega^*_{\theta_t})|s_{t,i}]\|^2 + \|\mathbb{E}[(\gamma\varphi(s_{t,i+1}) - \varphi(s_{t,i}))^T(\omega_t - \omega^*_{\theta_t})\frac{u_i}{\sigma}|s_{t,i}]\|^2 \tag{88}$$

We begin with the second term. The expectation can be written as:

$$\begin{aligned}\mathbb{E}\big[(\gamma\varphi(s_{t,i+1}) - \varphi(s_{t,i}))^T(\omega_t - \omega^*_{\theta_t})\frac{u_i}{\sigma}|s_{t,i}\big] &= \int_{s_{t,i+1}} (\gamma\varphi(s_{t,i+1}) - \varphi(s_{t,i}))^T(\omega_t - \omega^*_{\theta_t}) \\ &\quad \mathbb{E}_{u_i}[\frac{P(s_{t,i+1}|s_{t,i}, \mu_{\theta_t+\sigma u_i}(s_{t,i}))}{\sigma} u_i] ds_{t,i+1} \quad (89) \\ &= \int_{s_{t,i+1}} (\gamma\varphi(s_{t,i+1}) - \varphi(s_{t,i}))^T(\omega_t - \omega^*_{\theta_t}) \\ &\quad \mathbb{E}_{u_i}[\nabla_\theta P(s_{t,i+1}|s_{t,i}, \mu_{\theta_t+\sigma u_i}(s_{t,i}))] ds_{t,i+1} \quad (90)\end{aligned}$$

Now using lemma 16 we get:

$$\Big(\mathbb{E}[(\gamma\varphi(s_{t,i+1}) - \varphi(s_{t,i}))^T(\omega_t - \omega^*_{\theta_t})\frac{u_i}{\sigma}|s_{t,i}]\Big)^2 \leq \Phi^2_{\max} L_P^2 L_\mu^2 \frac{d}{\sigma^2}\|\omega_t - \omega^*_{\theta_t}\|^2 \tag{91}$$

For the first term, we use the Lipschitzness of $r(s, \mu_{\theta+\sigma u}(s))$ and ω^*_θ 23 to get:

$$\big(\mathbb{E}[g_{t,i}(\omega^*_{\theta_t})|s_{t,i}]\big)^2 \leq L_\mu L_r + 8R_{\max}\Phi^3_{\max}\frac{\gamma L_P L_\mu}{1-\gamma} \tag{92}$$

Using lemma 22, we get

$$\|\bar{h}(\omega_t) - H(\omega_t)\|^2 \leq \frac{C_1}{M} + \frac{dC_2}{\sigma^2 M}\|\omega_t - \omega^*_{\theta_t}\| \tag{93}$$

where $C_1 = C_\rho(L_\mu L_r + 8R_{\max}\Phi^3_{\max}\frac{\gamma L_P L_\mu}{1-\gamma})$ and $C_1' = C_\rho \Phi^2_{\max} L_P^2 L_\mu^2$. Let $C_2 = C_1' + 2\gamma^2\Phi^2_{\max}L_P^2L_\mu^2$

Upper bound of Term 2: We begin by writing:

$$\begin{aligned}H(\omega_t) - \nabla J_\sigma(\theta_t) &= \mathbb{E}_{s\sim d^{\pi^{\theta_t,\sigma}},u,s'}\left[\frac{\delta(\omega_t, s, u, s') - A^{\pi^{\theta_t,\sigma}}(s, \mu_{\theta_t+\sigma u}(s))}{\sigma} u\right] \\ &= \mathbb{E}_{s\sim d^{\pi^{\theta_t,\sigma}},u,s'}\left[\frac{\gamma\varphi(s')^T\omega_t - \varphi(s)^T\omega_t - \gamma V^{\pi^{\theta_t,\sigma}}(s') + V^{\pi^{\theta_t,\sigma}}(s)}{\sigma} u\right] \\ &= \gamma\mathbb{E}_{s\sim d^{\pi^{\theta_t,\sigma}},u,s'}\left[\frac{\varphi(s')^T\omega_t - V^{\pi^{\theta_t,\sigma}}(s')}{\sigma} u\right] \\ &= \gamma\mathbb{E}_{s\sim d^{\pi^{\theta_t,\sigma}},u,s'}\left[\frac{\varphi(s')^T(\omega_t - \omega^*_{\theta_t})}{\sigma} u\right] \\ &\quad + \gamma\mathbb{E}_{s\sim d^{\pi^{\theta_t,\sigma}},u,s'}\left[\frac{\varphi(s')^T\omega^*_{\theta_t} - V^{\pi^{\theta_t,\sigma}}(s')}{\sigma} u\right] \quad (94)\end{aligned}$$

Now, using the well-known result stated in remark 2, and we get:

$$H(\omega_t) - \nabla J_\sigma(\theta_t) = \gamma \mathbb{E}_{s\sim d^{\pi_{\theta_t,\sigma}}, u\sim N(0,I)}\Big[\int_{s'} \varphi(s')^T(\omega_t - \omega^*_{\theta_t})\nabla P(s'|s,\mu_{\theta_t+\sigma u}(s))\Big] + \gamma \mathbb{E}_{s\sim d^{\pi_{\theta_t,\sigma}}, u\sim N(0,I)}\Big[\int_{s'} \big(\varphi(s')^T\omega^*_{\theta_t} - V^{\pi_{\theta_t,\sigma}}(s')\big)\nabla P(s'|s,\mu_{\theta_t+\sigma u}(s))\Big] \tag{95}$$

By applying lemma 16 (resp. 17 and assumption 3) to upper bound the first term (resp. the second term) in 95, we get:

$$\mathbb{E}[\|H(\omega_t) - \nabla J_\sigma(\theta_t)\|^2|\mathcal{F}t] \leq 2\gamma^2\Phi^2_{\max}L^2_P L^2_\mu \|\omega_t - \omega^*_{\theta_t}\|^2 + 2\gamma\Phi^2_{\max}L^2_P L^2_\mu \Xi_{critic} \tag{96}$$

By setting:

$$\mathbf{J}_T = \frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}[\|\nabla J_\sigma(\theta_i)\|^2] \tag{97}$$

$$Z_T = \frac{1}{T}\sum_{i=0}^{T-1}\mathbb{E}[\|\omega_t - \omega_{\theta_t}\|^2] \tag{98}$$

and using $M \geq 1$

$$\mathbf{J}_T \leq 4\frac{J^*_\sigma - J_\sigma(\theta_0)}{\xi^\theta T} + \frac{dC_2}{\sigma^2}Z_T + \frac{C_1}{M} + L_{2,\sigma}\frac{\xi^\theta}{2}\sigma^2_M + \mathcal{O}(\Xi_{critic}) \tag{99}$$

and also, we get the following inequality that we will use in the analysis of the critic:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|h(\omega)\|^2] \leq 3\mathbf{J}_T + \frac{3C_1}{M} + \frac{3dC_2}{\sigma^2}Z_T + \sigma^2_M + \mathcal{O}(\Xi_{critic}) \tag{100}$$

D.2.2 Analysis of the critic: Second inequality

We make some notations:

$$O_{t,i} = (s_{t,i}, a_{t,i}, s_{t,i+1}) \tag{101}$$

$$z_t = \omega_t - \omega^*_t \tag{102}$$

$$g_t(O_{t,i},\omega_t) = \varphi(s_{t,i})\big[r_{t,i} + \gamma\varphi(s_{t,i+1})^T\omega_t - \varphi(s_{t,i})^T\omega_t\big] \tag{103}$$

$$g_t(\omega_t) = \frac{1}{M}\sum_{i=0}^{M-1} g_t(O_{t,i},\omega_t) \tag{104}$$

$$\bar{g}_t(\omega_t) = \mathbb{E}_{s\sim d^{\pi_{\theta,\sigma}}, u\sim N(0,I), s'\sim P(s,\mu_{\theta_t+\sigma u}(s))}[g_t(O,\omega_t)] \tag{105}$$

$$\Lambda_t = \langle z_t, g_t(\omega_t) - \bar{g}_t(\omega_t)\rangle \tag{106}$$

Using the update rule for critic in algorithm 1 and the inequality $\|a-b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we can write:

$$\|z_{t+1}\|^2 = \|\Pi_{R_\omega}[\omega_t + \xi^\omega g_t(\omega_t)] - \omega^*_{t+1}\|^2 \tag{107}$$

$$\leq \|\omega_t + \xi^\omega g_t(\omega_t) - \omega^*_{t+1}\|^2 \tag{108}$$

$$= \|z_t + \xi^\omega g_t(\omega_t) + (\omega^*_t - \omega^*_{t+1})\|^2 \tag{109}$$

$$= \|z_t\|^2 + 2\xi^\omega\langle z_t, g_t(\omega_t)\rangle + 2\omega\langle z_t, \omega^*_t - \omega^*_{t+1}\rangle + \|\xi^\omega g_t(\omega_t) + (\omega^*_t - \omega^*_{t+1})\|^2 \tag{110}$$

$$\leq \|z_t\|^2 + 2\xi^\omega\langle z_t, \bar{g}_t(\omega_t)\rangle + 2\xi^\omega\Lambda_t + \langle z_t, \omega^*_t - \omega^*_{t+1}\rangle + 2(\xi^\omega)^2\|g_t(\omega_t)\|^2 + 2\|\omega^*_t - \omega^*_{t+1}\|^2 \tag{111}$$

Taking the expectation conditionned on $\mathcal{F}_t$, we get:

$$\mathbb{E}[\|z_{t+1}\|^2|\mathcal{F}_t] \leq \|z_t\|^2 + 2\xi^\omega\mathbb{E}[\langle z_t, \bar{g}_t(\omega_t)\rangle|\mathcal{F}_t] + 2\xi^\omega\mathbb{E}[\Lambda_t|\mathcal{F}_t] + \mathbb{E}[\langle z_t, \omega^*_t - \omega^*_{t+1}\rangle|\mathcal{F}_t] + 2(\xi^\omega)^2\mathbb{E}[\|g_t(\omega_t)\|^2|\mathcal{F}_t] + 2\mathbb{E}[\|\omega^*_t - \omega^*_{t+1}\|^2|\mathcal{F}_t] \tag{112}$$

For the second term, we observe that we can rewrite $\bar{g}_t(\omega_t)$ as $A\omega_t + b$ which gives $\bar{g}(\omega_t^*) = 0$. Then:

$$\langle z_t, \bar{g}_t(\omega_t)\rangle = \langle z_t, \bar{g}_t(\omega_t) - \bar{g}_t(\omega_t^*)\rangle \tag{113}$$
$$= \langle z_t, Az_t\rangle \tag{114}$$
$$\leq -\lambda\|z_t\|^2 \tag{115}$$

To upper bound the third term, which represents the bias introduced by Markovian sampling, we invoke Assumption 3. Because our two time-scale approach evaluates a frozen policy over a batch of size M, we bypass the need for the auxiliary Markov chain construction used in standard single-step critic analyses Xu et al. (2019).

Conditioning on the filtration $\mathcal{F}_t$, we expand the bias:

$$|\mathbb{E}[\Lambda_t \mid \mathcal{F}_t]| = \left|\left\langle z_t, \mathbb{E}[g_t(\omega_t) \mid \mathcal{F}_t] - \bar{g}(\omega_t, \theta_t)\right\rangle\right| \tag{116}$$
$$= \left|\left\langle z_t, \frac{1}{M}\sum_{i=0}^{M-1}\int_{\mathcal{S}} \mathbb{E}[g(O_{t,i}, \omega_t) \mid s_{t,i} = s]\left(P_{t,i}^{\theta_t,\sigma}(s) - d^{\pi^{\theta_t,\sigma}}(s)\right) ds\right\rangle\right| \tag{117}$$
$$\leq ||z_t||\frac{1}{M}\sum_{i=0}^{M-1}\sup_s\left|\left|\mathbb{E}[g(O_{t,i}, \omega_t) \mid s]\right|\right|\int_s\left|P_{t,i}^{\theta_t,\sigma}(s) - d^{\pi^{\theta_t,\sigma}}(s)\right| ds \tag{118}$$

By the projection operator, $||z_t|| \leq 2R_\omega$. The expected gradient magnitude is uniformly bounded by $\Delta_{\max}\Phi_{\max}$, where $\Delta_{\max} = R_{\max} + (1+\gamma)\Phi_{\max}R_\omega$. Applying the Total Variation distance bound from the uniform ergodicity assumption 1 reduces the integral to $\kappa\rho^i$:

$$|\mathbb{E}[\Lambda_t \mid \mathcal{F}_t]| \leq 2R_\omega\Delta_{\max}\Phi_{\max}\frac{\kappa}{M}\sum_{i=0}^{M-1}\rho^i \tag{119}$$
$$\leq \frac{2R_\omega\Delta_{\max}\Phi_{\max}\kappa}{M(1-\rho)} \tag{120}$$

For the forth and the last term, we use lemma 23 and the update of tyhe actor:

$$\mathbb{E}[\langle z_t, \omega_t^* - \omega_{t+1}^*\rangle \,|\mathcal{F}_t] \leq 8R_{\max}\Phi_{\max}^3\frac{\gamma L_p L_\mu}{1-\gamma}\xi^\theta\|z_t\|\mathbb{E}[\|h(\omega_t)\||\mathcal{F}_t] \tag{121}$$

and

$$\mathbb{E}[\|\omega_t^* - \omega_{t+1}^*\|^2|\mathcal{F}_t] \leq 16R_{\max}^2\phi_{\max}^6\frac{\gamma^2L_P^2L_\mu^2}{(1-\gamma)^2}\mathbb{E}[\|h(\omega_t)\|^2|\mathcal{F}_t] \tag{122}$$

To upper bound the fifth term, we start by decomposing $g_t(\omega_t)$ into its stationary expectation $\bar{g}_t(\omega_t)$ and the finite-sample Markovian noise. Using the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we have:

$$\begin{aligned}\mathbb{E}\big[\|g_t(\omega_t)\|^2 \mid \mathcal{F}_t\big] &= \mathbb{E}\left[\left\|\bar{g}_t(\omega_t) + \frac{1}{M}\sum_{i=0}^{M-1}\big(g_t(O_{t,i},\omega_t) - \bar{g}_t(\omega_t)\big)\right\|^2 \,\middle|\, \mathcal{F}_t\right] \\ &\leq 2\|\bar{g}_t(\omega_t)\|^2 + 2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=0}^{M-1}e_{t,i}\right\|^2 \,\middle|\, \mathcal{F}_t\right]\end{aligned} \tag{123}$$

where $e_{t,i} = g_t(O_{t,i},\omega_t) - \bar{g}_t(\omega_t)$ represents the noise of the i-th sample relative to the stationary gradient.

For the first term, because the features and rewards are uniformly bounded, the expected gradient $\bar{g}_t(\cdot)$ is L_g-Lipschitz continuou, which gives:

$$\|\bar{g}_t(\omega_t) - \bar{g}_t(\omega_t^*)\| \leq L_g\|\omega_t - \omega_t^*\| \tag{124}$$

Since $\bar{g}_t(\omega_t^*) = 0$ and $z_t = \omega_t - \omega_t^*$, this simplifies to $\|\bar{g}_t(\omega_t)\| \leq L_g\|z_t\|$. Thus

$$2\|\bar{g}_t(\omega_t)\|^2 \leq 2L_g^2\|z_t\|^2 \tag{125}$$

For the second term, we first upper bound $\mathbb{E}[\|e_{t,i}\|^2|s] \leq G_{max} = \Phi_{max}(R_{\max} + (1+\gamma)\mathbb{R}_\omega \Phi_{\max})$. Then, under assumption 1, we get:

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=0}^{M-1} e_{t,i}\right\|^2 \middle| \mathcal{F}_t\right] \leq \frac{2G_{\max}}{M} + \frac{2G_{\max}}{M^2}\sum_{i>j}\kappa\rho^{i-j} \tag{126}$$

$$\leq \frac{2G_{\max}}{M} + \frac{2G_{\max}}{(1-\rho)M} \tag{127}$$

Therefore:

$$\mathbb{E}\big[\|g_t(\omega_t)\|^2 \mid \mathcal{F}_t\big] \leq 2L_g^2\|z_t\|^2 + \frac{2K}{M} \tag{128}$$

with $K = 2G_{\max}(1 + \frac{1}{1-\rho})$, and $L_g = 2\Phi_{\max}^2$

Gathering all the previous inequalities, and setting $\xi^\omega \leq \min\{\frac{\lambda}{2L_g^2}, \frac{1}{\lambda}\}$ we get:

$$\mathbb{E}[\|z_{t+1}\|^2] \leq \big(1 - \frac{\lambda\xi^\omega}{2}\big)\mathbb{E}[\|z_t\|^2] + A_1\frac{\xi^\theta}{1-\gamma}\sqrt{\mathbb{E}[\|z_t\|^2]}\sqrt{\mathbb{E}[\|h(\omega_t)\|^2]}$$
$$+ A_2\Big(\frac{\xi^\theta}{1-\gamma}\Big)\mathbb{E}[\|h(\omega_t)\|^2] + \frac{A_3\xi^\omega}{M} + \frac{2K}{M}(\xi^\omega)^2 \tag{129}$$

where

$$A_1 = 8R_{\max}\Phi_{\max}^3\gamma L_p L_\mu \tag{130}$$

$$A_2 = 32R_{\max}^2\phi_{\max}^6\gamma^2 L_P^2 L_\mu^2 \tag{131}$$

$$A_3 = \frac{4R_\omega\Delta_{\max}\Phi_{\max}\kappa}{1-\rho} \tag{132}$$

In order to upper bound Z_T, we will use lemma 24 which gives:

$$Z_T \leq \frac{8R_\omega}{\lambda\xi^\omega T} + 2\beta(\xi^\theta,\xi^\omega)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|h(\omega_t)\|^2] + \frac{8A_3}{\lambda M} + \frac{8K}{\lambda M}\xi^\omega \tag{133}$$

On the other hand, plugging the inequality 100 into 133 yields:

$$Z_T \leq \frac{8R_\omega}{\lambda\xi^\omega T} + 2\beta(\xi^\theta,\xi^\omega)\Big(3\mathbf{J}_T + \frac{3C_1}{M} + 3\frac{dC_2}{\sigma^2}Z_T + \sigma_M^2 + \mathcal{O}(\Xi_{critic})\Big)$$
$$+ \frac{8A_3}{\lambda M} + \frac{8K}{\lambda M}\xi^\omega \tag{134}$$

$$\big(1 - 6\beta(\xi^\theta,\xi^\omega)\frac{dC_2}{\sigma^2}\big)Z_T \leq \frac{8R_\omega}{\lambda\xi^\omega T} + 6\beta(\xi^\theta,\xi^\omega)\mathbf{J}_T + \frac{6C_1\beta(\xi^\theta,\xi^\omega) + 8A_3/\lambda}{M}$$
$$+ 2\beta(\xi^\theta,\xi^\omega)\sigma_M^2 + \frac{8K}{\lambda M}\xi^\omega + \mathcal{O}(\Xi_{critic}) \tag{135}$$

Let $\xi^\theta << \xi^\omega$, so that $\beta(\xi^\theta,\xi^\omega) \leq \frac{\sigma^2}{12dC_2} = \frac{\sigma^2}{dC}$. Such step sizes exists because the descriminant of the quadratic formula $\beta(\xi^\theta,\xi^\omega) - \frac{\sigma^2}{6dC_2}$ is positive. Assuming that the exploration scale $\sigma^2 < d$, the last inequality simplifies to:

$$Z_T \leq \frac{16R_\omega}{\lambda\xi^\omega T} + 12\beta(\xi^\theta,\xi^\omega)\mathbf{J}_T + \frac{C_1/C_2 + 16A_3/\lambda + 16K\xi^\omega/\lambda}{M} + 2\beta(\xi^\theta,\xi^\omega)\sigma_M^2 + \mathcal{O}(\Xi_{critic}) \tag{136}$$

D.3 PROOF OF THEOREM 9

To derive the overall convergence rate, we couple the bounds of theorem 8. Therfore:

$$\begin{aligned}(1 - 12C_2\beta(\xi^\theta, \xi^\omega))\mathbf{J}_T \leq \frac{1}{T}\left(\frac{4(J_\sigma^* - J_\sigma(\theta_0))}{\xi^\theta} + \frac{8C_2R_\omega}{\lambda\xi^\omega}\right) \\ + \frac{2C_1 + 16C_2A_3/\lambda + 16C_2K\xi^\omega/\lambda}{M} \\ + \left(L_{2,\sigma}\frac{\xi^\theta}{2} + C_2\beta(\xi^\theta, \xi^\omega)\right)\sigma_M^2 \\ + \mathcal{O}(\Xi_{critic})\end{aligned} \tag{137}$$

We set the actor step size $\xi^\theta << \xi^\omega$ so that $\beta(\xi^\theta, \xi^\omega) \leq \frac{\sigma^2}{dC}$, with $C = 24C_2$. We set $M > \frac{1}{\epsilon}$, and $T > \frac{1}{\xi^\theta\epsilon}$. It remains to choose the actor step size ξ^θ so that

$$C_2\beta(\xi^\theta, \xi^\omega) + L_{2,\sigma}\frac{\xi^\theta}{2} \leq \frac{\sigma^2}{d}\tilde{C} \tag{138}$$

for some MDP-dependent constante $\tilde{C}$. Therefore, ξ^θ must verifies:

$$\left(\frac{4C_2}{\lambda^2(1-\gamma)^2}\right)(\xi^\theta)^2 + \left(\frac{4C_2\xi^\omega}{\lambda(1-\gamma)} + \frac{L_{2,\sigma}(\xi^\omega)^2}{2}\right)\xi^\theta - \frac{\sigma^2(\xi^\omega)^2}{d} \leq 0 \tag{139}$$

which implies directly that $\xi^\theta = \mathcal{O}\Big(\frac{(1-\gamma)^2\sigma^2}{L_{2,\sigma}}\Big)$. It remains to substitute the formula of $L_{2,\sigma}$ for both regimes.

D.4 PROOF OF CORROLARY 10

Straightforward, since the difference in norme between $\nabla J_\sigma(\theta)$ and $\nabla J_D(\theta)$ is of order $\frac{d^{1.5}\sigma}{1-\gamma}$, it is sufficient to set $\sigma = \mathcal{O}(\frac{(1-\gamma)\sqrt{\epsilon}}{d^{1.5}})$.

E LIPSCHITZNESS AND SMOOTHNESS OF THE VALUE/ACTION-VALUE FUNCTIONS

Proposition 12 (Lipschitzness of the action-value function)**.** *Under assumptions 4 and 5, the action value function $Q^{\pi_{\theta,\sigma}}$ is L_Q-Lipschitz wrt the action variable:*

$$|Q^{\pi_{\theta,\sigma}}(s,a_1) - Q^{\pi_{\theta,\sigma}}(s,a_2)| \leq \left(L_r + \frac{\gamma R_{\max}}{1-\gamma} L_P\right)\|a_1 - a_2\| \tag{140}$$

for all $a_1, a_2 \in \mathcal{A}$, $s \in \mathcal{S}$ and $\theta \in \mathbb{R}^d$. $L_Q = L_r + \frac{\gamma R_{\max}}{1-\gamma} L_P$

Proof. Let $\theta \in \mathbb{R}^d$, $s \in \mathcal{S}$ and $a_1, a_2 \in \mathcal{A}$:

$$|Q^{\pi_{\theta,\sigma}}(s,a_1) - Q^{\pi_{\theta,\sigma}}(s,a_1)| \leq |r(s,a_1) - r(s,a_2)| + \gamma \int_{s'} \left| \left(P(s'|s,a_1) - P(s'|s,a_2)\right) V^{\pi_{\theta,\sigma}}(s') \right| ds' \tag{141}$$

$$\leq L_r \|a_1 - a_2\| + \gamma \frac{R_{\max}}{1-\gamma} \int_{s'} \left| P(s'|s,a_1) - P(s'|s,a_2) \right| ds' \tag{142}$$

$$\leq \left(L_r + \frac{\gamma R_{\max}}{1-\gamma} L_P\right)\|a_1 - a_2\| \tag{143}$$

□

Proposition 13 (Lipschitzness of the value function)**.** *Under assumptions 4 and 5, the value function $V^{\pi_{\theta,\sigma}}$ is L_V-Lipschitz wrt the parameter θ:*

$$|V^{\pi_{\theta_1,\sigma}}(s) - V^{\pi_{\theta_2,\sigma}}(s)| \leq \frac{1}{1-\gamma}\left(L_r + \frac{\gamma R_{\max}}{1-\gamma} L_P\right) L_\mu \|\theta_1 - \theta_2\| \tag{144}$$

for all $s \in \mathcal{S}$ and $\theta_1, \theta_2 \in \mathbb{R}^d$. $L_V = \left(L_r + \frac{\gamma R_{\max}}{1-\gamma} L_P\right) L_\mu = L_Q L_\mu$

Proof. We use the performance difference lemma Kakade & Langford (2002):

$$(1-\gamma)|V^{\pi_{\theta_1,\sigma}}(s) - V^{\pi_{\theta_2,\sigma}}(s)| \leq \mathbb{E}_{s\sim d^{\pi_{\theta_1,\sigma}}}\left[\left| \int_a Q^{\pi_{\theta_2,\sigma}}(s,a)\pi^{\theta_1,\sigma}(a|s)da - V^{\pi_{\theta_2,\sigma}}(s) \right|\right] \tag{145}$$

$$\leq \mathbb{E}_{s\sim d^{\pi_{\theta_1,\sigma}}}\left[\int_u \left| Q^{\pi_{\theta_2,\sigma}}(s,\mu_{\theta_1+\sigma u}(s)) - Q^{\pi_{\theta_2,\sigma}}(s,\mu_{\theta_2+\sigma u}(s)) \right| p(u)du\right] \tag{146}$$

$$\leq L_Q L_\mu \|\theta_1 - \theta_2\| \tag{147}$$

$$= \left(L_r + \frac{\gamma R_{\max}}{1-\gamma} L_P\right) L_\mu \|\theta_1 - \theta_2\| \tag{148}$$

□

Proposition 14 (Smoothness of the action-value function)**.** *Under assumptions 6 and 7, $Q^{\pi_{\theta,\sigma}}$ is $L_{2,Q}$-smooth wrt the action variable:*

$$\|\nabla_a Q^{\pi_{\theta,\sigma}}(s,a)|_{a=a_1} - \nabla_a Q^{\pi_{\theta,\sigma}}(s,a)|_{a=a_2}\| \leq \left(L_{2,r} + \gamma \frac{R_{\max}(L_{2,P} + L_P^2)}{1-\gamma}\right)\|a_1 - a_2\| \tag{149}$$

for all $s \in \mathcal{S}$, $\theta \in \mathbb{R}^d$ and $a_1, a_2 \in \mathcal{A}$. $L_{2,Q} = L_{2,r} + \gamma \frac{R_{\max}(L_{2,P} + L_P^2)}{1-\gamma}$

Proof. By definition, the action-value function satisfies the Bellman equation:

$$Q^{\pi^{\theta,\sigma}}(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(s'|s,a) V^{\pi^{\theta,\sigma}}(s') ds' \tag{150}$$

Under Assumption 6, both the reward function and the transition kernel are continuously differentiable. Taking the gradient with respect to the action variable a yields:

$$\nabla_a Q^{\pi^{\theta,\sigma}}(s,a) = \nabla_a r(s,a) + \gamma \int_{\mathcal{S}} \nabla_a P(s'|s,a) V^{\pi^{\theta,\sigma}}(s') ds' \tag{151}$$

Let $a_1, a_2 \in \mathcal{A}$. Using the triangle inequality, we have

$$\begin{aligned}
\|\nabla_a Q^{\pi^{\theta,\sigma}}(s,a_1) - \nabla_a Q^{\pi^{\theta,\sigma}}(s,a_2)\| = \Bigg\| & \nabla_a r(s,a_1) - \nabla_a r(s,a_2) \\
& + \gamma \int_{\mathcal{S}} \Big(\nabla_a P(s'|s,a_1) - \nabla_a P(s'|s,a_2)\Big) V^{\pi^{\theta,\sigma}}(s') ds' \Bigg\| \\
\leq \|\nabla_a r(s,a_1) - \nabla_a r(s,a_2)\| & \\
+ \frac{R_{\max}}{1-\gamma} \int_{s'} \|\nabla_a P(s'|s,a_1) - \nabla_a P(s'|s,a_2)\| ds' &
\end{aligned}$$

Where in the last inequality we used $\|V^{\pi^{\theta,\sigma}}(s)\|_\infty \leq \frac{R_{\max}}{1-\gamma}$. Now, we apply lemma 19, to to bound the second term and we get:

$$\|\nabla_a Q^{\pi^{\theta,\sigma}}(s,a_1) - \nabla_a Q^{\pi^{\theta,\sigma}}(s,a_2)\| \leq \Big(L_{2,r} + \gamma \frac{R_{\max}(L_{2,P} + L_P^2)}{1-\gamma}\Big) \|a_1 - a_2\| \tag{152}$$

□

Proposition 15 (Lipschitzness of $\nabla_a Q^{\pi^{\theta_1,\sigma}}$ wrt θ)**.** *$\nabla_a Q^{\pi^{\theta,\sigma}}$ is $L'_{2,Q}$-Lipschitz wrt the policy parameter θ:*

$$\|\nabla_a Q^{\pi^{\theta_1,\sigma}}(s,a) - \nabla_a Q^{\pi^{\theta_2,\sigma}}(s,a)\| \leq \tag{153}$$

for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $\theta_1, \theta_2 \in \mathbb{R}^d$. $L'_{2,Q} = L_Q L_\mu L_P$

Proof. Using the Bellman equation we can write:

$$\nabla_a Q^{\pi^{\theta,\sigma}}(s,a) = \nabla_a r(s,a) + \int_{s'} \nabla_a P(s'|s,a) V^{\pi^{\theta,\sigma}}(s,a) \tag{154}$$

so

$$\begin{aligned}
\|\nabla_a Q^{\pi^{\theta_1,\sigma}}(s,a) - \nabla_a Q^{\pi^{\theta_2,\sigma}}(s,a)\| &\leq \int_{s'} \|\nabla_a P(s'|s,a)\| . |V^{\pi^{\theta_1,\sigma}}(s') - V^{\pi^{\theta_2,\sigma}}(s')| && (155)\\
&\leq L_P L_V && (156)\\
&= L_Q L_\mu L_P && (157)
\end{aligned}$$

where in the last inequality we used lemma 16 and 13. □

F USEFUL LEMMAS

Lemma 16. *Under assumption 4, it holds:*

$$\int_{s'} |P(s'|s,a_1) - P(s'|s,a_2)| ds' \leq L_P \|a_1 - a_2\| \tag{158}$$

for all $s \in \mathcal{S}$ and $a_1, a_2 \in \mathcal{A}$.

Proof. First, we define the two sets

$$A_1 = \{s' \in \mathcal{S} \quad | \quad P(s'|s,a_1) \leq P(s'|s,a_2)\}$$
$$A_2 = \{s' \in \mathcal{S} \quad | \quad P(s'|s,a_2) < P(s'|s,a_1)\}$$

Therefor, for all $s' \in A_1$ we have $0 \leq \log P(s'|s,a_2) - \log P(s'|s,a_1) \leq L_P\|a_1 - a_2\|$, and for all $s' \in A_2$ we have $0 \leq \log P(s'|s,a_1) - \log P(s'|s,a_2) \leq L_P\|a_1 - a_2\|$.
We can write:

$$\int_{s'} |P(s'|s,a_1) - P(s'|s,a_2)|ds' = \int_{s'\in A_1} |P(s'|s,a_1) - P(s'|s,a_2)|ds' + \int_{s'\in A_2} |P(s'|s,a_1) - P(s'|s,a_2)|ds' \tag{159}$$

$$= \int_{s'\in A_1} P(s'|s,a_1)\left[\exp\left(\log P(s'|s,a_2) - \log P(s'|s,a_1)\right) - 1\right]ds' \tag{160}$$

$$+ \int_{s'\in A_2} P(s'|s,a_2)\left[\exp\left(\log P(s'|s,a_1) - \log P(s'|s,a_2)\right) - 1\right]ds' \tag{161}$$

$$\leq \left(\exp(L_P\|a_1 - a_2\|) - 1\right)\left(\int_{s'\in A_1} P(s'|s,a_1)ds' + \int_{s'\in A_2} P(s'|s,a_2)ds'\right) \tag{162}$$

$$= \left(\exp(L_P\|a_1 - a_2\|) - 1\right)\left(1 - \frac{1}{2}\int_{s'} |P(s'|s,a_1) - P(s'|s,a_2)|ds'\right) \tag{163}$$

The last equality holds because:

$$\int_{s'\in A_1} P(s'|s,a_1)ds' + \int_{s'\in A_2} P(s'|s,a_2)ds' = 1 - \int_{s'\in A_2}\left(P(s'|s,a_1) - P(s'|s,a_2)\right)ds' \tag{164}$$

$$= 1 - \int_{s'\in A_1}\left(P(s'|s,a_2) - P(s'|s,a_1)\right)ds' \tag{165}$$

we sum and divide by two the two equalities.
Now we get:

$$\int_{s'} |P(s'|s,a_1) - P(s'|s,a_2)|ds' \leq 2\frac{\exp(L_P\|a_1 - a_2\|) - 1}{\exp(L_P\|a_1 - a_2\|) + 1} \tag{166}$$

$$= 2\tanh\frac{L_P\|a_1 - a_2\|}{2} \tag{167}$$

$$\leq L_P\|a_1 - a_2\| \tag{168}$$

where we used the fact that $\tanh x \leq x$ for all $x \geq 0$. □

Lemma 17. *Under 4, 5 , it holds that:*

$$\int_{s'} f(s')\|\nabla_x P(s'|s,\mu_x(s))\|ds' \leq L_P L_\mu \int_{s'} f(s')P(s'|s,\mu_x(s)) \tag{169}$$

for all positive function f, $s \in \mathcal{S}$ and $x \in \mathbb{R}^d$

Proof. By applying the chain rule, the gradient of the transition kernel with respect to the parameter x is given by:

$$\nabla_x P(s'|s,\mu_x(s)) = \nabla_x \mu_x(s)^\top \nabla_a P(s'|s,a)\big|_{a=\mu_x(s)} \tag{170}$$

Taking the norm of both sides, we get:

$$\|\nabla_x P(s'|s,\mu_x(s))\| \leq \|\nabla_x \mu_x(s)\| \|\nabla_a P(s'|s,a)\|\big|_{a=\mu_x(s)} \tag{171}$$

From assumption 5, the policy is L_μ-Lipschitz, which implies $\|\nabla_x \mu_x(s)\| \leq L_\mu$.

For the transition kernel, we use the standard log-derivative trick:

$$\nabla_a P(s'|s,a) = P(s'|s,a)\nabla_a \log P(s'|s,a) \tag{172}$$

As established in the proof of Lemma 16, assumption 4 implies that the function $a \mapsto \log P(s'|s,a)$ is L_P-Lipschitz. Therefore, its gradient is bounded by L_P:

$$\|\nabla_a \log P(s'|s,a)\| \leq L_P \tag{173}$$

Substituting this back, we bound the norm of the action gradient:

$$\|\nabla_a P(s'|s,a)\| \leq L_P P(s'|s,a) \tag{174}$$

Combining these inequalities gives a bound on the parameter gradient:

$$\|\nabla_x P(s'|s,\mu_x(s))\| \leq L_P L_\mu P(s'|s,\mu_x(s)) \tag{175}$$

Since $f(s')$ is a positive function, we can multiply both sides by $f(s')$ and integrate over s' without changing the direction of the inequality:

$$\begin{aligned}\int_{s'} f(s')\|\nabla_x P(s'|s,\mu_x(s))\| ds' &\leq \int_{s'} f(s') \left(L_P L_\mu P(s'|s,\mu_x(s))\right) ds' \\ &= L_P L_\mu \int_{s'} f(s') P(s'|s,\mu_x(s)) ds'\end{aligned} \tag{176}$$

□

Now using lemma 16 we can bound the total variation between $d^{\pi^{\theta_1,\sigma}}$ and $d^{\pi^{\theta_2,\sigma}}$:

Lemma 18. *Under assumptions 4, it holds:*

$$\int_s |d^{\pi^{\theta_1,\sigma}}(s) - d^{\pi^{\theta_2,\sigma}}(s)| ds \leq \frac{\gamma L_P L_\mu}{1-\gamma}\|\theta_1 - \theta_2\| \tag{177}$$

for all $s \in \mathcal{S}$, and $\theta_1, \theta_2 \in \mathbb{R}^d$.

Proof. We recall that the state visistation distribution $d^{\pi^{\theta,\sigma}}(s)$ is given by:

$$d^{\pi^{\theta,\sigma}}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s|\pi^{\theta,\sigma}) \tag{178}$$

where

$$\mathbb{P}(s_t = s|\pi^{\theta,\sigma}) = \int_{s_0,u_0,\ldots,s_{t-1},u_{t-1}} P_0(s_0)\Pi_{i=0}^{t-1} P(s_{i+1}|s_i,\mu_{\theta+\sigma u_i}(s_i))p(u_i) ds_0 du_0 \ldots ds_{t-1} du_{t-1} \tag{179}$$

which can be written as:

$$\mathbb{P}(s_t = s|\pi^{\theta,\sigma}) = \int_{s',u'} \mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) P(s|s',\mu_{\theta+\sigma u'}(s'))p(u') ds' du' \tag{180}$$

giving:

$$\begin{aligned}\mathbb{P}(s_t = s|\pi^{\theta_1,\sigma}) - \mathbb{P}(s_t = s|\pi^{\theta_2,\sigma}) &= \int_{s',u'} \mathbb{P}(s_{t-1} = s'|\pi^{\theta_1,\sigma}) \Big(P(s|s',\mu_{\theta_1+\sigma u'}(s')) - P(s|s',\mu_{\theta_2+\sigma u'}(s')) \Big) p(u') ds' \\ &+ \int_{s',u'} \Big(\mathbb{P}(s_{t-1} = s'|\pi^{\theta_1,\sigma}) - \mathbb{P}(s_{t-1} = s'|\pi^{\theta_2,\sigma}) \Big) P(s|s',\mu_{\theta_2+\sigma u'}(s'))p(u') ds'\end{aligned} \tag{181}$$

Integrating both sides over all states s, using lemma 16 and assumption 5, we get:

$$\int_s |\mathbb{P}(s_t = s|\pi^{\theta_1,\sigma}) - \mathbb{P}(s_t = s|\pi^{\theta_2,\sigma})|ds \leq L_P L_\mu \|\theta_1 - \theta_2\| + \int_s |\mathbb{P}(s_{t-1} = s|\pi^{\theta_1,\sigma}) - \mathbb{P}(s_{t-1} = s|\pi^{\theta_2,\sigma})|ds \tag{182}$$

By induction on t, we get:

$$\int_s |\mathbb{P}(s_t = s|\pi^{\theta_1,\sigma}) - \mathbb{P}(s_t = s|\pi^{\theta_2,\sigma})|ds \leq L_P L_\mu \|\theta_1 - \theta_2\| t \tag{183}$$

Using the last inequality, we conclude that:

$$\int_s |d^{\pi^{\theta_1,\sigma}}(s) - d^{\pi^{\theta_2,\sigma}}(s)|ds \leq (1-\gamma) L_P L_\mu \|\theta_1 - \theta_2\| \sum_{t=0}^{\infty} \gamma^t t \tag{184}$$

$$= \frac{\gamma L_P L_\mu}{1-\gamma} \|\theta_1 - \theta_2\| \tag{185}$$

□

Our aim is to upper bound the total variation between $d^{\pi^{\theta,\sigma}}$ and d^{μ_θ}. First, we prove the following lemma that is an implication of the smoothness of the log-transition probability.

Lemma 19. *Under assumptions 4 and 6, it holds that:*

$$\int_s |P(s'|s,a_1) - P(s'|s,a_2) - \langle \nabla_a P(s'|s,a)|_{a=a_2}, a_1 - a_2 \rangle| ds' \leq \frac{L_{2,P} + L_P^2}{2} \|a_1 - a_2\|^2 \tag{186}$$

Proof. Let $P_1 = P(s'|s,a_1)$ and $P_2 = P(s'|s,a_2)$. It suffices to prove that

$$\int_{s'} \|\nabla P_1 - \nabla P_2\| ds' \leq (L_{2,P} + L_P^2) \|a_1 - a_2\| \tag{187}$$

We use the log-derivative trick:

$$\nabla P_1 - \nabla P_2 = P_1 \nabla \log P_1 - P_2 \nabla \log P_2 \tag{188}$$

$$= P_1(\nabla \log P_1 - \nabla \log P_2) + (P_1 - P_2) \nabla \log P_2 \tag{189}$$

$$\int_{s'} \|\nabla P_1 - \nabla P_2\| ds' \leq \int_{s'} |P_1|.\|\nabla \log P_1 - \nabla \log P_2\| + \int_{s'} |P_1 - P_2|.\|\nabla \log P_1\| \tag{190}$$

$$\overset{(i)}{\leq} L_{2,P} \|a_1 - a_2\| + L_P \int_{s'} |P_1 - P_2| \tag{191}$$

$$\overset{(ii)}{\leq} (L_{2,P} + L_P^2) \|a_1 - a_2\| \tag{192}$$

where in (i) we used assumption 6, and in (ii) we used lemma 18.
To get the desired inequality, we can write

$$P_1 - P_2 = \int_0^1 \langle \nabla_a P(s'|s, a_2 + t(a_1 - a_2)), a_1 - a_2 \rangle \, dt \tag{193}$$

and

$$\langle \nabla P_2, a_1 - a_2 \rangle = \int_0^1 \langle \nabla P_2, a_1 - a_2 \rangle \, dt \tag{194}$$

which gives

$$P_1 - P_2 - \langle \nabla P_2, a_1 - a_2 \rangle = \int_0^1 \langle \nabla_a P(s'|s, a_2 + t(a_1 - a_2)) - \nabla P_2, a_1 - a_2 \rangle \, dt \tag{195}$$

$$\int_{s'} |P_1 - P_2 - \langle \nabla P_2, a_1 - a_2 \rangle| ds' = \int_{s'} \left| \int_0^1 \langle \nabla_a P(s'|s, a_2 + t(a_1 - a_2)) - \nabla P_2, a_1 - a_2 \rangle \, d \right| ds' \tag{196}$$

$$\leq \int_0^1 \int_{s'} \|\nabla_a P(s'|s, a_2 + t(a_1 - a_2)) - \nabla P_2\| . \|a_1 - a_2\| ds' dt \tag{197}$$

$$\overset{(i)}{\leq} (L_{2,P} + L_P^2) \int_0^1 t \|a_1 - a_2\|^2 dt \tag{198}$$

$$= \frac{L_{2,P} + L_P^2}{2} \|a_1 - a_2\|^2 \tag{199}$$

where in (i) we used the previous inequality we proved 187. □

Lemma 20. *Under assumptions 4, 6 and 7, it holds that*

$$\int_s |d^{\pi^{\theta,\sigma}}(s) - d^{\mu_\theta}(s)| ds \leq \gamma \frac{L_{2,P} L_\mu + L_P L_{2,\mu} + L_P^2 L_\mu}{2(1-\gamma)} d\sigma^2 \tag{200}$$

Proof. We recall that for the deterministic policy μ_θ we have

$$\mathbb{P}(s_t = s|\mu_\theta) = \int_{s'} \mathbb{P}(s_{t-1} = s') P(s|s', \mu_\theta(s')) ds' \tag{201}$$

and for the stochastic policy $\pi^{\theta,\sigma}$, $\mathbb{P}(s_t = s|\pi^{\theta,\sigma})$ is given in equation 180. Now we have:

$$\mathbb{P}(s_t = s|\pi^{\theta,\sigma}) - \mathbb{P}(s_t = s|\mu_\theta) = \int_{s'} \Bigg[\mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) \int_{u'} P(s|s', \mu_{\theta+\sigma u'}(s')) p(u') du' - \mathbb{P}(s_{t-1} = s'|\mu_\theta) P(s|s', \mu_\theta(s')) \Bigg] ds' \tag{202}$$

$$= \int_{s'} \Bigg[P(s|s', \mu_\theta(s')) \Big(\mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) - \mathbb{P}(s_{t-1} = s'|\mu_\theta) \Big) + \mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) \Big(\int_{u'} P(s|s', \mu_{\theta+\sigma u'}(s')) p(u') du' - P(s|s', \mu_\theta(s')) \Big) \Bigg] ds' \tag{203}$$

$$= \int_{s'} \Bigg[P(s|s', \mu_\theta(s')) \Big(\mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) - \mathbb{P}(s_{t-1} = s'|\mu_\theta) \Big) \Bigg] ds' + \int_{s'} \Bigg[\mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) \Big(\int_{u'} \Big(P(s|s', \mu_{\theta+\sigma u'}(s')) - P(s|s', \mu_\theta(s')) - \langle \nabla P(s|s', \mu_\theta(s')), \mu_{\theta+\sigma u'}(s) - \mu_\theta(s) \rangle \Big) p(u') du' \Bigg] ds' + \int_{s', u'} \mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) \langle \nabla P(s|s', \mu_\theta(s')), \mu_{\theta+\sigma u'}(s) - \mu_\theta(s) - \sigma \nabla_\theta \mu_\theta(s) u' \rangle p(u') du' ds \tag{204}$$

Where in the last term we used the fact that $\mathbb{E}[u] = 0$. Now we use lemma 19 and the smoothness of the smoothness of μ_θ (assumption 7) we get

$$\int_s \Big|\mathbb{P}(s_t = s|\pi^{\theta,\sigma}) - \mathbb{P}(s_t = s|\mu_\theta)\Big| ds \leq \int_{s'} \Big|\mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) - \mathbb{P}(s_{t-1} = s'|\mu_\theta)\Big| ds' + \frac{L_{2,P} + L_P^2}{2} \int_{u'} \|\mu_{\theta+\sigma u'}(s) - \mu_\theta(s)\|^2 p(u') du' + \frac{L_P L_{2,\mu}}{2} \sigma^2 d \tag{205}$$

$$\leq \int_{s'} \Big|\mathbb{P}(s_{t-1} = s'|\pi^{\theta,\sigma}) - \mathbb{P}(s_{t-1} = s'|\mu_\theta)\Big| ds' + \frac{L_{2,P} L_\mu^2 + L_P L_{2,\mu} + L_P^2 L_\mu^2}{2} d\sigma^2 \tag{206}$$

By induction on t, we get

$$\int_s \Big|\mathbb{P}(s_t = s|\pi^{\theta,\sigma}) - \mathbb{P}(s_t = s|\mu_\theta)\Big| ds \leq \frac{L_{2,P} L_\mu^2 + L_P L_{2,\mu} + L_P^2 L_\mu^2}{2} d\sigma^2 t \tag{207}$$

which gives:

$$\int_s |d^{\pi^{\theta,\sigma}}(s) - d^{\mu_\theta}(s)| ds \leq \gamma \frac{L_{2,P} L_\mu^2 + L_P L_{2,\mu} + L_P^2 L_\mu^2}{2(1-\gamma)} d\sigma^2 \tag{208}$$

□

Lemma 21. *The variance $\mathbb{E}[\|h(\omega_t) - \bar{h}(\omega_t)\|^2 | \mathcal{F}_t]$ is bounded by a constante σ_M, with $\sigma_M = \mathcal{O}\Big(\frac{d}{\sigma^2 M}\Big)$.*

Proof. To bound the variance of the batch estimator $\sigma_M^2 = \mathbb{E}\big[\|h(\omega_t) - \bar{h}(\omega_t)\|^2 | \mathcal{F}_t\big]$, we must account for the Markovian correlation of the state samples $s_{t,0}, \ldots, s_{t,M-1}$. We recall that $h(\omega_t) = \frac{1}{M}\sum_{i=0}^{M-1} g_{t,i}$, where $g_{t,i} = \frac{\delta_{t,i}}{\sigma} u_i$, and $\bar{h}(\omega_t)$ is its conditional expectation given the initial state of the batch $s_{t,0}$ (i.e., the filtration $\mathcal{F}_t$).

We start by expanding the variance of the sample mean into diagonal and cross terms:

$$\sigma_M^2 = \frac{1}{M^2} \sum_{i=0}^{M-1} \mathbb{E}\big[\|g_{t,i} - \mathbb{E}[g_{t,i}|\mathcal{F}_t]\|^2 \big| \mathcal{F}_t\big] + \frac{1}{M^2} \sum_{i \neq j} \mathbb{E}\big[\langle g_{t,i} - \mathbb{E}[g_{t,i}|\mathcal{F}_t], g_{t,j} - \mathbb{E}[g_{t,j}|\mathcal{F}_t]\rangle \big| \mathcal{F}_t\big]. \tag{209}$$

Because we condition on the initial state $s_{t,0}$, the chain is in a transient regime, meaning the samples are not drawn from the stationary distribution d^π. Therefore, we cannot directly apply the uniform ergodicity assumption 22 (which assume the variables are centered around the stationary mean). Instead, we must bound the covariance directly.

Let $\mu_i = \mathbb{E}[g_{t,i}|\mathcal{F}_t]$ be the conditional expectation of the i-th sample, and let $Z_i = g_{t,i} - \mu_i$ be the centered random variable. Assume without loss of generality that $i < j$. By the tower property of expectations, conditioning on the intermediate state $s_{t,i}$, we have:

$$\mathbb{E}\big[\langle Z_i, Z_j\rangle \big| \mathcal{F}_t\big] = \mathbb{E}_{s_{t,i}}\Big[\big\langle g_{t,i} - \mu_i,\ \mathbb{E}[g_{t,j} - \mu_j | s_{t,i}]\big\rangle \Big| \mathcal{F}_t\Big] \leq \mathbb{E}_{s_{t,i}}\Big[\|g_{t,i} - \mu_i\| \cdot \|\mathbb{E}[g_{t,j}|s_{t,i}] - \mu_j\| \Big| \mathcal{F}_t\Big]. \tag{210}$$

Under the uniform ergodicity assumption 1, there exist constants $C > 0$ and $\rho \in (0,1)$ such that:

- Starting from $s_{t,i}$, the j-th sample is $(j-i)$ steps away: $\|\mathbb{E}[g_{t,j}|s_{t,i}] - \mu_{\text{stat}}\| \leq C\rho^{j-i}$.
- Starting from $s_{t,0}$, the j-th sample is j steps away: $\|\mu_j - \mu_{\text{stat}}\| \leq C\rho^j$.

We can now bound the right side of the inner product in Equation 210 by adding and subtracting μ_{stat} and applying the triangle inequality:

$$\begin{aligned}\|\mathbb{E}[g_{t,j}|s_{t,i}] - \mu_j\| &= \big\|\big(\mathbb{E}[g_{t,j}|s_{t,i}] - \mu_{\text{stat}}\big) - \big(\mu_j - \mu_{\text{stat}}\big)\big\| \\ &\le \|\mathbb{E}[g_{t,j}|s_{t,i}] - \mu_{\text{stat}}\| + \|\mu_j - \mu_{\text{stat}}\| \\ &\le C\rho^{j-i} + C\rho^j \\ &\le 2C\rho^{j-i}, \end{aligned} \tag{211}$$

where the last inequality holds because $\rho^j < \rho^{j-i}$ for $i > 0$.

Substituting this back into Equation 210, and using the fact that $\mathbb{E}[\|Z_i\|] \le \mathbb{E}[\|Z_i\|^2] \le \max_k \mathbb{E}[\|Z_k\|^2|\mathcal{F}_t]$, we obtain :

$$\mathbb{E}\big[\langle Z_i, Z_j\rangle\big|\mathcal{F}_t\big] \le 2C\rho^{j-i} \max_k \mathbb{E}\big[\|g_{t,k} - \mu_k\|^2\big|\mathcal{F}_t\big]. \tag{212}$$

Now we return to the full variance expansion in Equation 209. Using the bound for the geometric series $\sum_{i\neq j} \rho^{|i-j|} \le M\frac{2\rho}{1-\rho}$ and defining $C'_\rho = 1 + 2C\frac{2\rho}{1-\rho}$, we get:

$$\begin{aligned}\sigma_M^2 &\le \frac{1}{M^2}\left(M + M\cdot 2C\frac{2\rho}{1-\rho}\right)\max_k \mathbb{E}\big[\|g_{t,k} - \mu_k\|^2\big|\mathcal{F}_t\big] \\ &\le \frac{\tau_{\text{mix}}}{M}\max_k \mathbb{E}\big[\|g_{t,k}\|^2\big|\mathcal{F}_t\big]. \end{aligned} \tag{213}$$

Finally, using $\mathbb{E}[\|g_{t,k}\|^2|\mathcal{F}_t] \le \frac{\delta_{\max}^2 d}{\sigma^2}$, we arrive at the explicit variance bound:

$$\sigma_M^2 \le \frac{C'_\rho \delta_{\max}^2 d}{\sigma^2 M} = \mathcal{O}\left(\frac{d}{\sigma^2 M}\right). \tag{214}$$

□

Lemma 22. *Following the same notations as in Appendix D, assume that there exists $K > 0$ such that:*

$$\sup_{s\in\mathcal{S}} \|\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})|s_{t,i}]\|^2 \le K$$

Then, under assumption 1 it holds:

$$\|\frac{1}{M}\sum_{i=0}^{M-1} \mathbb{E}[g(s_{t,i+1}, u_i, s_{t,i})|\mathcal{F}_t] - \mathbb{E}_{s',u,s}[g(s', u, s)]\|^2 \le \frac{C_\rho K}{M} \tag{215}$$

where $s \sim d^{\pi^{\theta,\sigma}}$, $u \sim N(0, I)$, $s' \sim P(.|s, \mu_{\theta+\sigma u}(s))$, $C_\rho = 4 + \frac{2}{(1-\rho)^2}$

Proof. We have

$$\begin{aligned}\|\bar{h}(\omega_t) - H(\omega_t)\|^2 \le \frac{1}{M^2}\sum_{i=0}^{M-1} &\|\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})] - H(\omega_t)\|^2 \\ &+ \frac{1}{M^2}\sum_{i\neq j} \|\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})] - H(\omega_t)\|.\|\mathbb{E}[g(s_{t,j}, u_j, s_{t,j+1})] - H(\omega_t)\| \end{aligned} \tag{216}$$

For the first sum, we write $\|\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})] - H(\omega_t)\|^2 \le 2\|\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})|s_{t,i}]\|^2 + 2\|H(\omega_t)\|^2 \le 2K + 2\|H(\omega_t)\|^2$, and $\|H(\omega_t)\|^2 \le \mathbb{E}_s[\|\mathbb{E}_{u,s'}[g(s, u, s')]\|^2] \le K$, therefore $\|\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})] - H(\omega_t)\|^2 \le 4K$. For the second term, we use assumption 1 to get:

$$\begin{aligned}\|\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})] - H(\omega_t)\| &= \|\mathbb{E}_{s_{t,i}\sim P^i(.|s_{t,0})}[\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})|s_{t,i}]] \\ &\quad - \mathbb{E}_{s_{t,i}\sim d^{\pi^{\theta,\sigma}}}[\mathbb{E}[g(s_{t,i}, u_i, s_{t,i+1})|s_{t,i}]]\| \qquad (217) \\ &\le \sqrt{K}\int_{s_{t,i}} |P^i(s_{t,i}|s_{t,0}) - d^{\pi^{\theta,\sigma}}(s_{t,i}))|ds_{t,i} \qquad (218) \\ &\le \sqrt{K}\kappa\rho^i \qquad (219) \end{aligned}$$

Therfore

$$\|\frac{1}{M}\sum_{i=0}^{M-1}\mathbb{E}[g(s_{t,i+1},u_i,s_{t,i})|\mathcal{F}_t]-\mathbb{E}_{s',u,s}[g(s',u,s)]\|^2 \leq \frac{4K}{M}+\frac{2K\kappa}{M}\sum_{i>j}\rho^{i+j} \tag{220}$$

$$\leq C_\rho\frac{K}{M} \tag{221}$$

□

Lemma 23. *let A_1, A_2, ω_1 and ω_2 be the expected feature transition matrices and the TD fixed points corresponding to the policies $\pi^{\theta_1,\sigma}$ and $\pi^{\theta_2,\sigma}$. Then*

$$\|\omega_1^*-\omega_2^*\| \leq 4R_{\max}\Phi_{\max}^3\frac{\gamma L_P L_\mu}{1-\gamma}\|\theta_1-\theta_2\| \tag{222}$$

Proof. Let $\hat{\omega}$ be the unique solution of the linear system:

$$A_1\hat{\omega}=b_2 \tag{223}$$

With the triangle inequality, we have

$$\|\omega_1^*-\omega_2^*\| \leq \|\omega_1^*-\hat{\omega}\|+\|\omega_2^*-\hat{\omega}\| \tag{224}$$

$$\leq \|A_1(b_1-b_2)\|+\|(A_1^{-1}-A_2^{-1})b_2\| \tag{225}$$

$$\leq 2\Phi_{\max}^2\|b_1-b_2\|+R_{\max}\Phi_{\max}\|A_1^{-1}-A_2^{-1}\| \tag{226}$$

Now, using lemma 18, we can upper bound each term:

$$\|b_1-b_2\| \leq R_{\max}\Phi_{\max}\int_s |d^{\pi^{\theta_1,\sigma}}(s)-d^{\pi^{\theta_1,\sigma}}(s)|ds \tag{227}$$

$$\leq R_{\max}\Phi_{\max}\frac{\gamma L_P L_\mu}{1-\gamma}\|\theta_1-\theta_2\| \tag{228}$$

and similarly

$$\|A_1^{-1}-A_2^{-1}\| \leq 2\Phi_{\max}^2\frac{\gamma L_P L_\mu}{1-\gamma}\|\theta_1-\theta_2\| \tag{229}$$

□

Lemma 24. *Let $\alpha_1 \in]0,1[$, $\alpha_2>0$, $\alpha_3>0$, $c>0$, and $(r_t),(s_t)$ be two positive sequences satisfying:*

$$r_{t+1} \leq (1-\alpha_1)r_t+\alpha_2\sqrt{r_t}\sqrt{s_t}+\alpha_3 s_t+c \tag{230}$$

We define $R_T=\frac{1}{T}\sum_{t=0}^{T-1}r_t$ and $S_T=\frac{1}{T}\sum_{t=0}^{T-1}s_t$. Then:

$$R_T \leq \frac{2(r_0-r_T)}{\alpha_1 T}+\beta S_T+\frac{2c}{\alpha_1} \tag{231}$$

where $\beta=\frac{\alpha_2^2}{\alpha_1^2}+\frac{2\alpha_3}{\alpha_1}$.

Proof. We begin by rearranging the recurrence relation to isolate the difference $r_{t+1}-r_t$:

$$r_{t+1}-r_t \leq -\alpha_1 r_t+\alpha_2\sqrt{r_t}\sqrt{s_t}+\alpha_3 s_t+c \tag{232}$$

To bound the cross term $\alpha_2\sqrt{r_t}\sqrt{s_t}$, we apply Young's inequality, $ab\leq\frac{a^2}{2\eta}+\frac{\eta b^2}{2}$ for any $\eta>0$. By choosing $\eta=\frac{\alpha_1}{\alpha_2}$ and setting $a=\sqrt{r_t}$ and $b=\alpha_2\sqrt{s_t}$, we obtain:

$$\alpha_2\sqrt{r_t}\sqrt{s_t} \leq \frac{\alpha_1}{2}r_t+\frac{\alpha_2^2}{2\alpha_1}s_t \tag{233}$$

Substituting this bound back into the original inequality yields:

$$r_{t+1} - r_t \leq -\alpha_1 r_t + \frac{\alpha_1}{2} r_t + \frac{\alpha_2^2}{2\alpha_1} s_t + \alpha_3 s_t + c \tag{234}$$

$$r_{t+1} - r_t \leq -\frac{\alpha_1}{2} r_t + \left(\frac{\alpha_2^2}{2\alpha_1} + \alpha_3 \right) s_t + c \tag{235}$$

Next, we sum this inequality from $t = 0$ to $t = T - 1$. The left-hand side forms a telescoping sum:

$$r_T - r_0 \leq -\frac{\alpha_1}{2} \sum_{t=0}^{T-1} r_t + \left(\frac{\alpha_2^2}{2\alpha_1} + \alpha_3 \right) \sum_{t=0}^{T-1} s_t + cT \tag{236}$$

Dividing the entire inequality by T and substituting the definitions of the time-averaged sequences R_T and S_T:

$$\frac{r_T - r_0}{T} \leq -\frac{\alpha_1}{2} R_T + \left(\frac{\alpha_2^2}{2\alpha_1} + \alpha_3 \right) S_T + c \tag{237}$$

Finally, we rearrange the terms to isolate R_T:

$$\frac{\alpha_1}{2} R_T \leq \frac{r_0 - r_T}{T} + \left(\frac{\alpha_2^2}{2\alpha_1} + \alpha_3 \right) S_T + c \tag{238}$$

$$R_T \leq \frac{2(r_0 - r_T)}{\alpha_1 T} + \left(\frac{\alpha_2^2}{\alpha_1^2} + \frac{2\alpha_3}{\alpha_1} \right) S_T + \frac{2c}{\alpha_1} \tag{239}$$

This matches the desired result with $\beta = \frac{\alpha_2^2}{\alpha_1^2} + \frac{2\alpha_3}{\alpha_1}$. □

G Deep Parameter Step-Based Exploration

While actor-critic architectures have lower variance compared with Monte Carlo full-trajectory returns, as it is the case in the PGPE algorithm (Montenegro et al., 2024; 2025), they inherently rely on biased gradient estimators. In Algorithm 1, this bias stems from three primary sources: the restricted representational capacity of the linear state-value function, the bootstrapping mechanism of TD(0), and the continuous Markovian sampling.

To mitigate the representation bias, we propose the Deep Parameter Step-Based Exploration (Deep-PSBE) in Algorithm 2. This extension replaces the linear critic with a highly expressive Multi-Layer Perceptron (MLP) to approximate the state-value function. However, replacing a linear function with a non-linear neural network is instable when bootstrapping. To stabilize the learning process, we adopt the target network mechanism introduced in DDPG (Lillicrap et al., 2016). More precisely, we decouple the TD target computation by evaluating the next state's value $V_{\omega^-}(s')$ using a separate, slow-moving target network parameterized by ω^-. This target network is updated via Polyak averaging ($\omega^- \leftarrow \tau\omega + (1-\tau)\omega^-$) with a smoothing coefficient $\tau \ll 1$. This constrains the target values to change gradually, preventing the destructive feedback loops that typically cause divergence in deep TD learning. Furthermore, to explicitly control the bootstrapping bias of TD(0), we incorporate Generalized Advantage Estimation (GAE) (Schulman et al., 2015), utilizing the decay parameter $\lambda \in [0, 1]$ to smoothly interpolate between the low-variance one-step TD formulation and the unbiased full-trajectory Monte Carlo returns.

Furthermore, Deep-PSBE completely bypasses the need for a target actor. Standard methods like DDPG or TD3 heavily rely on a target actor to keep the policy from shifting wildly and destabilizing the critic's targets. We handle this inherently through our two-time-scale architecture. By setting the actor's step size significantly smaller than the critic's ($\xi^\theta \ll \xi^\omega$), we force the policy to evolve slowly and smoothly. From the critic's perspective, this slow-moving policy creates a highly stable environment. This allows the faster-updating critic to easily track and converge to the true value function. Therefore and because the actor naturally stabilizes itself through this timescale separation, adding the complexity of a target actor is simply unnecessary.

Algorithm 2 Deep Parameter Step-Based Exploration (Deep-PSBE)

Initialize:
Actor parameter θ_0, Critic MLP parameter ω_0
Initialize Target Network:
$\omega_0^- \leftarrow \omega_0$
Hyperparameters:
Batch size M, GAE parameter $\lambda \in [0, 1]$, target smoothing coefficient $\tau \ll 1$
Draw $s_{0,0}$ from the initial distribution P_0
for $t = 0, \dots, T-1$ **do**
  1. Sample M Markovian transitions:
  $s_{t,0} = s_{t-1,M}$
  for $i = 0, \dots, M-1$ **do**
    $u_i \sim \mathcal{N}(0, I_d), a_{t,i} = \mu_{\theta_t + \sigma u_i}(s_{t,i}), s_{t,i+1} \sim \tilde{P}(\cdot|s_{t,i}, a_{t,i})$
  end for
  2. Generalized Advantage Estimation (GAE):
  Compute TD errors using the target network for stability:
  $\forall i : \delta_{t,i} = r_{t,i} + \gamma V_{\omega^-}(s_{t,i+1}) - V_\omega(s_{t,i})$
  Compute GAE backward through the mini-batch: $\hat{A}_{t,M-1} = \delta_{t,M-1}$
  for $i = M-2$ **down to** 0 **do**
    $\hat{A}_{t,i} = \delta_{t,i} + \gamma\lambda\hat{A}_{t,i+1}$
  end for
  3. Update the Actor:
  $\theta_{t+1} = \theta_t + \xi^\theta \frac{1}{M} \sum_{i=0}^{M-1} \frac{\hat{A}_{t,i}}{\sigma} u_i$
  4. Update the Critic:
  Compute value targets: $\hat{V}_{t,i} = \hat{A}_{t,i} + V_{\omega^-}(s_{t,i})$
  Minimize the MSE loss by semi-gradient descent:
  $\omega_{t+1} = \omega_t + \xi^\omega \frac{1}{M} \sum_{i=0}^{M-1} (\hat{V}_{t,i} - V_\omega(s_{t,i})) \nabla_\omega V_\omega(s_{t,i})$
  5. Update Target Network:
  $\omega_{t+1}^- \leftarrow \tau\omega_{t+1} + (1-\tau)\omega_t^-$
end for

H Experimental details

In this section, we present the full experimental setup, the hyperparameter grids searched, and the per-σ learning curves summarised in Figure 1.

H.1 Experimental Setup

All runs use a discount factor of $\gamma = 0.99$ and a budget of 2×10^6 environment transitions, obtained as $T = 1000$ iterations of $M = 2000$ transitions for PSBE and Deep-PSBE. Performance is measured on evaluation episodes of horizon 200: returns are undiscounted and every reported quantity is a mean over 5 seeds with 95% confidence intervals.

H.2 Hyperparameter grids

Table 1 gives the grid searched for each algorithm, identical across the three environments. In PSBE and Deep-PSBE, the critic step size is coupled to the actor's, we take $\xi^\omega = c\,\xi^\theta$ with $c \in \{10, 100\}$.

H.3 Per-σ learning curves

Figure 3 provides the full learning curves underlying the summary in Figure 1, showing J_D and J_σ at each exploration level in the grid across all environments.

Algorithm	Hyperparameter	Values
PSBE	iterations T	1000
	transitions per iteration M	2000
	actor step size ξ^θ	$\{10^{-4},\ 10^{-3},\ 3\cdot10^{-3},\ 3\cdot10^{-2}\}$
	critic step size ξ^ω	$c\,\xi^\theta$, $c \in \{10, 100\}$
	exploration level σ	$\{0.01,\ 0.03,\ 0.1,\ 0.3,\ 1,\ 3\}$
	discount factor γ	0.99
	projection radius R_ω	$\{10,\ 100\}$
Deep-PSBE	iterations T	1000
	transitions per iteration M	2000
	actor step size ξ^θ	$\{3\cdot10^{-5},\ 10^{-4},\ 3\cdot10^{-3}\}$
	critic step size ξ^ω	$c\,\xi^\theta$, $c \in \{10, 100\}$
	exploration level σ	$\{0.03,\ 0.1,\ 0.3,\ 1\}$
	GAE λ	0.95
	target coefficient τ	0.005
	discount factor γ	0.99
	critic architecture	128×128
	activation function	ReLU
PGPE	iterations	1000
	trajectories per iteration	10
	horizon	200
	step size	$\{3\cdot10^{-3},\ 10^{-2},\ 3\cdot10^{-2},\ 10^{-1}\}$
	exploration level σ	$\{0.01,\ 0.03,\ 0.1,\ 0.3,\ 1\}$
	discount factor γ	0.99
DDPG, TD3	environment steps	$2\cdot10^6$
	learning rate	$\{10^{-4},\ 3\cdot10^{-4},\ 10^{-3}\}$
	action noise	$\{0.1,\ 0.2\}$
	discount factor γ	0.99
	batch size	256
	replay buffer capacity	10^6
	warm-up steps	1000

Table 1: Hyperparameter grids of the compared algorithms.

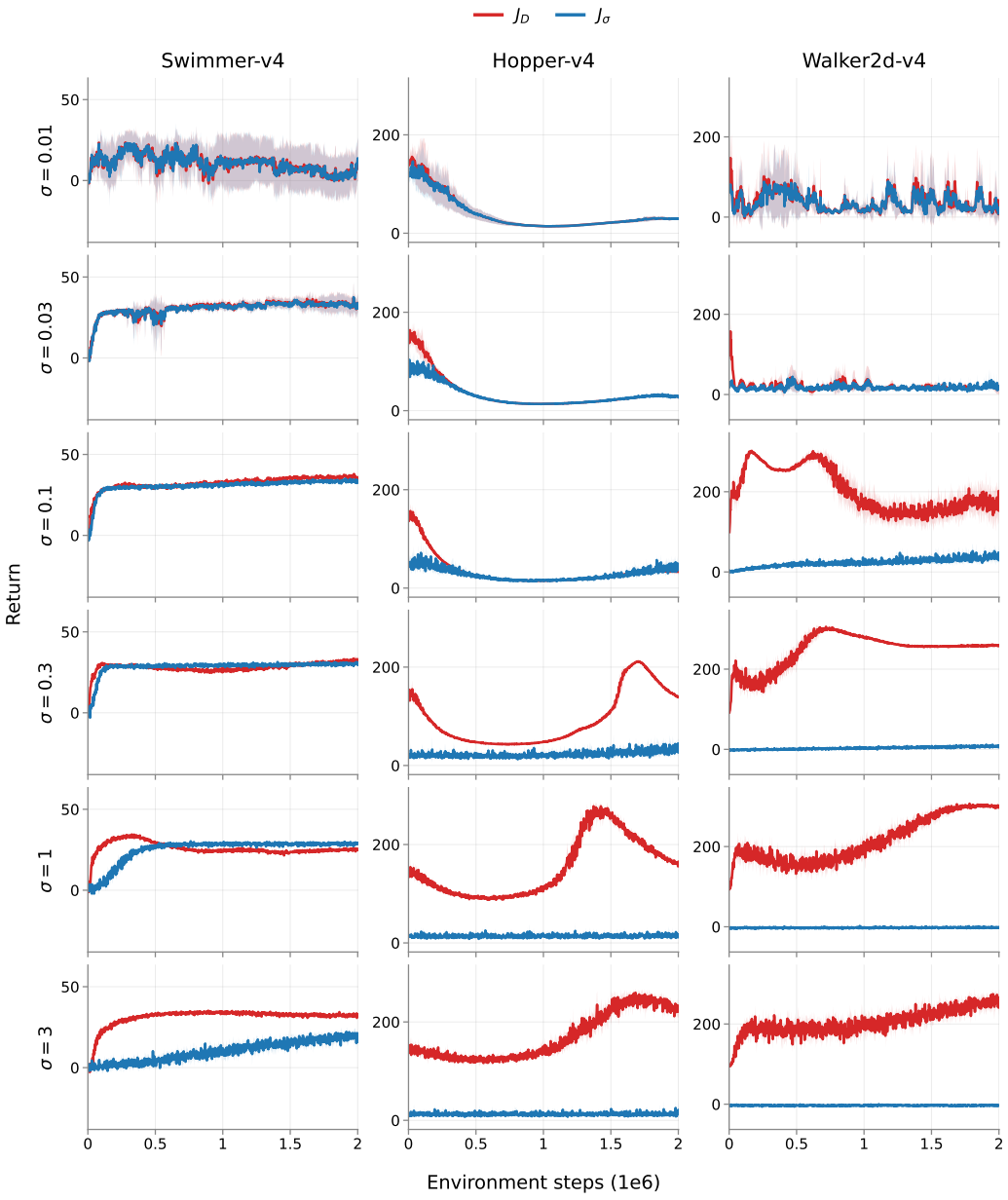

Figure 3: J_σ and J_D learning curves at every exploration level, showing the gap between the two objectives widen as σ grows (5 seeds, mean $\pm$ 95% C.I.).