Policy Gradient with Active Importance Sampling

Matteo Papini
matteo.papini@polimi.it
Politecnico di Milano

Giorgio Manganini
giorgio.manganini@gssi.it
Gran Sasso Science Institute

Alberto Maria Metelli
albertomaria.metelli@polimi.it
Politecnico di Milano

Marcello Restelli
marcello.restelli@polimi.it
Politecnico di Milano

Abstract

Importance sampling (IS) represents a fundamental technique for a large surge of off-policy reinforcement learning approaches. Policy gradient (PG) methods, in particular, significantly benefit from IS, enabling the effective reuse of previously collected samples, thus increasing sample efficiency. However, classically, IS is employed in RL as a passive tool for re-weighting historical samples. However, the statistical community employs IS as an active tool combined with the use of behavioral distributions that allow the reduction of the estimate variance even below the sample mean one. In this paper, we focus on this second setting by addressing the behavioral policy optimization (BPO) problem. We look for the best behavioral policy from which to collect samples to reduce the policy gradient variance as much as possible. We provide an iterative algorithm that alternates between the cross-entropy estimation of the minimum-variance behavioral policy and the actual policy optimization, leveraging on defensive IS. We theoretically analyze such an algorithm, showing that it enjoys a convergence rate of order $O(\epsilon^{-4})$ to a stationary point, but depending on a more convenient variance term w.r.t. standard PG methods. We then provide a practical version that is numerically validated, showing the advantages in the policy gradient estimation variance and on the learning speed.

1 Introduction

Policy gradient (PG, Peters & Schaal, 2006) algorithms represent a large class of *reinforcement learning* (RL, Sutton & Barto, 2018) approaches that are particularly suitable to address complex control problems thanks to their ability to deal with continuous state and action spaces natively. PG methods address the RL problem by considering a parametric control *policy* $\pi_{\boldsymbol{\theta}}$ and formulate the learning process as a particular stochastic optimization problem by updating the policy parameters $\boldsymbol{\theta}$ in the ascent direction of the policy gradient. Clearly, the policy gradient needs to be estimated from samples, making the accuracy of such an estimate crucial for the actual performance of the PG approaches (Zhao et al., 2011; Papini et al., 2022).

In this direction, a significant line of research is represented by the approach to sample reuse. Borrowing the techniques from the statistical simulation community, *importance sampling* (IS, Owen, 2013) has been imported to the PG methods. The majority of the approaches that apply IS to PG methods are based on the idea of reweighting the data collected in the past (i.e., with *behavioral policies*) proportionally to the probability of being generated by the current policy (i.e., *target policy*), whose gradient needs to be estimated (e.g., Thomas et al., 2015; Metelli et al., 2018). Theoretical results about the advantages in terms of variance reduction have been provided in Metelli et al. (2020). However, these approaches can be considered *passive* since the focus is on reusing in the most effective way the sample collected in the past without considering the possibility of *choosing* the behavioral policy to improve the estimation of the gradient of the current target policy.

Indeed, this is the main use of IS for in the Monte Carlo simulation community, where this technique takes an *active* role. Specifically, in these scenarios, the objective is to find the best behavioral policy from which to collect samples in order to reduce the estimate variance as much as possible. It can be proved that under specific assumptions on the random variable whose expectation is to be estimated, such off-policy variance can be reduced even below that of the standard sample mean estimate Owen (2013). Although this line represents an appealing direction within a class of approaches (like RL) that suffer from an inherent sample inefficiency, the community has not deeply studied this direction.

Original Contributions In this paper, we focus on the active role of IS in the PG family of RL algorithms. Specifically, we investigate if we can actively learn the behavioral policy from which to collect samples in order to control the variance of the PG estimator effectively. We call this problem *behavioral policy optimization* (BPO). The contributions of the paper can be stated as follows:

- We formulate the BPO problem as finding the behavioral policy that minimizes the variance of the off-policy gradient estimate of a given target policy. After showing that this optimization problem allows for a closed-form solution under restrictive conditions, we introduce an approach for estimating such a behavioral policy based on cross-entropy minimization (Section 3).
- We provide a theoretical analysis of a principled algorithm that alternates two phases: behavioral policy learning based on cross-entropy and actual performance optimization based on the off-policy gradient update. We show that a careful sample partition between the two phases allows for achieving convergence rates of order $O(\epsilon^{-4})$ but depending on a more convenient variance term compared to standard REINFORCE (Section 4).
- We provide a practical version of the analyzed algorithm that uses all the samples collected. Then, we empirically evaluate such an algorithm, showing a significant reduction in the variance of the gradient estimate that translates into a faster learning curve (Section 6).

The proofs of all the results reported in the main paper can be found in Appendix B.

2 Preliminaries

Notation Let $n \in \mathbb{N}$, we denote with $[n] := \{1, \dots, n\}$. For a measurable set $\mathcal{X}$, we denote with $\Delta^{\mathcal{X}}$ the set of probability measures over $\mathcal{X}$. Let $P, Q \in \Delta^{\mathcal{X}}$ be two probability measures such that $P \ll Q$, that is, P is absolutely continuous with respect to Q. When the reference measure λ is clear from the context (Lebesgue measure for continuous $\mathcal{X}$ and counting measure for discrete $\mathcal{X}$, respectively), we use p to denote the Radon-Nikodym derivative $\mathrm{d}P/\mathrm{d}\lambda$ (density and mass function, respectively) and $\int_{\mathcal{X}} \cdot \, \mathrm{d}x$ to denote integration with respect to λ (Lebesgue integral and summation, respectively). We define the KL-divergence D_{KL} and the chi-square divergence χ^2 as:

$$D_{\mathrm{KL}}(P\|Q) := \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x, \quad \chi^2(P\|Q) := \int_{\mathcal{X}} \frac{(p(x)-q(x))^2}{q(x)} \mathrm{d}x. \tag{1}$$

Markov Decision Processes A discounted Markov decision problem (MDP, Puterman, 2014) is defined as a 6-tuple $(\mathcal{S}, \mathcal{A}, P, R, \mu_0, \gamma)$, where $\mathcal{S}$ is the measurable state space, $\mathcal{A}$ is the measurable action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ is the transition model defining for every $(s, a) \in \mathcal{S} \times \mathcal{A}$ the probability distribution of the next state $s' \sim P(\cdot|s, a)$, $R : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$ is the reward function $R(s, a)$ when performing action a in state s, uniformly bounded by $R_{\max} < +\infty$ defining the reward $R(s, a)$ obtained when playing action a in state s, $\mu_0 \in \Delta^{\mathcal{S}}$ is the initial-state distribution prescribing the state at which interaction begins, $s_0 \sim \mu_0$, and $\gamma \in [0, 1]$ is the discount factor.

Actor-only Policy Gradient We consider an agent whose behavior is described by a parametric policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta^{\mathcal{A}}$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is the parameter belonging to the parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$, assumed to be convex. In this setting, the agent's goal consists of maximizing the expected return:

$$\boldsymbol{\theta}^* \in \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} J(\boldsymbol{\theta}) := \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}} \left[R(\boldsymbol{\tau})\right], \qquad \text{where} \qquad R(\boldsymbol{\tau}) := \sum_{t=0}^{T-1} \gamma^t R(s_t, a_t),$$

and $\boldsymbol{\tau} = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}) \in \boldsymbol{\mathcal{T}}$ is the trajectory whose probability density function is given by $p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) = \mu_0(s_0) \prod_{t=0}^{T-1} \pi_{\boldsymbol{\theta}}(a_t|s_t) P(s_{t+1}|s_t, a_t)$, T is the trajectory length, and $\boldsymbol{\mathcal{T}} = (\mathcal{S} \times \mathcal{A})^T$ is the trajectory set.[1] If $\pi_{\boldsymbol{\theta}}$ is differentiable in $\boldsymbol{\theta}$, we can express the *policy gradient* (Williams, 1992), that is the gradient of the expected return $J(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$\nabla J(\boldsymbol{\theta}) = \underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}}{\mathbb{E}} \left[\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) R(\boldsymbol{\tau}) \right].$$

Actor-only methods (Peters & Schaal, 2006) perform learning by updating the policy parameters in the direction of the ascending policy gradient $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$, where $\alpha > 0$ is the step size.

On-policy gradient estimators The policy gradient $\nabla J(\boldsymbol{\theta})$ needs to be estimated from a set of collected trajectories. If the trajectories $\mathcal{D}_{\text{on}} = \{\boldsymbol{\tau}_i\}_{i \in [n]}$ are collected with the same policy $\pi_{\boldsymbol{\theta}}$ of which we seek to estimate the policy gradient, we speak of *on-policy* gradient estimation:

$$\widehat{\nabla} J(\boldsymbol{\theta}; \mathcal{D}_{\text{on}}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_i), \qquad \boldsymbol{\tau}_i \sim p_{\boldsymbol{\theta}}, \quad \forall i \in [n], \tag{2}$$

where $\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})$ is a single-trajectory estimator of the policy gradient. Classical unbiased estimators include: REINFORCE (Williams, 1992) where $\mathbf{g}^{\text{R}}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) = (\sum_{t=0}^{T-1} \nabla \log \pi_{\boldsymbol{\theta}}(a_t|s_t)) R(\boldsymbol{\tau})$ and G(PO)MPD (Baxter & Bartlett, 2001) where $\mathbf{g}^{\text{G}}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) = \sum_{t=0}^{T-1} \gamma^t R(s_t, a_t) \sum_{l=0}^{t} \nabla \log \pi_{\boldsymbol{\theta}}(a_l|s_l)$.

Off-policy gradient estimators with Single behavioral policy When, instead, we seek to estimate the policy gradient $\nabla J(\boldsymbol{\theta})$ of a *target* policy $\pi_{\boldsymbol{\theta}}$ having collected n trajectories $\mathcal{D}_{\text{off}} = \{\boldsymbol{\tau}_i\}_{i \in [n]}$ with a different *behavioral* policy $\pi_{\boldsymbol{\theta}^b}$, under the assumption that $\pi_{\boldsymbol{\theta}}(\cdot|s) \ll \pi_{\boldsymbol{\theta}^b}(\cdot|s)$ for every $s \in \mathcal{S}$, we speak of *(single) off-policy* gradient estimation:[2]

$$\widehat{\nabla} J(\boldsymbol{\theta}; \mathcal{D}_{\text{off}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}_i)}{p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau}_i)} (\boldsymbol{\tau}_i) \mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_i), \qquad \boldsymbol{\tau}_i \sim p_{\boldsymbol{\theta}^b}, \quad \forall i \in [n], \tag{3}$$

where $\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau})}$ is the trajectory *(simple) importance weight* (Owen, 2013), defined as:

$$\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau})} = \prod_{t=0}^{T-1} \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}^b}(a_t|s_t)}. \tag{4}$$

Off-policy gradient estimators with Multiple behavioral policies It is possible to extend these estimators to the case in which trajectories are collected from multiple $m \in \mathbb{N}$ behavioral policies parameters $\{\boldsymbol{\theta}^b_j\}_{j \in [m]}$. In such a case, for every $j \in [m]$, we have collected n_j trajectories $\{\boldsymbol{\tau}_{ij}\}_{i \in [n_j]}$ from the behavioral policy $\pi_{\boldsymbol{\theta}^b_j}$ and such that $\beta_j(\cdot) \pi_{\boldsymbol{\theta}}(\cdot|s) \ll \pi_{\boldsymbol{\theta}^b_j}(\cdot|s)$ for every $s \in \mathcal{S}$, we speak of *multiple off-policy* gradient estimation:

$$\widehat{\nabla} J(\boldsymbol{\theta}; \mathcal{D}_{\text{off}}; \boldsymbol{\beta}) = \sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} \beta_j(\boldsymbol{\tau}_{ij}) \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{ij})}{p_{\boldsymbol{\theta}^b_j}(\boldsymbol{\tau}_{ij})} \mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{ij}), \quad \boldsymbol{\tau}_{ij} \sim p_{\boldsymbol{\theta}^b_j}, \quad \forall i \in [n_j], \ \forall j \in [m], \tag{5}$$

where $\mathcal{D}_{\text{off}} = \{\{\boldsymbol{\tau}_{ij}\}_{i \in [n_j]}\}_{j \in [m]}$ and $\beta_j(\boldsymbol{\tau}) \geqslant 0$ for every $j \in [m]$ and $\sum_{j=1}^{m} \beta_j(\boldsymbol{\tau}) = 1$ for every trajectory $\boldsymbol{\tau} \in \boldsymbol{\mathcal{T}}$ is a *partition of the unity.* A common choice for the coefficients β_j which enjoys desirable theoretical properties is the *balance heuristic* (BH, Veach & Guibas, 1995):

$$\beta^{\text{BH}}_j(\boldsymbol{\tau}) := \frac{n_j p_{\boldsymbol{\theta}^b_j}(\boldsymbol{\tau})}{\sum_{k=1}^{m} n_k p_{\boldsymbol{\theta}^b_k}(\boldsymbol{\tau})} = \frac{n_j \prod_{t=0}^{T-1} \pi_{\boldsymbol{\theta}^b_j}(a_t|s_t)}{\sum_{k=1}^{m} n_k \prod_{t=0}^{T-1} \pi_{\boldsymbol{\theta}^b_k}(a_t|s_t)}. \tag{6}$$

[1] For a sufficiently large length, namely $T \geqslant (1-\gamma)^{-1} \log\left(\epsilon^{-1} R_{\max}(1-\gamma)^{-1}\right)$, the finite-horizon γ-discounted expected return is ϵ-close to its infinite-horizon counterpart (Kearns & Singh, 2002). For this reason, we will use the two interchangeably, and just make sure $T \simeq (1-\gamma)^{-1}$ in our simulations.

[2] if dataset $\mathcal{D}_{\text{off}}$ is made of just one trajectory $\boldsymbol{\tau}$, with little abuse of notation, we denote the estimator by $\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})$.

The resulting estimator becomes:

$$\widehat{\nabla} J(\boldsymbol{\theta}; \mathcal{D}_{\text{off}}) = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{ij})}{\sum_{k=1}^{m} \frac{n_j}{n} p_{\boldsymbol{\theta}_k^b}(\boldsymbol{\tau}_{ij})} \mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{ij}), \quad \boldsymbol{\tau}_{ij} \sim p_{\boldsymbol{\theta}_j^b}, \quad \forall i \in [n_j],\ \forall j \in [m], \tag{7}$$

where $n = \sum_{j=1}^{m} n_j$ is the total number of trajectories. The *(multiple) importance weight* can be interpreted as the (single) importance weight having as a behavioral distribution the mixture of the m behavioral distributions with weights $\frac{n_j}{n}$, i.e., $\Phi_m := \sum_{k=1}^{m} \frac{n_j}{n} p_{\boldsymbol{\theta}_k^b}$ (Metelli et al., 2020).

When the set of behavioral policy parameters contains the target policy parameter $\boldsymbol{\theta}$ too, we speak of *defensive (multiple) off-policy* gradient estimation Owen (2013). In such a case, the importance weight is guaranteed to be bounded.

3 Behavioral Policy Optimization

In this section, we introduce the *behavioral policy optimization* (BPO) problem we aim to solve in this paper. The BPO problem consists in finding the "best behavioral policy" $\pi_{\boldsymbol{\theta}^b}$ to be used for collecting the trajectories $\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^b}$ for estimating the policy gradient $\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})$ of the target policy $\pi_{\boldsymbol{\theta}}$. We formalize the notion of "best behavioral policy" as the one that minimizes the trace of the covariance matrix of the off-policy gradient estimator $\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})$ where $\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^b}$ (that we will refer to as *gradient variance*) induced by the candidate behavioral policy $\pi_{\boldsymbol{\theta}^b}$:[3]

$$p_{*,\boldsymbol{\theta}} \in \underset{p_{\boldsymbol{\theta}^b}\,:\,\boldsymbol{\theta}^b \in \boldsymbol{\Theta}}{\arg\min} \underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^b}}{\mathbb{V}\text{ar}} \left[\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau}) \right] := \underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^b}}{\mathbb{E}} \left[\left\| \widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau}) - \nabla J(\boldsymbol{\theta}) \right\|_2^2 \right]. \tag{8}$$

The trace is a common scalarization of the covariance matrix. Moreover, controlling the trace of the covariance of the gradient estimate is enough to establish finite-time convergence guarantees for SGD algorithms (Ghadimi & Lan, 2013). The optimization problem of Equation (8) can be challenging since it involves a minimization over the parameter space $\boldsymbol{\Theta}$, which can determine, in general, a non-convex optimization problem. In Section 3.1, we show that when extending the optimization over the full set of distributions over the trajectory space $\boldsymbol{\mathcal{T}}$, we can solve the BPO problem in closed form. In Section 3.2, we illustrate how the closed-form solution can be employed to learn a policy that induces a trajectory distribution representable within the policy parameters space $\boldsymbol{\Theta}$ approximately close to the best one.

3.1 Closed-form solution

In this section, we study the solution of the problem of Equation (8) when no restriction to the representable trajectory distributions is enforced. Although this assumption is not realistic from the policy gradient perspective, given the fact that the transition model of the environment is not under control and the policy space might be constrained to the specific parametrization $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, it represents an important preliminary step for obtaining a practical algorithm. The following result provides a closed-form solution to the BPO problem.

Theorem 1. *Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{g}_{\boldsymbol{\theta}} : \boldsymbol{\mathcal{T}} \to \mathbb{R}^d$ be the single-trajectory gradient estimator used to compute $\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})$. The solution $p_{*,\boldsymbol{\theta}} \in \Delta^{\boldsymbol{\mathcal{T}}}$ to the BPO problem (Equation 8) is given by:*

$$p_{*,\boldsymbol{\theta}}(\boldsymbol{\tau}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2}{\int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \mathrm{d}\boldsymbol{\tau}}. \tag{9}$$

The optimal value of Equation (8) is given by:

$$\underset{\boldsymbol{\tau} \sim p_{*,\boldsymbol{\theta}}}{\mathbb{V}\text{ar}} \left[\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau}) \right] = \underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}}{\mathbb{E}} \left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \right]^2 - \|\nabla J(\boldsymbol{\theta})\|_2^2 . \tag{10}$$

[3] In the following, we will continue employing the policy gradient notation, although the presented result hold for the estimation of the expected value of a general vector-valued function.

It is worth comparing the result of Equation (10) with the variance of the on-policy gradient estimator that can be easily computed from Equation (8):

$$\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}\left[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})\right] = \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\right] - \|\nabla J(\boldsymbol{\theta})\|_2^2 . \tag{11}$$

Although the subtracted term $\|\nabla J(\boldsymbol{\theta})\|_2^2$ is the same in (11) and (10), the first one presents some differences. Indeed, in Equation (11) we have an *expectation of the squared L_2-norm* of the single-trajectory gradient estimator, i.e., $\mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\right]$, whereas in Equation (10), we have the *squared expectation of the L_2-norm* of the single-trajectory gradient estimator, i.e., $\mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right]^2$. From Jensen's inequality, we immediately observe that:

$$\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right]^2 \leqslant \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\right], \tag{12}$$

and, consequently, we conclude that the off-policy gradient estimator with $p_{*,\boldsymbol{\theta}}$ as behavioral distribution suffers a smaller variance compared with the on-policy gradient estimator.

Furthermore, it is worth comparing the result of Theorem 1 with the well-known result for minimum-variance estimation of expectation for non-negative scalar functions (Kahn, 1950). Indeed, Theorem 1 generalizes this result for vector-valued functions, reducing to the classical result for non-negative scalar functions, with the standard zero-variance estimator.

As already noted at the beginning of the section, although a convenient closed-form expression for the trajectory density function exists, it cannot be used in practice to collect trajectories since no policy exists inducing such a trajectory distribution. Nevertheless, it can be employed to learn a policy that induces a distribution as close as possible to this one.

3.2 Cross-entropy minimization

In this section, we illustrate how to employ the closed-form solution of the BPO problem derived in Section 3.1 in order to obtain a practical algorithm. Since, in practice, the parameter space $\boldsymbol{\Theta}$, together with the transition model, allows to span of a subset of the trajectory distributions $\Delta^{\mathcal{T}}$, we cannot represent the optimal behavioral distribution p_* by means of a parametrization, i.e., there not exists $\boldsymbol{\theta}^b_* \in \boldsymbol{\Theta}$ such that $p_{*,\boldsymbol{\theta}} = p_{\boldsymbol{\theta}^b_*}$ a.s. However, we can conveniently project it into the space of representable behavioral distributions by minimizing the KL divergence:

$$\boldsymbol{\theta}^b_\dagger \in \underset{\boldsymbol{\theta}^b\in\boldsymbol{\Theta}}{\arg\min}\, D_{\mathrm{KL}}\left(p_{*,\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}^b}\right). \tag{13}$$

This minimization problem can be further simplified into a weighted cross-entropy minimization by exploiting the functional form of $p_{*,\boldsymbol{\theta}}$, as shown in the following result.

Proposition 3.1. *Let $p_{*,\boldsymbol{\theta}}$ as defined in Equation (9). Then, the solution to the problem in Equation (13) can be obtained via the weighted cross-entropy minimization:*

$$\boldsymbol{\theta}^b_\dagger \in \underset{\boldsymbol{\theta}^b\in\boldsymbol{\Theta}}{\arg\min} \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}\left[-\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\| \log p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau})\right] = \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}\left[-\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\| \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}^b}(a_t|s_t)\right]. \tag{14}$$

This alternative formulation has the advantage that the objective function is expressed as an expected value w.r.t. the trajectory distribution induced by the target policy, which can be estimated either on- or off-policy. In the most general case, we can resort to (multiple) off-policy estimation:

$$\widehat{\boldsymbol{\theta}}^b_\dagger \in \underset{\boldsymbol{\theta}^b\in\boldsymbol{\Theta}}{\arg\min} \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{ij})}{\boldsymbol{\Phi}_m(\boldsymbol{\tau}_{ij})} \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_{ij})\| \log p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau}_{ij}), \quad \boldsymbol{\tau}_{ij} \sim p_{\boldsymbol{\theta}^b_j}, \quad \forall i \in [n_j],\ \forall j \in [m]. \tag{15}$$

In the general case, a closed-form solution may not be available, but we can still resort to iterative optimization techniques such as gradient descent. In practice, it is common to use Gaussian or softmax policies parametrized by neural networks. In this case, by using over-parametrized networks, we expect to find good behavior policies even if the objective is non-convex (Du et al., 2019).

Algorithm 1 Policy Gradient with Behavioral Policy Optimization.

1: **Input:** initial target policy parameters $\boldsymbol{\theta}_0$, batch sizes $N_{\text{BPO}}, N_{\text{PG}}$, step size α, defensive parameter β
2: **for** $k = 0, \dots, K-1$ **do**
3: $\quad\mathcal{D}_k^{\text{BPO}} = \{N_{\text{BPO}}$ trajectories collected with $\boldsymbol{\theta}_k\}$
4: $\quad\widetilde{\boldsymbol{\theta}}_k \leftarrow$ Solve (approximately) Equation (13) with $\mathcal{D}_k^{\text{BPO}}$
5: $\quad\mathcal{D}_k^{\text{PG}} = \left\{(1-\beta)N_{\text{PG}} \text{ trajectories } \boldsymbol{\tau} \sim p_{\widetilde{\boldsymbol{\theta}}_k} \text{ and } \beta N_{\text{PG}} \text{ trajectories } \boldsymbol{\tau} \sim p_{\boldsymbol{\theta}_k}\right\}$
6: $\quad\boldsymbol{v}_k \leftarrow \widehat{\nabla} J(\boldsymbol{\theta}_k; \mathcal{D}_k^{\text{PG}})$
7: $\quad\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \alpha \boldsymbol{v}_k$
8: **end for**
9: **return** $\boldsymbol{\theta}_L$ with $L \sim \text{Uni}([K])$

4 Theoretical Analysis

In this section, we study the theoretical properties of Algorithm 1, with a focus on the variance reduction granted by the active-IS estimator and how this impacts the rate of convergence of policy gradient to stationary points of the expected-return objective.

The quality of the policy gradient update will ultimately depend on how close our behavior policy is to the optimal one, and this cannot be ignored when deciding how many samples N_{BPO} are allocated to approximately solving Equation (13) in Line 4 of the algorithm. In Section 4.1, we first study the problem in full generality, assuming access to an ϵ-minimizer of Equation (13). We remove this assumption in Section 4.2, studying the convergence rate for a specific but broad class of policies.

4.1 Behavior Policy Optimization Oracle

The following lemma shows the relationship between the variance of the off-policy estimator and the distance, in terms of chi-square divergence, between the chosen behavior distribution and the optimal one. It is given in terms of the variance reduction over Monte Carlo (on-policy) estimation.

Lemma 4.1. *Fix a target policy $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a behavior trajectory distribution $q \in \Delta^{\mathcal{T}}$. Let $\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \boldsymbol{\tau})$ be the importance-weighted estimate of $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ computed with $\boldsymbol{\tau} \sim q$. Then the variance reduction from using q in place of $p_{\boldsymbol{\theta}}$ is given by:*

$$\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\text{ar}}\left[\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta};\boldsymbol{\tau})\right] - \underset{\boldsymbol{\tau}\sim q}{\mathbb{V}\text{ar}}\left[\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta};\boldsymbol{\tau})\right] = \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\text{ar}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right] - Z_{\boldsymbol{\theta}}^2 \chi^2(p_{*,\boldsymbol{\theta}} \| q),$$

where $Z_{\boldsymbol{\theta}} := \mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2]$.

This lemma shows that the variance reduction depends on how closely we can approximate the optimal behavior distribution in terms of chi-square divergence. Unfortunately, the latter is hard to optimize from data. Using defensive samples reduces this to a KL-divergence error, which is much easier to control. In this section, we just observe that the KL divergence can be made small using the approach proposed in Section 3.2, and operate under the following, more abstract:

Assumption 1 (BPO Oracle)**.** *For any target policy parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $p_{*,\boldsymbol{\theta}}$ be the corresponding optimal behavior distribution as defined in Equation (8). We assume access to a Behavioral Policy Optimization oracle* BPO : $\boldsymbol{\Theta} \to \boldsymbol{\Theta}$ *that takes a target policy parameter $\boldsymbol{\theta}$ and returns a behavior policy parameter $\widetilde{\boldsymbol{\theta}}$ such that:*

$$D_{\text{KL}}\left(p_{*,\boldsymbol{\theta}} \| p_{\widetilde{\boldsymbol{\theta}}}\right) \leqslant \epsilon_{\text{KL}},$$

for some constant $\epsilon_{\text{KL}} \geqslant 0$ independent of $\boldsymbol{\theta}$.

The following theorem upper-bounds the excess variance in terms of the KL-divergence and provides a principled way to choose the defensive parameter β in Algorithm 1.

Theorem 2. *Fix a target policy $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a behavior policy $\widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$. Let $\beta \in [0,1]$ and let $\Phi = \beta p_{\boldsymbol{\theta}} + (1-\beta) p_{\widetilde{\boldsymbol{\theta}}}$ be the mixture trajectory distribution. Let $\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta};\boldsymbol{\tau})$ be the β-defensive importance-weighted estimate of $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ computed with $\boldsymbol{\tau} \sim \Phi$. Then the variance reduction from* *using* Φ *in place of* $p_{\boldsymbol{\theta}}$ *is at least*

$$\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}\left[\widehat{\nabla}J(\boldsymbol{\theta};\boldsymbol{\tau})\right]-\underset{\boldsymbol{\tau}\sim\Phi}{\mathbb{V}\mathrm{ar}}\left[\widehat{\nabla}_{\boldsymbol{\theta}}J(\boldsymbol{\theta};\boldsymbol{\tau})\right]\geqslant\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right]-4Z_{\boldsymbol{\theta}}(Z_{\boldsymbol{\theta}}+\beta G_{\boldsymbol{\theta}})\left(2+\frac{1-\beta}{\beta}D_{\mathrm{KL}}(p_{*,\boldsymbol{\theta}}\|p_{\widetilde{\boldsymbol{\theta}}})\right),$$

where $Z_{\boldsymbol{\theta}}=\mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2]$ *and* $G_{\boldsymbol{\theta}}=\operatorname{ess\,sup}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}\{\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\}$. *Under Assumption 1, provided* $\epsilon_{\mathrm{KL}}\leqslant 1$, *by setting* $\beta=\sqrt{\frac{\epsilon_{\mathrm{KL}}}{2-\epsilon_{\mathrm{KL}}}}$, *the variance reduction is at least*

$$\begin{aligned}\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}\left[\widehat{\nabla}J(\boldsymbol{\theta};\boldsymbol{\tau})\right]-\underset{\boldsymbol{\tau}\sim\Phi}{\mathbb{V}\mathrm{ar}}\left[\widehat{\nabla}_{\boldsymbol{\theta}}J(\boldsymbol{\theta};\boldsymbol{\tau})\right]&\geqslant\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right]-4Z_{\boldsymbol{\theta}}^2(2-\epsilon_{\mathrm{KL}})-4Z_{\boldsymbol{\theta}}G_{\boldsymbol{\theta}}\epsilon_{\mathrm{KL}}\\&\quad-4Z_{\boldsymbol{\theta}}(Z_{\boldsymbol{\theta}}+G_{\boldsymbol{\theta}})\sqrt{\epsilon_{\mathrm{KL}}(2-\epsilon_{\mathrm{KL}})}\qquad(16)\\&\geqslant\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right]-8Z_{\boldsymbol{\theta}}^2-4Z_{\boldsymbol{\theta}}(Z_{\boldsymbol{\theta}}+2G_{\boldsymbol{\theta}})\sqrt{\epsilon_{\mathrm{KL}}}.\qquad(17)\end{aligned}$$

Remark 4.1. *As* $\epsilon_{\mathrm{KL}}\to 0$, *we have* $\mathbb{V}\mathrm{ar}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\widehat{\nabla}J(\boldsymbol{\theta};\boldsymbol{\tau})]-\mathbb{V}\mathrm{ar}_{\tilde{\boldsymbol{\tau}}\sim\Phi}[\widehat{\nabla}J(\boldsymbol{\theta};\boldsymbol{\tau})]\geqslant\mathbb{V}\mathrm{ar}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2]-8Z_{\boldsymbol{\theta}}^2-o(\sqrt{\epsilon_{\mathrm{KL}}})$. *Thus, if the KL-divergence is small enough, we there is variance reduction if*

$$\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right]=\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}\left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\right]-Z_{\boldsymbol{\theta}}^2>9Z_{\boldsymbol{\theta}}^2,\qquad(18)$$

that is, when $\mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2]>10Z_{\boldsymbol{\theta}}^2$. *To see that variance reduction is indeed possible, consider the example: let* $\boldsymbol{\mathcal{T}}=\{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2\}$ *and the target distribution is* p_θ *such that* $p_\theta(\boldsymbol{\tau}_1)=\theta$ *and* $p_\theta(\boldsymbol{\tau}_2)=1-\theta$, *with* $\theta\in[0,1]$. *Suppose* $g_\theta(\boldsymbol{\tau}_1)\in\{1,-1\}$ *and* $g_\theta(\boldsymbol{\tau}_2)=0$ *for all* θ. *Then* $\mathbb{E}_{\boldsymbol{\tau}\sim p_\theta}[|g_\theta(\boldsymbol{\tau})|^2]=\theta$, *while* $Z_\theta^2=\mathbb{E}_{\boldsymbol{\tau}\sim p_\theta}[|g_\theta(\boldsymbol{\tau})|]^2=\theta^2$. *So we can be sure there is variance reduction as long as* $\theta<1/10$.

We can use this result on variance reduction to upper bound the variance of the policy gradient estimates computed by our algorithm. In the following, let $\mathcal{F}_k$ denote the sigma-algebra generated by all the random variables from Algorithm 1 up to iteration $k-1$ included, and all the trajectories from $\mathcal{D}_k^{\mathrm{BPO}}$. Note that both $\boldsymbol{\theta}_k$ and $\widetilde{\boldsymbol{\theta}}_k$ are $\mathcal{F}_k$-measurable. For brevity, we will write $\mathbb{E}_k[X]$ for the conditional expectation $\mathbb{E}[X|\mathcal{F}_k]$, and $\mathbb{V}\mathrm{ar}_k[X]$ for the conditional variance $\mathbb{V}\mathrm{ar}[X|\mathcal{F}_k]=\mathbb{E}_k[\|X-\mathbb{E}_k[X]\|_2^2]$ of a random element X.

Theorem 3. *Fix an iteration* $k\in[K]$ *of Algorithm 1 and let* $\mathcal{D}_{\mathrm{ON}}$ *denote a dataset of* N_{PG} *independent trajectories collected with* $\boldsymbol{\theta}_k$. *Under Assumption 1, the variance reduction granted by using the off-policy estimator* $\mathbf{v}_k:=\widehat{\nabla}J(\boldsymbol{\theta}_k;\mathcal{D}_k^{\mathrm{PG}})$ *with respect to an on-policy estimator is given by:*

$$\underset{k}{\mathbb{V}\mathrm{ar}}\left[\widehat{\nabla}J(\boldsymbol{\theta}_k;\mathcal{D}_{\mathrm{ON}})\right]-\underset{k}{\mathbb{V}\mathrm{ar}}[\mathbf{v}_k]\geqslant\frac{1}{N_{\mathrm{PG}}}\left(V_k-8Z_k^2-4Z_k(Z_k+2G_k)\sqrt{\epsilon_{\mathrm{KL}}}\right),\qquad(19)$$

where $Z_k:=\mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}_k}}[\|\mathbf{g}_{\boldsymbol{\theta}_k}(\boldsymbol{\tau})\|_2|\mathcal{F}_k]$, $V_k:=\mathbb{V}\mathrm{ar}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}_k}}[\|\mathbf{g}_{\boldsymbol{\theta}_k}(\boldsymbol{\tau})\|_2|\mathcal{F}_k]$, *and* $G_k:=\operatorname{ess\,sup}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}_k}}\{\|\mathbf{g}_{\boldsymbol{\theta}_k}(\boldsymbol{\tau})\|_2\}$. *Thus, the conditional variance of* $\mathbf{v}_k$ *is upper-bounded as follows:*

$$\underset{k}{\mathbb{V}\mathrm{ar}}[\mathbf{v}_k]\leqslant\frac{1}{N_{\mathrm{PG}}}\left(9Z_k^2+Z_k(Z_k+2G_k)\sqrt{\epsilon_{\mathrm{KL}}}-\|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right).\qquad(20)$$

4.2 Convergence Rate

So far, we studied the variance of the active-IS estimator from Algorithm 1, showing that variance reduction is possible whenever the KL divergence between the optimal behavior distribution $p_{\boldsymbol{\theta},*}$ and its estimate $p_{\widetilde{\boldsymbol{\theta}}}$ is small enough. We now give a more concrete characterization of the variance reduction in terms of how many on-policy samples are used to compute $p_{\widetilde{\boldsymbol{\theta}}}$. We are only able to do so for a restricted class of policies, namely *exponential-family* policies with linear sufficient statistics. However, this is a broad class that includes linear Gaussian and Softmax policies. Furthermore, this is the class of policies for which the (empirical) cross-entropy minimization problem described in Section 3.2 admits a closed-form solution. Thus, it represents a setting where sample and computational efficiency can be achieved at the same time. Our analysis will also provide a principled way to allocate a per-iteration budget of N trajectories in Algorithm 1, that is, how to split them into N_{BPO} trajectories for behavior policy optimization, and N_{PG} trajectories for gradient estimation.

We begin by listing all the assumptions that we will use in this section.

Assumption 2 (Exponential-Family Policy)**.** *The target policy is of the form:*

$$\pi_{\boldsymbol{\theta}}(a|s) = h(a) \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(s,a) - A(\boldsymbol{\theta}, s)\right), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *is the sufficient statistic,* $h : \mathcal{A} \to \mathbb{R}_+$*, and* $A(\boldsymbol{\theta}, s) = \log \int_{\mathcal{A}} h(a) \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(s,a)\right) \mathrm{d}a$ *is the log-partition function.*

This general model allows to conveniently represent widely used policies, including Gaussian policies with linear mean and Softmax policies (Metelli et al., 2023). Note that, for a policy satisfying Assumption 2, the score function is $\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) = \boldsymbol{\varphi}(s,a) - \mathbb{E}_{a' \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{\varphi}(s,a')] =: \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s,a)$, and also $\nabla^2_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) = -\operatorname{\mathbb{C}ov}_{a' \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{\varphi}(s,a')]$. We will refer to $\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}$ as the *centered* sufficient statistic. We now introduce a necessary assumption to guarantee that the optimal behavioral distribution over trajectories $p_{*,\boldsymbol{\theta}^\dagger}$ is representable within the ones induced by the policies $\pi_{\boldsymbol{\theta}^*}$ with $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$.

Assumption 3 (Realizability)**.** *For any target policy* $\boldsymbol{\theta}^\dagger \in \boldsymbol{\Theta}$*, there exists* $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ *s.t. the optimal behavior distribution w.r.t.* $\boldsymbol{\theta}^\dagger$ *is* $p_{*,\boldsymbol{\theta}^\dagger} = p_{\boldsymbol{\theta}^*}$*, the trajectory distribution induced by policy* $\pi_{\boldsymbol{\theta}^*}$*.*

The next assumption is related to the tail behavior of the noise

Assumption 4 (Subgaussianity)**.** *For any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and* $s \in \mathcal{S}$*, the centered sufficient statistic* $\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s,\cdot)$ *is* σ*-subgaussian in the sense that, for any* $\boldsymbol{\lambda} \in \mathbb{R}^d$*:*

$$\mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\exp\left(\boldsymbol{\lambda}^\top \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s,a)\right)\right] \leqslant \exp\left(\frac{\|\boldsymbol{\lambda}\|_2^2 \sigma^2}{2}\right), \qquad \forall s \in \mathcal{S}.$$

Finally, we enforce the following assumption that prescribes an exploration condition of the played policy encoded in a property of the spectrum of the empirical Fisher information matrix.

Assumption 5 (Explorability)**.** *For a fixed target policy* $\boldsymbol{\theta}^\dagger \in \boldsymbol{\Theta}$ *and a dataset of* n *trajectories* $\{\boldsymbol{\tau}_i\}_{i \in [n]}$ *collected with* $\pi_{\boldsymbol{\theta}^\dagger}$ *let*

$$\widehat{\mathcal{F}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{g}_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau}_i)\|_2 \sum_{t=0}^{T-1} \mathop{\mathbb{C}\mathrm{ov}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t^i)}[\boldsymbol{\varphi}(s_t^i, a)]. \tag{21}$$

We assume that, for all $n \geqslant 1$ *and* $\boldsymbol{\theta}^\dagger, \boldsymbol{\theta} \in \boldsymbol{\Theta}$*,* $\mathbb{E}\left[\lambda_{\min}(\widehat{\mathcal{F}}(\boldsymbol{\theta}))\right] \geqslant \lambda_* > 0$*.*

Given the previously listed assumptions, we are able to provide a meaningful bound on the expected error expressed in KL-divergence between the optimal behavioral trajectory distribution $p_{*,\boldsymbol{\theta}}$ and the one estimated by the cross entropy minimization procedure $\widetilde{\boldsymbol{\theta}}$.

Lemma 4.2. *Fix a target policy parameter* $\boldsymbol{\theta}^\dagger \in \boldsymbol{\Theta}$ *and let* $\{\boldsymbol{\tau}_i\}_{i \in [n]}$ *be a dataset of* n *i.i.d. trajectories collected with* $\pi_{\boldsymbol{\theta}^\dagger}$*. Let*

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \|\mathbf{g}_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau}_i)\|_2 \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}}(a_t^i | s_t^i),$$

and if $\operatorname{ess\,sup}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}} \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \leqslant G$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*. Then, under Assumptions 2, 3, 4, 5 it holds that:*

$$\mathbb{E}\left[D_{\mathrm{KL}}(p_{*,\boldsymbol{\theta}^\dagger} \| p_{\widetilde{\boldsymbol{\theta}}})\right] \leqslant \frac{G^2 T^3 \sigma^4}{2\lambda_*^2 n}.$$

We are now ready to quantify the complete variance of the defensive off-policy estimator.

Theorem 4. *Assuming* $N_{\mathrm{BPO}} > \frac{G^2 T^3 \sigma^4}{2\lambda_*^2}$*, let* $\epsilon^* = \frac{G^2 T^3 \sigma^4}{2\lambda_*^2 N_{\mathrm{BPO}}}$*. Then, under Assumptions 2, 3, 4, 5, Algorithm 1 with* $\beta = \sqrt{\epsilon^*/(2-\epsilon^*)}$ *guarantees*

$$\mathop{\mathbb{V}\mathrm{ar}}_{k}[\mathbf{v}_k] \leqslant \frac{1}{N_{\mathrm{PG}}} \left(9 Z_k^2 + \frac{Z_k (Z_k + 2G) G T^{3/2} \sigma^2}{\lambda_* \sqrt{2 N_{\mathrm{BPO}}}} - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right). \tag{22}$$

Furthermore, by setting $N_{\text{BPO}} = N_{\text{PG}} = \frac{N}{2}$ *and* $\beta \in (0,1)$*, provided* $N > \frac{G^2T^3\sigma^4(1+\beta^2)}{2\lambda_*^2\beta^2}$ *we have:*

$$\underset{k}{\mathbb{V}\text{ar}}[\mathbf{v}_k] \leqslant \frac{1}{N}\left(18Z_k^2 - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right) + \frac{Z_k(Z_k+2G)GT^{3/2}\sigma^2}{2\lambda_* N^{3/2}}. \tag{23}$$

We are finally able to provide the convergence rate of the corresponding iterative optimization.

Corollary 1. *Let* $\widetilde{V} := 18Z_k^2 - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2$ *denote the residual variance left by the BPO process. Under the assumptions of Theorem 4, a total number of trajectories*

$$NK \leqslant \left\lceil 12(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))\left(\frac{3C_1\widetilde{V}}{\epsilon^4} + \frac{C_1+3C_2}{\epsilon^{10/3}}\right)\right\rceil$$

is sufficient for Algorithm 1 to obtain $\mathbb{E}[\|\nabla J(\boldsymbol{\theta}_{out})\|_2] \leqslant \epsilon$*, where* $\boldsymbol{\theta}_{out}$ *is chosen uniformly at random from the iterates* $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_K$ *of the algorithm, where* $C_1 = \frac{R_{\max}\sigma^2}{(1-\gamma)^2}$ *and* $C_2 = \frac{R_{\max}^4\sigma^5\|\boldsymbol{\varphi}\|_\infty(\sqrt{T}\sigma+2T\|\boldsymbol{\varphi}\|_\infty)T^3}{2\lambda_*(1-\gamma)^5}$.

Remark 4.2. *Although, in the worst case, the sample complexity is* $O(\epsilon^{-4})$ *like on-policy REINFORCE (Yuan et al., 2022), when the residual variance* $\widetilde{V}$ *is negligible, namely,* $\widetilde{V} = O(\epsilon^{2/3})$*, Algorithm 1 can achieve an improved sample complexity of* $O(\epsilon^{-10/3})$*, the same as SVRPG (Papini et al., 2018). Examples of this can be constructed as in Remark 4.1. Even though the optimal sample complexity for first-order policy optimization is* $O(\epsilon^{-3})$ *(Xu et al., 2020a) and our* $\epsilon^{2/3}$ *improvement does not hold in full generality, we are not aware of any other case of provable acceleration of policy gradient algorithms following from behavior-policy optimization.*

5 Related Works

Baselines A common technique from statistical simulation to reduce variance in policy gradient estimation is using the *baselines*. A baseline $\mathbf{b}$ is a non-random quantity that is subtracted from the return $R(\boldsymbol{\tau})$ based on the observation that $\mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{\tau})R(\boldsymbol{\tau})] = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\nabla \log p_{\boldsymbol{\theta}}(\boldsymbol{\tau})(R(\boldsymbol{\tau})-\mathbf{b})]$. Optimal baselines for the REINFORCE and G(PO)MDP estimators have been derived by Peters & Schaal (2006). Other approaches exploit a baseline that is obtained from a moving average of the most recent returns (Weaver & Tao, 2001; Zhao et al., 2011). This approach is similar to using a critic to estimate the value function (Mei et al., 2022). The effectiveness of a baseline is highly problem-dependent and, in the end, does not change the convergence rate of the policy gradient algorithm, which remains of order $O(\epsilon^{-4})$, being ϵ the expected norm of the policy gradient reached.

Variance-Reduced Policy Gradient Algorithms *Variance reduction* techniques have been first introduced for supervised learning, having SVRG (Johnson & Zhang, 2013) as progenitor. The idea consists of re-using snapshots of gradients computed in the past to exploit the correlation in order to reduce the variance. Still, in the supervised learning community, several variations and improvements have been presented, which include SARAH (Nguyen et al., 2017), STORM (Cutkosky & Orabona, 2019) and PAGE (Li et al., 2021). Each of these has been adapted to the policy gradient setting, giving rise to SVRPG (Papini et al., 2018), SRVR-PG (Xu et al., 2020c), STORMPG (Yuan et al., 2020), and PAGEPG (Gargiani et al., 2022), respectively. These approaches have succeeded in strictly improving the convergence rate over standard PGs. Indeed, SVRPG archives a convergence rate of order $O(\epsilon^{-10/3})$, as shown by Xu et al. (2020b), while SRVR-PG, STORMPG, and PAGEPG outperform it with a convergence rate of order $O(\epsilon^{-3})$, which is currently conjectured to be optimal.

Active Importance Sampling In Hanna et al. (2017), the problem of *behavioral policy search* is addressed with the goal of finding the most effective (i.e., minimum variance) behavioral policy to estimate the expected return of a given target policy. The approach is based on a gradient method that optimizes the policy parameters in order to find the minimum-variance behavioral policy. Although the approach demonstrated advantages from the policy evaluation perspective, it struggles to extend to policy optimization. In Hanna (2019), the extension to the optimization perspective has been provided with a *parallel policy search* approach that simultaneously optimizes over the parameters of the behavioral and target policies. Unfortunately, the algorithm enjoys no theoretical guarantees and shows limited empirical advantages. Recently, in Metelli et al. (2023), the authors have deepened the connections between minimum-variance behavioral policy and the policy optimization have been studied. Specifically, under certain assumptions on the policy space, it is possible to show that the minimum variance behavioral policy attains a performance improvement. However, these works lack a comprehensive theoretical analysis capable of quantifying analytically the actual advantage of *active IS*, possibly in terms of convergence rate.

6 Numerical Simulations

In this section, we first provide a practical version of Algorithm 1 and then provide the experimental results on classical control tasks.

6.1 Practical Algorithm

Here, we present some practical aspects related to the implementation of Algorithm 1, based on the above-introduced idea of IS estimators. In particular, in Algorithm 1, we face two estimation problems: the estimation of KL divergence in Line 4 and the off-policy gradient estimation in Line 6. Both can benefit from effectively reusing already collected trajectories during the algorithm execution so as to reduce the overall number of samples generated per iteration.

Offline KL divergence, Line 4 In place of collecting, at every iteration k, new N_{BPO} trajectories with the current target policy $\pi_{\boldsymbol{\theta}_k}$ to build the dataset $\mathcal{D}_k^{\text{BPO}}$, we reuse the samples for the off-policy gradient estimation at the previous iteration $k-1$, namely $\mathcal{D}_{k-1}^{\text{PG}}$. We call this KL estimation *offline*, as it employs trajectories from the previous target and behavioral policies $\pi_{\boldsymbol{\theta}_{k-1}}$ and $\pi_{\tilde{\boldsymbol{\theta}}_{k-1}}$. Such offline samples need to be re-weighted proportionally to the probability of being generated by the current target policy $\pi_{\boldsymbol{\theta}_k}$, for which we resort to the (multiple) off-policy estimator in Equation 15.

Biased off-policy gradient, Line 6 The off-policy gradient estimation in Algorithm 1 is computed with the only behavioral policy $\pi_{\tilde{\boldsymbol{\theta}}_k}$ and, when the defensive strategy is used $\beta > 0$,[4] with the current target policy $\pi_{\boldsymbol{\theta}_k}$. To increase the number of trajectories available for the gradient estimation, we can reuse the already collected trajectories for the (offline) KL divergence estimation, namely $\mathcal{D}_k^{\text{BPO}}$. This approach is a multiple off-policy gradient estimator. If the offline KL strategy is employed, this means using the target policy $\pi_{\boldsymbol{\theta}_{k-1}}$ at the previous iteration as an additional behavioral policy. Otherwise, $\mathcal{D}_k^{\text{BPO}}$ contains *biased* defensive samples from the current target policy $\pi_{\boldsymbol{\theta}_k}$, as they were already used to compute the current behavioral policy $\pi_{\tilde{\boldsymbol{\theta}}_k}$.

6.2 Experimental Results

All experiments are conducted with Gaussian policies with fixed diagonal variance, and the mean is linearly parametrized in the state so that $\pi_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}^\top s, \sigma \mathbf{I})$. We first provide a set of numerical results on the Linear Quadratic Regulator (LQ) environment, quantifying the variance reduction of the single target policy gradient estimate; we then show the impact of such variance reduction on the learning iterations for solving the full control task in the Cartpole benchmark. We employed the G(PO)MDP gradient estimator and its optimal baselines as derived in Peters & Schaal (2006).

Variance Reduction In this set of experiments, we want to analyze the impact of the optimal behavioral policy in estimating the target policy gradient. In particular, we compare the gradient variance (as defined in Equation (8)) in the on-policy and the proposed off-policy setting. The optimal behavioral policy parameters $\widehat{\boldsymbol{\theta}}_{\dagger}^{b}$ were computed by solving (15), where the cross-entropy term was estimated by sampling N_{BPO} trajectories from the target policy $\pi_{\boldsymbol{\theta}}$. Afterwards, N_{PG}

[4] From theory, one can set some $\epsilon < 1$ as the desired accuracy of the BPO subroutine and then set $\beta = \sqrt{\frac{\epsilon}{2-\epsilon}}$. In practice, one can tune β like any other hyperparameter (e.g., the step size). See Appendix D for additional experimental results obtained with different choices of β.

Table 1: LQ environment, with horizon = 2 and state dimension = 1. Variance reduction in off-policy gradient, expressed as $\Delta\mathbb{V}\text{ar}$ and its 95% Gaussian confidence interval $(\Delta\mathbb{V}\text{ar}^-, \Delta\mathbb{V}\text{ar}^+)$, with different hyper-parameters.

(a) Target policy with $\log\sigma = 0$ and varying $\boldsymbol{\theta}$.

$\Delta\mathbb{V}\text{ar}$	$\Delta\mathbb{V}\text{ar}^-$	$\Delta\mathbb{V}\text{ar}^+$	biased	β	N_{BPO}	N_{PG}	$\boldsymbol{\theta}$
2.05	1.13	2.97	True	0.8	50	50	1.0
1.64	−0.10	3.39	True	0.0	10	90	1.0
1.50	0.78	2.23	True	0.4	50	50	1.0
1.39	0.32	2.45	False	0.0	10	90	1.0
1.26	0.63	1.89	True	0.0	50	50	1.0
1.15	−0.62	2.91	True	0.8	10	90	1.0
0.70	0.25	1.15	True	0.0	30	70	−1.0
0.56	−0.71	1.84	False	0.0	50	50	1.0
0.56	0.30	0.82	True	0.8	50	50	−1.0
0.51	0.03	0.98	True	0.0	10	90	−1.0
0.47	0.26	0.68	True	0.4	50	50	−1.0
0.41	0.14	0.67	True	0.4	50	50	0.5
0.40	0.18	0.61	True	0.0	50	50	−1.0
0.39	−0.02	0.80	False	0.0	10	90	0.5
0.32	0.16	0.49	True	0.0	30	70	0.5
0.32	−0.17	0.81	False	0.4	10	90	−1.0
0.31	−0.07	0.69	False	0.0	10	90	−1.0
0.31	−0.16	0.77	False	0.4	50	50	−1.0
0.30	0.07	0.52	True	0.8	50	50	0.5
0.29	−0.14	0.72	False	0.8	10	90	−1.0
0.27	−0.14	0.68	True	0.4	10	90	−1.0

(b) Target policy with $\boldsymbol{\theta} = 0$ and varying $\log\sigma$.

$\Delta\mathbb{V}\text{ar}$	$\Delta\mathbb{V}\text{ar}^-$	$\Delta\mathbb{V}\text{ar}^+$	biased	β	N_{BPO}	N_{PG}	$\log\sigma$
4.04	2.02	6.07	True	0.8	50	50	1.0
3.77	2.40	5.15	True	0.4	50	50	1.0
3.25	1.63	4.86	True	0.0	30	70	1.0
3.18	1.95	4.40	True	0.0	50	50	1.0
2.70	0.72	4.68	True	0.8	30	70	1.0
2.36	−0.39	5.11	True	0.4	30	70	1.0
2.06	0.52	3.59	True	0.0	10	90	1.0
1.54	−0.38	3.45	False	0.0	10	90	1.0
1.19	−0.69	3.06	False	0.0	30	70	1.0
0.60	−1.28	2.49	False	0.8	30	70	1.0
0.59	0.24	0.94	True	0.8	50	50	0.5
0.58	−1.89	3.05	False	0.0	50	50	1.0
0.56	0.22	0.90	True	0.0	50	50	0.5
0.48	0.19	0.78	True	0.4	50	50	0.5
0.42	0.15	0.69	True	0.8	30	70	0.5
0.39	−1.54	2.32	False	0.4	30	70	1.0
0.24	−0.04	0.52	True	0.8	10	90	0.5
0.23	−0.03	0.48	True	0.4	30	70	0.5
0.16	−0.28	0.60	True	0.0	10	90	0.5
0.16	−0.25	0.56	False	0.0	50	50	0.5
0.15	−0.20	0.50	False	0.0	30	70	0.5

trajectories were sampled from the behavioral $\pi_{\hat{\boldsymbol{\theta}}^b_\dagger}$ to build the data-set $\mathcal{D}_{\text{off}}$ and compute the off-policy gradient as in equation (3). The on-policy gradient estimations were instead obtained with a batch of $N_{\text{BPO}} + N_{\text{PG}}$ trajectories forming the data-set $\mathcal{D}_{\text{on}}$.

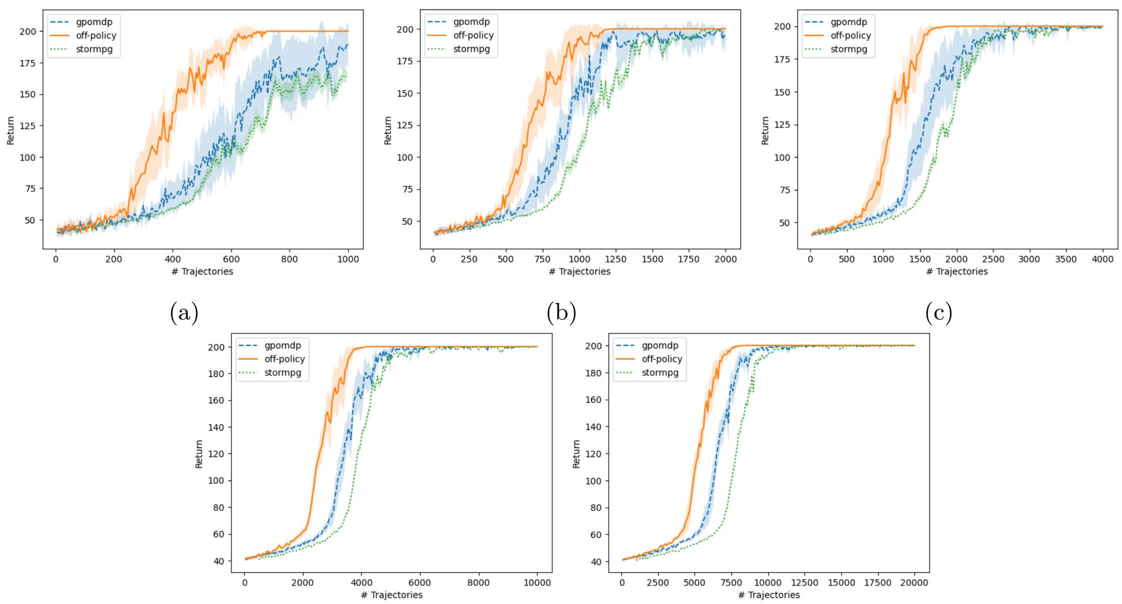
We run exhaustive experiments by varying the LQ horizon and the state dimensions. The complete results are reported in Appendix D. Here, we fix the horizon to 2 and consider mono-dimensional LQ problems varying parameters of the target policy, i.e., various $\boldsymbol{\theta} \in \{-1.0, -0.5, 0.0, 0.5, 1.0\}$ and log standard deviations $\log\sigma \in \{-1.0, -0.5, 0.0, 0.5, 1.0\}$. Finally, we varied also the hyper-parameters of our off-policy method, i.e. the defensive coefficient $\beta \in \{0, 0.4, 0.8\}$, the biased off-policy practical gradient calculation (the offline estimation of the KL divergence here is not possible), and the batch sizes N_{BPO} (10, 30 and 50) and $N_{\text{PG}} \in \{90, 70, 50\}$. Tables 1a and 1b report, for each environment and policy configuration, the first 20 results ordered by the average variance gap between the on-policy and off-policy methods (over 100 repetitions), i.e.:

$$\Delta\mathbb{V}\text{ar} = \frac{1}{100}\sum_{i=1}^{100}\left(\mathbb{V}\text{ar}\left[\widehat{\nabla} J(\boldsymbol{\theta}; \mathcal{D}^{(i)}_{\text{on}})\right] - \mathbb{V}\text{ar}\left[\widehat{\nabla} J(\boldsymbol{\theta}; \mathcal{D}^{(i)}_{\text{off}})\right]\right). \tag{24}$$

Across all the results, we can notice a few prevalent patterns. Firstly, as may be expected, the variance reduction is numerically more significant for "extreme" values of the policy parameters ($\boldsymbol{\theta}$ and $\log\sigma$ close to 1), as the gradient estimation problem becomes more and more difficult and prone to high variance, thus leading to significant margin of improvement (see also the complete results in Appendix D). Secondly, the biased off-policy gradient calculation is predominant in most of the highest variance reduction results, as it allows the use of the same number of samples of the on-policy counterpart. Lastly, the other off-policy hyper-parameters do not seem to impact these variance reduction results clearly, alternating different combinations in the best experiments reported in all the tables.

Learning Speed-up In this second set of experiments, we want to measure the impact of the variance reduction provided by our off-policy method in the learning process for solving the classic

Cartpole balancing problem and compare our results with the state-of-art variance reduction algorithm STORMPG. For our off-policy algorithm, we chose $\beta = 0$, and employed both the practical aspects of the offline KL divergence estimation (hence we do not use N_{BPO}) and of biased off-policy gradient estimation (see Section 6.1). All the experiments were run with a fixed budget of N_{PG} samples for each iteration, which also correspond to the mini batch-size employed by the STORMPG (the initial batch-size was set to 10 times N_{PG}). Figure 1 shows that our off-policy method outperforms the STOMRPG in all the different configurations, enjoying both a more stable behavior at convergence and a lower variance during the learning iterations.



(a) (b) (c)

(d) (e)

Figure 1: Cartpole. Average return and its 95% Gaussian CI (30 repetitions) over the learning iterations. Different policy gradient batch-sizes were used: (a) $N_{\mathrm{PG}} = 5$, (b) $N_{\mathrm{PG}} = 10$, (c) $N_{\mathrm{PG}} = 20$, (d) $N_{\mathrm{PG}} = 50$, (e) $N_{\mathrm{PG}} = 100$.

7 Discussion and Conclusions

In this paper, we have presented a novel approach to control the variance of the PG estimator. Leveraging the idea of looking for the best behavioral policy that minimizes the variance of the IS estimator, we have introduced a novel algorithm that exploits a two-phase procedure, alternating between the cross-entropy estimation of such a policy and the actual off-policy performance improvement. We have shown that, thanks to the defensive estimate, we are able to achieve a convergence rate of order $O(\epsilon^{-4})$ to a stationary point. Compared to the standard REINFORCE convergence rate, our algorithm enjoys a smaller residual variance. Then, we provided a practical version of such an algorithm, which uses all the samples collected so far at the price of an estimation bias. This algorithm was evaluated on benchmark continuous control tasks compared to standard baselines, showing a significant reduction of the estimation variance and a faster learning curve. Future works include studying other kinds of scalarization than the trace of the covariance matrix, the extension of the provided algorithm in combination with variance reduction techniques, such as SVRPG, and the conception of a more practical adaptation that suitably combines with deep architectures.

Acknowledgments

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.3 DD. 341 - 15 march 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 - PE00000013 - D53C22002380006.

References

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *J. Artif. Intell. Res.*, 15:319–350, 2001.

Sébastien Bubeck and Mark Sellke. First-order bayesian regret analysis of thompson sampling. In *ALT*, volume 117 of *Proceedings of Machine Learning Research*, pp. 196–233. PMLR, 2020.

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

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Dylan J. Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *NeurIPS*, pp. 18907–18919, 2021.

Matilde Gargiani, Andrea Zanelli, Andrea Martinelli, Tyler Summers, and John Lygeros. Page-pg: A simple and loopless variance-reduced policy gradient method with probabilistic gradient estimation. In *International Conference on Machine Learning*, pp. 7223–7240. PMLR, 2022.

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

Josiah P. Hanna, Philip S. Thomas, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1394–1403. PMLR, 2017.

Josiah Paul Hanna. *Data efficient reinforcement learning with off-policy and simulated data*. The University of Texas at Austin, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Herman Kahn. Random sampling (monte carlo) techniques in neutron attenuation problems. i. *Nucleonics (US) Ceased publication*, 6(See also NSA 3-990), 1950.

Michael J. Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 49(2-3):209–232, 2002.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pp. 6286–6295. PMLR, 2021.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. The role of baselines in policy gradient optimization. *Advances in Neural Information Processing Systems*, 35:17818–17830, 2022.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5447–5459, 2018.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21:141:1–141:75, 2020.

Alberto Maria Metelli, Samuele Meta, and Marcello Restelli. On the relation between policy improvement and off-policy minimum-variance policy evaluation. In *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1423–1433. PMLR, 2023.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takác. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2613–2621. PMLR, 2017.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4023–4032. PMLR, 2018.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *Mach. Learn.*, 111(11):4081–4137, 2022. URL https://doi.org/10.1007/s10994-022-06232-6.

Jan Peters and Stefan Schaal. Policy gradient methods for robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2006, October 9-15, 2006, Beijing, China*, pp. 2219–2225. IEEE, 2006.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pp. 3000–3006. AAAI Press, 2015.

Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1995, Los Angeles, CA, USA, August 6-11, 1995*, pp. 419–428. ACM, 1995.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pp. 538–545, 2001.

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

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *ICLR*. OpenReview.net, 2020a.

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

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *ICLR*. OpenReview.net, 2020c.

Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3332–3380. PMLR, 2022.

Tingting Zhao, Hirotaka Hachiya, Gang Niu, and Masashi Sugiyama. Analysis and improvement of policy gradient estimation. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 262–270, 2011.

A Hellinger Distance

The Hellinger distance between two distributions $P \ll Q$ is defined as[5]

$$D_H(P,Q) = \sqrt{\int_{\mathcal{T}} \left(\sqrt{p(\boldsymbol{\tau})} - \sqrt{q(\boldsymbol{\tau})}\right)^2 \mathrm{d}\boldsymbol{\tau}}. \tag{25}$$

In the following we list some known properties of the Hellinger distance that will be useful in our proofs. See, for instance, (Foster & Krishnamurthy, 2021).

- Boundedness: $D_H(P,Q) \leqslant \sqrt{2}$.
- The Hellinger distance is a metric. In particular, we will use symmetry, $D_H(P,Q) = D_H(Q,P)$, and the fact that $D_H(P,P) = 0$.
- The squared Hellinger distance is an f-divergence. In particular, we will use the joint convexity of f-divergences: $D_H^2(\beta P_1 + (1-\beta)P_2, \beta Q_1 + (1-\beta)Q_2) \leqslant \beta D_H^2(P_1,Q_1) + (1-\beta)D_H^2(P_2,Q_2)$. By taking $P_2 = Q_1 = Q_2$, we have $D_H^2(P, \beta P + (1-\beta)Q) \leqslant (1-\beta)D_H(P,Q)$.
- Pinsker-style inequality: $D_H(P,Q) \leqslant \sqrt{\min\{D_{\mathrm{KL}}(P\|Q), D_{\mathrm{KL}}(Q\|P)\}}$.

B Omitted Proofs

B.1 Proofs of Section 3

Theorem 1. *Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{g}_{\boldsymbol{\theta}} : \mathcal{T} \to \mathbb{R}^d$ be the single-trajectory gradient estimator used to compute $\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})$. The solution $p_{*,\boldsymbol{\theta}} \in \Delta^{\mathcal{T}}$ to the BPO problem (Equation 8) is given by:*

$$p_{*,\boldsymbol{\theta}}(\boldsymbol{\tau}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2}{\int_{\mathcal{T}} p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \mathrm{d}\boldsymbol{\tau}}. \tag{9}$$

The optimal value of Equation (8) is given by:

$$\underset{\boldsymbol{\tau} \sim p_{*,\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}} \left[\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})\right] = \underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}}{\mathbb{E}} \left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right]^2 - \|\nabla J(\boldsymbol{\theta})\|_2^2. \tag{10}$$

Proof. We consider a probability measure over the trajectory space $p \in \Delta^{\mathcal{T}}$. Let us first observe that since the off-policy estimator is unbiased, we can focus on the second moment:

$$\underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^b}}{\mathbb{V}\mathrm{ar}} \left[\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})\right] = \underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^b}}{\mathbb{E}} \left[\left\| \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau})} \mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau}) - \nabla J(\boldsymbol{\theta}) \right\|_2^2\right] \tag{26}$$

$$= \underset{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}}{\mathbb{E}} \left[\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau})}\right)^2 \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\right] - \|\nabla J(\boldsymbol{\theta})\|_2^2 \tag{27}$$

where the first inequality follows from the independence of the trajectories. Thus, we consider the following optimization problem, where the expectations are written with the corresponding integrals for convenience:

$$\min_{p \in \Delta^{\mathcal{T}}} \int_{\mathcal{T}} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})^2}{p(\boldsymbol{\tau})} \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2 \, \mathrm{d}\boldsymbol{\tau} \tag{28}$$

$$\text{s.t.} \quad \int_{\mathcal{T}} p(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} = 1 \tag{29}$$

$$p(\boldsymbol{\tau}) \geqslant 0 \qquad \forall \boldsymbol{\tau} \in \mathcal{T} \tag{30}$$

[5] In some texts, the Hellinger distance is normalized by $\sqrt{2}$ to be in $[0,1]$.

The problem has a convex objective function and linear constraints. Thus, we approach it with the Lagrange multipliers, dropping the non-negativity constraint that, as we shall see, will be already ensured by the derived solution. Let $\lambda \in \mathbb{R}$:

$$L(p(\cdot), \lambda) = \int_{\mathcal{T}} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})^2}{p(\boldsymbol{\tau})} \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2 \, \mathrm{d}\boldsymbol{\tau} + \lambda \left(\int_{\mathcal{T}} p(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} - 1 \right). \tag{31}$$

By vanishing the functional derivative w.r.t. $p(\cdot)$, we obtain for every $\boldsymbol{\tau} \in \mathcal{T}$:

$$\frac{\delta L(p(\cdot), \lambda)}{\delta p(\cdot)}(\boldsymbol{\tau}) = -\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})^2}{p(\boldsymbol{\tau})^2} \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2 + \lambda = 0 \implies p(\boldsymbol{\tau}) = \sqrt{\lambda} p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2, \tag{32}$$

having retained the non-negative solution only. Since for constraint 30, the density must integrate up to 1, we have that for every $\boldsymbol{\tau} \in \mathcal{T}$:

$$p(\boldsymbol{\tau}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2}{\int_{\mathcal{T}} p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \mathrm{d}\boldsymbol{\tau}}. \tag{33}$$

□

Proposition 3.1. *Let $p_{*,\boldsymbol{\theta}}$ as defined in Equation (9). Then, the solution to the problem in Equation (13) can be obtained via the weighted cross-entropy minimization:*

$$\boldsymbol{\theta}_{\dagger}^b \in \operatorname*{arg\,min}_{\boldsymbol{\theta}^b \in \boldsymbol{\Theta}} \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}} \left[-\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\| \log p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau}) \right] = \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}} \left[-\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\| \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}^b}(a_t|s_t) \right]. \tag{14}$$

Proof. We simply exploit the form of the optimal behavioral distribution p_* and the definition of KL divergence:

$$\begin{aligned}
\operatorname*{arg\,min}_{\boldsymbol{\theta}^b \in \boldsymbol{\Theta}} D_{\mathrm{KL}}\left(p_{*,\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}^b}\right) &= \operatorname*{arg\,min}_{\boldsymbol{\theta}^b \in \boldsymbol{\Theta}} \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{*,\boldsymbol{\theta}}} \left[\log \left(\frac{p_{*,\boldsymbol{\theta}}(\boldsymbol{\tau})}{p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau})} \right) \right] && (34)\\
&= \operatorname*{arg\,min}_{\boldsymbol{\theta}^b \in \boldsymbol{\Theta}} - \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{*,\boldsymbol{\theta}}} \left[\log p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau}) \right] && (35)\\
&= \operatorname*{arg\,min}_{\boldsymbol{\theta}^b \in \boldsymbol{\Theta}} - \int_{\mathcal{T}} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}) \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2}{p_{\boldsymbol{\theta}}(\boldsymbol{\tau}') \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau}')\|_2 \mathrm{d}\boldsymbol{\tau}'} \log p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} && (36)\\
&= \operatorname*{arg\,min}_{\boldsymbol{\theta}^b \in \boldsymbol{\Theta}} - \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}} \left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\| \log p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau}) \right], && (37)
\end{aligned}$$

which proves the first equality. For the second equality, we observe that:

$$\log p_{\boldsymbol{\theta}^b}(\boldsymbol{\tau}) = \log \mu_0(s_0) + \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}^b}(a_t|s_t) + \sum_{t=0}^{T-1} \log P(s_{t+1}|s_t, a_t), \tag{38}$$

and that the addenda of the initial-state distribution and of the transition model do not depend on $\boldsymbol{\theta}^b$. □

B.2 Proofs of Section 4.1

Lemma 4.1. *Fix a target policy $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a behavior trajectory distribution $q \in \Delta^{\mathcal{T}}$. Let $\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \boldsymbol{\tau})$ be the importance-weighted estimate of $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ computed with $\boldsymbol{\tau} \sim q$. Then the variance reduction from using q in place of $p_{\boldsymbol{\theta}}$ is given by:*

$$\mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}} \left[\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \boldsymbol{\tau}) \right] - \mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau} \sim q} \left[\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \boldsymbol{\tau}) \right] = \mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}} \left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \right] - Z_{\boldsymbol{\theta}}^2 \chi^2(p_{*,\boldsymbol{\theta}} \| q),$$

where $Z_{\boldsymbol{\theta}} := \mathbb{E}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}}[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2]$.

Proof. Let p_* be short for $p_{*,\boldsymbol{\theta}}$. First, we know from Theorem 1 that the variance reduction granted by the optimal behavior distribution w.r.t. on-policy estimation is

$$\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] - \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] = \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}[\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2] - \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{E}}[\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2]^2 = \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}[\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2].$$

Let $\boldsymbol{v} =$, so the variance reduction granted by sampling from q is

$$\begin{aligned}\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] - \underset{\boldsymbol{\tau}\sim q}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] &= \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] - \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] \\ &\quad + \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};p_*;\boldsymbol{\tau})] - \underset{\boldsymbol{\tau}\sim q}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] \qquad (39)\\ &= \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}{\mathbb{V}\mathrm{ar}}[\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2] - \left(\underset{\boldsymbol{\tau}\sim q}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] - \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};p_*;\boldsymbol{\tau})]\right), \qquad (40)\end{aligned}$$

which is the variance reduction granted by p_* minus the excess variance due to using a proxy q of p_*. We can characterize this excess variance as follows. Since both estimates are unbiased:

$$\begin{aligned}\underset{\boldsymbol{\tau}\sim q}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] - \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{V}\mathrm{ar}}[\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})] &= \underset{\boldsymbol{\tau}\sim q}{\mathbb{E}}\left[\left\|\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})\right\|_2^2\right] - \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{E}}\left[\left\|\widehat{\nabla} J(\boldsymbol{\theta};\boldsymbol{\tau})\right\|_2^2\right] \qquad (41)\\ &= \int_{\boldsymbol{\mathcal{T}}} q(\boldsymbol{\tau})\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})^2}{q(\boldsymbol{\tau})^2}\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\,\mathrm{d}\boldsymbol{\tau} - \int_{\boldsymbol{\mathcal{T}}} p_*(\boldsymbol{\tau})\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})^2}{p_*(\boldsymbol{\tau})^2}\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\,\mathrm{d}\boldsymbol{\tau} \qquad (42)\\ &= \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\theta}}(\boldsymbol{\tau})\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{q(\boldsymbol{\tau})}\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\,\mathrm{d}\boldsymbol{\tau} - \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\theta}}(\boldsymbol{\tau})\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{p_*(\boldsymbol{\tau})}\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2^2\,\mathrm{d}\boldsymbol{\tau} \qquad (43)\\ &= Z_{\boldsymbol{\theta}}\int_{\boldsymbol{\mathcal{T}}} p_*(\boldsymbol{\tau})\frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{q(\boldsymbol{\tau})}\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\,\mathrm{d}\boldsymbol{\tau} - Z_{\boldsymbol{\theta}}\int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\,\mathrm{d}\boldsymbol{\tau} \qquad (44)\\ &= Z_{\boldsymbol{\theta}}\int_{\boldsymbol{\mathcal{T}}} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{\tau})}{q(\boldsymbol{\tau})}\|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\,(p_*(\boldsymbol{\tau}) - q(\boldsymbol{\tau}))\,\mathrm{d}\boldsymbol{\tau} \qquad (45)\\ &= Z_{\boldsymbol{\theta}}^2\int_{\boldsymbol{\mathcal{T}}} \frac{p_*(\boldsymbol{\tau})}{q(\boldsymbol{\tau})}(p_*(\boldsymbol{\tau}) - q(\boldsymbol{\tau}))\,\mathrm{d}\boldsymbol{\tau} \qquad (46)\\ &= Z_{\boldsymbol{\theta}}^2\left(\int_{\boldsymbol{\mathcal{T}}} \frac{p_*(\boldsymbol{\tau})^2}{q(\boldsymbol{\tau})}\mathrm{d}\boldsymbol{\tau} - 1\right) \qquad (47)\\ &= Z_{\boldsymbol{\theta}}^2\chi^2(p_*\|q), \qquad (48)\end{aligned}$$

where Equation (44) and (46) are by definition of p_*. ☐

Unfortunately, it is not possible to upper bound the chi-square divergence in terms of the KL in general. To obtain an upper bound for the special case of defensive estimators, we will need the following technical lemma, a generalization of Lemma 8 by Bubeck & Sellke (2020).

Lemma B.1. *For any $\eta > 0$,*

$$\int_{\boldsymbol{\mathcal{T}}} \frac{(q(\boldsymbol{\tau}) - p(\boldsymbol{\tau}))^2}{q(\boldsymbol{\tau})}\mathbf{1}_{\{q(\boldsymbol{\tau})\geqslant \eta p(\boldsymbol{\tau})\}}\mathrm{d}\boldsymbol{\tau} \leqslant 4\eta^{-3/2}D_H^2(p,q).$$

Proof. Let $f_t(s) = (\sqrt{t} - \sqrt{s})^2$. Its second derivative is $f_t''(s) = \frac{\sqrt{t}}{2s\sqrt{s}}$. We can see that, restricted to $s \leqslant \eta^{-1}t$, f_t is $\frac{\eta^{3/2}}{2t}$-strongly convex. Hence:

$$f_t(s) \geqslant \frac{\eta^{3/2}(t-s)^2}{4t}. \qquad (49)$$

Letting $t = q(\boldsymbol{\tau})$ and $s = p(\boldsymbol{\tau})$ and using the definition of Hellinger distance:

$$D_H^2(p,q) = \int_{\mathcal{T}} \left(\sqrt{q(\boldsymbol{\tau})} - \sqrt{p(\boldsymbol{\tau})}\right)^2 \mathrm{d}\boldsymbol{\tau} \geqslant \int_{\mathcal{T}} \left(\sqrt{q(\boldsymbol{\tau})} - \sqrt{p(\boldsymbol{\tau})}\right)^2 \mathbf{1}_{\{q(\boldsymbol{\tau}) \geqslant \eta p(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau} \tag{50}$$

$$\geqslant \frac{\eta^{3/2}}{4} \int_{\mathcal{T}} \frac{(q(\boldsymbol{\tau}) - p(\boldsymbol{\tau}))^2}{q(\boldsymbol{\tau})} \mathbf{1}_{\{q(\boldsymbol{\tau}) \geqslant \eta p(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau}. \tag{51}$$

□

We are now ready to prove Theorem 2.

Theorem 2. *Fix a target policy $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a behavior policy $\widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$. Let $\beta \in [0,1]$ and let $\Phi = \beta p_{\boldsymbol{\theta}} + (1-\beta) p_{\widetilde{\boldsymbol{\theta}}}$ be the mixture trajectory distribution. Let $\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \boldsymbol{\tau})$ be the β-defensive importance-weighted estimate of $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ computed with $\boldsymbol{\tau} \sim \Phi$. Then the variance reduction from using Φ in place of $p_{\boldsymbol{\theta}}$ is at least*

$$\mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}} \left[\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})\right] - \mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau}\sim \Phi} \left[\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \boldsymbol{\tau})\right] \geqslant \mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}} \left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right] - 4 Z_{\boldsymbol{\theta}} (Z_{\boldsymbol{\theta}} + \beta G_{\boldsymbol{\theta}}) \left(2 + \frac{1-\beta}{\beta} D_{\mathrm{KL}}(p_{*,\boldsymbol{\theta}} \| p_{\widetilde{\boldsymbol{\theta}}})\right),$$

where $Z_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}}[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2]$ and $G_{\boldsymbol{\theta}} = \operatorname{ess\,sup}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}} \{\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\}$. Under Assumption 1, provided $\epsilon_{\mathrm{KL}} \leqslant 1$, by setting $\beta = \sqrt{\frac{\epsilon_{\mathrm{KL}}}{2-\epsilon_{\mathrm{KL}}}}$, the variance reduction is at least

$$\begin{aligned}\mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}} \left[\widehat{\nabla} J(\boldsymbol{\theta}; \boldsymbol{\tau})\right] - \mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau}\sim \Phi} \left[\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \boldsymbol{\tau})\right] &\geqslant \mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}} \left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right] - 4 Z_{\boldsymbol{\theta}}^2 (2-\epsilon_{\mathrm{KL}}) - 4 Z_{\boldsymbol{\theta}} G_{\boldsymbol{\theta}} \epsilon_{\mathrm{KL}} \\ &\quad - 4 Z_{\boldsymbol{\theta}} (Z_{\boldsymbol{\theta}} + G_{\boldsymbol{\theta}}) \sqrt{\epsilon_{\mathrm{KL}}(2-\epsilon_{\mathrm{KL}})} \qquad (16) \\ &\geqslant \mathop{\mathbb{V}\mathrm{ar}}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}} \left[\|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2\right] - 8 Z_{\boldsymbol{\theta}}^2 - 4 Z_{\boldsymbol{\theta}} (Z_{\boldsymbol{\theta}} + 2 G_{\boldsymbol{\theta}}) \sqrt{\epsilon_{\mathrm{KL}}}. \qquad (17)\end{aligned}$$

Proof. To prove the first lower bound on variance reduction, we use Lemma 4.1 with ϕ (density of Φ) in place of q and upper bound the negative term as follows, applying Lemma B.1 twice:

$$Z_{\boldsymbol{\theta}}^2 \chi^2(p_* \| \Phi) = Z_{\boldsymbol{\theta}}^2 \int_{\mathcal{T}} \frac{(\phi(\boldsymbol{\tau}) - p_*(\boldsymbol{\tau}))^2}{\phi(\boldsymbol{\tau})} \mathbf{1}_{\{\phi(\boldsymbol{\tau}) \geqslant \beta^{2/3} p_*(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau} + Z_{\boldsymbol{\theta}}^2 \int_{\mathcal{T}} \frac{(\phi(\boldsymbol{\tau}) - p_*(\boldsymbol{\tau}))^2}{\phi(\boldsymbol{\tau})} \mathbf{1}_{\{\phi(\boldsymbol{\tau}) \leqslant \beta^{2/3} p_*(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau} \tag{52}$$

$$\begin{aligned}&\leqslant Z_{\boldsymbol{\theta}}^2 \frac{4}{\beta} D_H^2(p_*, \phi) + Z_{\boldsymbol{\theta}}^2 \int_{\mathcal{T}} \frac{(\phi(\boldsymbol{\tau}) - p_*(\boldsymbol{\tau}))^2}{\phi(\boldsymbol{\tau})} \mathbf{1}_{\{\phi(\boldsymbol{\tau}) \leqslant \beta^{2/3} p_*(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau} && (53) \\ &\leqslant \frac{4 Z_{\boldsymbol{\theta}}^2}{\beta} D_H^2(p_*, \phi) + \frac{Z_{\boldsymbol{\theta}}^2}{\beta} \int_{\mathcal{T}} \frac{(\phi(\boldsymbol{\tau}) - p_*(\boldsymbol{\tau}))^2}{p(\boldsymbol{\tau})} \mathbf{1}_{\{\phi(\boldsymbol{\tau}) \leqslant \beta^{2/3} p_*(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau} && (54) \\ &= \frac{4 Z_{\boldsymbol{\theta}}^2}{\beta} D_H^2(p_*, \phi) + \frac{Z_{\boldsymbol{\theta}}}{\beta} \int_{\mathcal{T}} \|g_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \frac{(\phi(\boldsymbol{\tau}) - p_*(\boldsymbol{\tau}))^2}{p_*(\boldsymbol{\tau})} \mathbf{1}_{\{p_*(\boldsymbol{\tau}) \geqslant \beta^{-2/3} \phi(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau} && (55) \\ &\leqslant \frac{4 Z_{\boldsymbol{\theta}}^2}{\beta} D_H^2(p_*, \phi) + \frac{Z_{\boldsymbol{\theta}} G_{\boldsymbol{\theta}}}{\beta} \int_{\mathcal{T}} \frac{(\phi(\boldsymbol{\tau}) - p_*(\boldsymbol{\tau}))^2}{p_*(\boldsymbol{\tau})} \mathbf{1}_{\{p_*(\boldsymbol{\tau}) \geqslant \beta^{-2/3} \phi(\boldsymbol{\tau})\}} \mathrm{d}\boldsymbol{\tau} && (56) \\ &\leqslant \frac{4 Z_{\boldsymbol{\theta}}^2}{\beta} D_H^2(p_*, \phi) + 4 Z_{\boldsymbol{\theta}} G_{\boldsymbol{\theta}} D_H^2(\phi, p_*) && (57) \\ &\leqslant 4 Z_{\boldsymbol{\theta}} \frac{Z_{\boldsymbol{\theta}} + \beta G_{\boldsymbol{\theta}}}{\beta} \left(\beta D_H^2(p_*, p) + (1-\beta) D_H^2(p_*, p_{\widetilde{\boldsymbol{\theta}}})\right) && (58) \\ &\leqslant 4 Z_{\boldsymbol{\theta}} \frac{Z_{\boldsymbol{\theta}} + \beta G_{\boldsymbol{\theta}}}{\beta} \left(2\beta + (1-\beta) D_{\mathrm{KL}}(p_* \| p_{\widetilde{\boldsymbol{\theta}}})\right) && (59) \\ &= 4 Z_{\boldsymbol{\theta}} (Z_{\boldsymbol{\theta}} + \beta G) \left(2 + \frac{1-\beta}{\beta} D_{\mathrm{KL}}(p_* \| p_{\widetilde{\boldsymbol{\theta}}})\right), && (60)\end{aligned}$$

where the inequalities (53) and (57) are by Lemma B.1. The latter expression is convex in β, but the optimal value $\beta^* = \sqrt{\frac{Z_{\boldsymbol{\theta}} \epsilon_{\mathrm{KL}}}{(2-\epsilon_{\mathrm{KL}}) G_{\boldsymbol{\theta}}}}$ cannot be computed since $Z_{\boldsymbol{\theta}}$ is unknown. However, upperbounding $Z_{\boldsymbol{\theta}}$ by $G_{\boldsymbol{\theta}}$ and setting[6] $\beta = \sqrt{\frac{\epsilon_{\text{KL}}}{2-\epsilon_{\text{KL}}}}$ yields, provided $\epsilon_{\text{KL}} \leqslant 1$:

$$Z_{\boldsymbol{\theta}}^2 \chi^2(p_* \| \Phi) \leqslant 4Z_{\boldsymbol{\theta}}^2(2-\epsilon_{\text{KL}}) + 4Z_{\boldsymbol{\theta}} G_{\boldsymbol{\theta}} \epsilon_{\text{KL}} + 4Z_{\boldsymbol{\theta}}(Z_{\boldsymbol{\theta}} + G_{\boldsymbol{\theta}})\sqrt{\epsilon_{\text{KL}}(2-\epsilon_{\text{KL}})}, \tag{61}$$

proving the second bound. The third and final bound follows from the fact that $\epsilon \leqslant \sqrt{\epsilon}$ for $\epsilon \leqslant 1$. □

Theorem 3. *Fix an iteration $k \in [K]$ of Algorithm 1 and let $\mathcal{D}_{\text{ON}}$ denote a dataset of N_{PG} independent trajectories collected with $\boldsymbol{\theta}_k$. Under Assumption 1, the variance reduction granted by using the off-policy estimator $\mathbf{v}_k := \widehat{\nabla} J(\boldsymbol{\theta}_k; \mathcal{D}_k^{\text{PG}})$ with respect to an on-policy estimator is given by:*

$$\underset{k}{\mathbb{V}\text{ar}}\left[\widehat{\nabla} J(\boldsymbol{\theta}_k; \mathcal{D}_{\text{ON}})\right] - \underset{k}{\mathbb{V}\text{ar}}[\mathbf{v}_k] \geqslant \frac{1}{N_{\text{PG}}}\left(V_k - 8Z_k^2 - 4Z_k(Z_k + 2G_k)\sqrt{\epsilon_{\text{KL}}}\right), \tag{19}$$

where $Z_k := \mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}_k}}[\|\mathbf{g}_{\boldsymbol{\theta}_k}(\boldsymbol{\tau})\|_2 | \mathcal{F}_k]$, $V_k := \mathbb{V}\text{ar}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}_k}}[\|\mathbf{g}_{\boldsymbol{\theta}_k}(\boldsymbol{\tau})\|_2 | \mathcal{F}_k]$, *and* $G_k := \operatorname{ess\,sup}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}_k}}\{\|\mathbf{g}_{\boldsymbol{\theta}_k}(\boldsymbol{\tau})\|_2\}$. *Thus, the conditional variance of $\mathbf{v}_k$ is upper-bounded as follows:*

$$\underset{k}{\mathbb{V}\text{ar}}[\mathbf{v}_k] \leqslant \frac{1}{N_{\text{PG}}}\left(9Z_k^2 + Z_k(Z_k + 2G_k)\sqrt{\epsilon_{\text{KL}}} - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right). \tag{20}$$

Proof. Assumption 1 allows Algorithm 1 to query the BPO oracle at Line 4, obtaining $\widetilde{\boldsymbol{\theta}}_k = \text{BPO}(\boldsymbol{\theta}_k)$ with $D_{\text{KL}}(p_{*,\boldsymbol{\theta}_k} \| p_{\widetilde{\boldsymbol{\theta}}_k}) \leqslant \epsilon_{\text{KL}}$. So, the first statement follows immediately from Theorem 2 and the properties of variance (just notice that $\mathbf{v}_k$ can also be written as the average of N_{PG} independent random variables). Then, the second statement follows by rearranging the terms and noting that:

$$N_{\text{PG}} \underset{k}{\mathbb{V}\text{ar}}\left[\widehat{\nabla} J(\boldsymbol{\theta}_k; \mathcal{D}_{\text{ON}})\right] - V_k = Z_k^2 - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2. \tag{62}$$

□

B.3 Proofs of Section 4.2

For the scope of this section, fix a target policy $\boldsymbol{\theta}^\dagger$, let p_* be the corresponding optimal behavior policy $p^*_{\boldsymbol{\theta}^\dagger}$, and let $F(\boldsymbol{\tau}) = \|g_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau})\|_2$ for brevity. Let $\widehat{L} : \boldsymbol{\Theta} \to \mathbb{R}_+$ be the empirical loss defined as:

$$\widehat{L}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{\tau}_i) \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}}(a_t^i | s_t^i), \tag{63}$$

where $\boldsymbol{\tau}_i = (s_0^i, a_0^i, \ldots, s_{T-1}^i, a_{T-1}^i)$, so that $\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \widehat{L}(\boldsymbol{\theta})$. Also, let

$$L(\boldsymbol{\theta}) = \mathbb{E}\left[\widehat{L}(\boldsymbol{\theta})\right] = -\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^\dagger}}{\mathbb{E}}\left[F(\boldsymbol{\tau})\sum_{t=0}^{T-1}\log \pi_{\boldsymbol{\theta}}(a_t|s_t)\right], \tag{64}$$

where $\boldsymbol{\tau} = (s_0, a_0, \ldots, a_{T-1}, s_{T-1})$, and $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$.

Lemma B.2. *Under Assumptions 2 and 4:*

$$\nabla L(\boldsymbol{\theta}) = -\underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^\dagger}}{\mathbb{E}}\left[F(\boldsymbol{\tau})\sum_{t=0}^{T-1}\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s_t, a_t)\right], \tag{65}$$

$$\nabla^2 L(\boldsymbol{\theta}) = \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^\dagger}}{\mathbb{E}}\left[F(\boldsymbol{\tau})\sum_{t=0}^{T-1}\underset{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t)}{\mathbb{C}\text{ov}}[\boldsymbol{\varphi}(s_t, a)]\right], \tag{66}$$

$$\left\|\nabla^2 L(\boldsymbol{\theta})\right\|_2 \leqslant GT\sigma^2. \tag{67}$$

[6] Note that we do not actually need to know $G_{\boldsymbol{\theta}}$, nor an upper bound.

Proof. The first two statements follow immediately from Assumption 2. As for the third statement:

$$\left\|\nabla^2 L(\boldsymbol{\theta})\right\|_2 \leqslant \mathbb{E}\left[G \sum_{t=0}^{T-1} \left\| \underset{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t)}{\mathbb{E}}\left[\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s_t,a)\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s_t,a)^\top\right]\right\|_2\right] \tag{68}$$

$$\leqslant \mathbb{E}\left[G \sum_{t=0}^{T-1} \underset{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t)}{\mathbb{E}}\left[\|\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s_t,a)\|_2^2\right]\right] \leqslant GT\sigma^2, \tag{69}$$

where the last inequality is by Assumption 4 and Proposition 1. □

Lemma B.3. *Under Assumptions 2, 3 and 4,*

$$\mathbb{E}\left[\left\|\nabla \widehat{L}(\boldsymbol{\theta}^*)\right\|_2^2\right] \leqslant \frac{Z_{\boldsymbol{\theta}^\dagger} G T^2 \sigma^2}{n}. \tag{70}$$

Proof. First notice that, for policies of the exponential family (Assumption 2):

$$\mathbb{E}\left[\nabla \widehat{L}(\boldsymbol{\theta}^*)\right] = \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^\dagger}}{\mathbb{E}}\left[\|g_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau})\|_2 \sum_{t=0}^{T-1} \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t,a_t)\right] \tag{71}$$

$$= Z_{\boldsymbol{\theta}^\dagger} \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{E}}\left[\sum_{t=0}^{T-1} \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t,a_t)\right] \tag{72}$$

$$= Z_{\boldsymbol{\theta}^\dagger} \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^*}}{\mathbb{E}}\left[\sum_{t=0}^{T-1} \underset{a\sim\pi_{\boldsymbol{\theta}^*}(\cdot|s_t)}{\mathbb{E}}\left[\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t,a)|s_t\right]\right] \tag{73}$$

$$= 0, \tag{74}$$

where the second-to-last equality is by Assumption 3. Then

$$\mathbb{E}\left[\left\|\nabla \widehat{L}(\boldsymbol{\theta}^*)\right\|_2^2\right] = \mathbb{V}\mathrm{ar}\left[\nabla \widehat{L}(\boldsymbol{\theta}^*)\right] \tag{75}$$

$$= \frac{1}{n} \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^\dagger}}{\mathbb{V}\mathrm{ar}}\left[\|g_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau})\|_2 \sum_{t=0}^{T-1} \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t,a_t)\right] \tag{76}$$

$$= \frac{1}{n} \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^\dagger}}{\mathbb{E}}\left[\|g_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau})\|_2^2 \left\|\sum_{t=0}^{T-1} \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t,a_t)\right\|_2^2\right] \tag{77}$$

$$= \frac{Z_{\boldsymbol{\theta}^\dagger}}{n} \underset{\boldsymbol{\tau}\sim p_*}{\mathbb{E}}\left[\|g_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau})\|_2 \left\|\sum_{t=0}^{T-1} \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t,a_t)\right\|_2^2\right] \tag{78}$$

$$\leqslant \frac{Z_{\boldsymbol{\theta}^\dagger} G T}{n} \underset{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}^*}}{\mathbb{E}}\left[\sum_{t=0}^{T-1} \underset{a\sim\pi_{\boldsymbol{\theta}^*}(\cdot|s_t)}{\mathbb{E}}\left[\|\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t,a)\|_2^2 \middle| s_t\right]\right] \tag{79}$$

$$\leqslant \frac{Z_{\boldsymbol{\theta}^\dagger} G T^2 \sigma^2}{n}, \tag{80}$$

where the last inequality is by Assumption 4 and the second-to-last relies on Assumption 3. □

Lemma 4.2. *Fix a target policy parameter $\boldsymbol{\theta}^\dagger \in \boldsymbol{\Theta}$ and let $\{\boldsymbol{\tau}_i\}_{i\in[n]}$ be a dataset of n i.i.d. trajectories collected with $\pi_{\boldsymbol{\theta}^\dagger}$. Let*

$$\widetilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \sum_{i=1}^{n} \|\mathbf{g}_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau}_i)\|_2 \sum_{t=0}^{T-1} \log \pi_{\boldsymbol{\theta}}(a_t^i|s_t^i),$$

and if $\operatorname{ess\,sup}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\theta}}} \|\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\tau})\|_2 \leqslant G$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Then, under Assumptions 2, 3, 4, 5 it holds that:

$$\mathbb{E}\left[D_{\mathrm{KL}}(p_{*,\boldsymbol{\theta}^\dagger} \| p_{\widetilde{\boldsymbol{\theta}}})\right] \leqslant \frac{G^2 T^3 \sigma^4}{2\lambda_*^2 n}.$$

Proof. By the mean value theorem, there exists a $c \in [0,1]$ such that

$$L(\widetilde{\boldsymbol{\theta}}) = L(\boldsymbol{\theta}^*) + \langle \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \nabla L(\boldsymbol{\theta}^*) \rangle + \frac{1}{2}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \nabla^2 L(\boldsymbol{\theta}_c)(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \tag{81}$$

$$\leqslant L(\boldsymbol{\theta}^*) + \frac{1}{2} G T \sigma^2 \left\| \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2^2, \tag{82}$$

where $\boldsymbol{\theta}_c = c\widetilde{\boldsymbol{\theta}} + (1-c)\boldsymbol{\theta}^*$ for some $c \in [0,1]$ and the last inequality is by Lemma B.2 under Assumptions 2 and 4.

Now let

$$\widehat{\mathcal{G}}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} F(\boldsymbol{\tau}_i) \sum_{t=0}^{T-1} \left(\overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s_t^i, a_t^i) - \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}^*}(s_t^i, a_t^i) \right), \tag{83}$$

and notice that $\widehat{\mathcal{G}}(\boldsymbol{\theta}^*) = 0$, and that $\nabla \widehat{\mathcal{G}}(\boldsymbol{\theta}) = \widehat{\mathcal{F}}(\boldsymbol{\theta})$ where $\widehat{\mathcal{F}}$ is defined in Assumption 5. Then, from the mean value theorem, there exists a $c \in [0,1]$ such that:

$$\widehat{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) = \widehat{\mathcal{G}}(\boldsymbol{\theta}^*) + (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \nabla \widehat{\mathcal{G}}(\boldsymbol{\theta}_c) = (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \widehat{F}(\boldsymbol{\theta}_c), \tag{84}$$

where $\boldsymbol{\theta}_c = c\widetilde{\boldsymbol{\theta}} + (1-c)\boldsymbol{\theta}^*$. Hence, by Assumption 5,

$$\mathbb{E}\left[\left\| \widehat{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) \right\|_2^2 \right] \geqslant \lambda_*^2 \, \mathbb{E}\left[\left\| \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2^2 \right]. \tag{85}$$

Next, notice that $\widehat{\mathcal{G}}(\boldsymbol{\theta}) = \nabla \widehat{L}(\boldsymbol{\theta}) - \nabla \widehat{L}(\boldsymbol{\theta}^*)$ by Assumption 2. Thus, by definition of $\widetilde{\boldsymbol{\theta}}$, $\widehat{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) = \nabla \widehat{L}(\widetilde{\boldsymbol{\theta}}) - \nabla \widehat{L}(\boldsymbol{\theta}^*) = \nabla \widehat{L}(\boldsymbol{\theta}^*)$, and

$$\mathbb{E}\left[\left\| \widehat{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) \right\|_2^2 \right] = \mathbb{E}\left[\left\| \nabla \widehat{L}(\boldsymbol{\theta}^*) \right\|_2^2 \right] \leqslant \frac{Z_{\boldsymbol{\theta}^\dagger} G T^2 \sigma^2}{n}. \tag{86}$$

where the last inequality is by Lemma B.3 under Assumptions 2, 3 and 4.

Finally, chaining the inequalities from Equations (82), (85), and (86):

$$\mathbb{E}[L(\widetilde{\boldsymbol{\theta}})] \leqslant L(\boldsymbol{\theta}^*) + \frac{1}{2} G T \sigma^2 \, \mathbb{E}\left[\left\| \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2^2 \right] \tag{87}$$

$$\leqslant L(\boldsymbol{\theta}^*) + \frac{G T \sigma^2}{2\lambda_*^2} \, \mathbb{E}\left[\left\| \widehat{\mathcal{G}}(\widetilde{\boldsymbol{\theta}}) \right\|_2^2 \right] \tag{88}$$

$$\leqslant L(\boldsymbol{\theta}^*) + \frac{Z_{\boldsymbol{\theta}^\dagger} G^2 T^3 \sigma^4}{2\lambda_*^2 n}. \tag{89}$$

Finally:

$$D_{\mathrm{KL}}(p_* \| p_{\widetilde{\boldsymbol{\theta}}}) = \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_*} \left[\log p_*(\boldsymbol{\tau}) - \log p_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\tau}) \right] \tag{90}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^*}} \left[\log p_{\boldsymbol{\theta}^*}(\boldsymbol{\tau}) - \log p_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\tau}) \right] \tag{91}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^*}} \left[\sum_{t=0}^{T-1} \left(\log \pi_{\boldsymbol{\theta}^*}(a_t|s_t) - \log \pi_{\widetilde{\boldsymbol{\theta}}}(a_t|s_t) \right) \right] \tag{92}$$

$$= \frac{1}{Z_{\boldsymbol{\theta}^\dagger}} \mathop{\mathbb{E}}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}^\dagger}} \left[\| g_{\boldsymbol{\theta}^\dagger}(\boldsymbol{\tau}) \|_2 \sum_{t=0}^{T-1} \left(\log \pi_{\boldsymbol{\theta}^*}(a_t|s_t) - \log \pi_{\widetilde{\boldsymbol{\theta}}}(a_t|s_t) \right) \right] \tag{93}$$

$$= \frac{L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)}{Z_{\boldsymbol{\theta}^\dagger}}, \tag{94}$$

and by Equation (89):

$$\mathbb{E}[D_{\mathrm{KL}}(p_* \| p_{\widetilde{\boldsymbol{\theta}}})] = \frac{\mathbb{E}[L(\widetilde{\boldsymbol{\theta}})] - L(\boldsymbol{\theta}^*)}{Z_{\boldsymbol{\theta}^\dagger}} \leqslant \frac{G^2 T^3 \sigma^4}{2\lambda_*^2 n}. \tag{95}$$

□

Lemma B.4. *Under Assumptions 2 and 4, for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$:*

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geqslant \langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \nabla J(\boldsymbol{\theta}) \rangle - \frac{R_{\max}\sigma^2}{(1-\gamma)^2} \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} \right\|_2^2.$$

Proof. Under Assumption 2,

$$\mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\| \nabla \log \pi_{\boldsymbol{\theta}}(a|s) \|_2^2] = \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\| \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s,a) \|_2^2] \leqslant \sigma^2,$$

where the last inequality is by sub-Gaussianity of the centered sufficient statistic (Assumption 4 and Proposition 1). Similarly:

$$\begin{aligned}
\mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\left\| \nabla^2 \log \pi_{\boldsymbol{\theta}} \right\|_2] &= \left\| \mathop{\mathbb{C}\mathrm{ov}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{\varphi}(s,a)] \right\|_2 \\
&\leqslant \operatorname{trace} \left(\mathop{\mathbb{C}\mathrm{ov}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{\varphi}(s,a)] \right) \\
&= \mathop{\mathbb{V}\mathrm{ar}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{\varphi}(s,a)] \\
&= \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\| \overline{\boldsymbol{\varphi}}_{\boldsymbol{\theta}}(s,a) \|_2^2] \leqslant \sigma^2.
\end{aligned}$$

Hence, by Proposition 2, $\left\| \nabla^2 J(\boldsymbol{\theta}) \right\|_2 \leqslant \frac{2R_{\max}\sigma^2}{(1-\gamma)^2}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Finally, by the mean value theorem, there exists $c \in (0,1)$ such that:

$$\begin{aligned}
J(\boldsymbol{\theta}') &= J(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \nabla J(\boldsymbol{\theta}) \rangle + \frac{1}{2} (\boldsymbol{\theta}' - \boldsymbol{\theta})^\top \nabla^2 J(\boldsymbol{\theta}_c)(\boldsymbol{\theta}' - \boldsymbol{\theta}) \\
&\geqslant J(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \nabla J(\boldsymbol{\theta}) \rangle - \frac{1}{2} \left\| \nabla^2 J(\boldsymbol{\theta}) \right\|_2 \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} \right\|_2^2 \\
&\geqslant J(\boldsymbol{\theta}) + \langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \nabla J(\boldsymbol{\theta}) \rangle - \frac{R_{\max}\sigma^2}{(1-\gamma)^2} \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} \right\|_2^2,
\end{aligned}$$

where $\boldsymbol{\theta}_c = c\boldsymbol{\theta} + (1-c)\boldsymbol{\theta}'$ for some $c \in [0,1]$. □

Theorem 4. *Assuming $N_{\mathrm{BPO}} > \frac{G^2T^3\sigma^4}{2\lambda_*^2}$, let $\epsilon^* = \frac{G^2T^3\sigma^4}{2\lambda_*^2 N_{\mathrm{BPO}}}$. Then, under Assumptions 2, 3, 4, 5, Algorithm 1 with $\beta = \sqrt{\epsilon^*/(2-\epsilon^*)}$ guarantees*

$$\mathop{\mathbb{V}\mathrm{ar}}_k[\mathbf{v}_k] \leqslant \frac{1}{N_{\mathrm{PG}}} \left(9Z_k^2 + \frac{Z_k(Z_k + 2G)GT^{3/2}\sigma^2}{\lambda_*\sqrt{2N_{\mathrm{BPO}}}} - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2 \right). \tag{22}$$

Furthermore, by setting $N_{\mathrm{BPO}} = N_{\mathrm{PG}} = \frac{N}{2}$ and $\beta \in (0,1)$, provided $N > \frac{G^2T^3\sigma^4(1+\beta^2)}{2\lambda_*^2\beta^2}$ we have:

$$\mathop{\mathbb{V}\mathrm{ar}}_k[\mathbf{v}_k] \leqslant \frac{1}{N} \left(18Z_k^2 - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2 \right) + \frac{Z_k(Z_k + 2G)GT^{3/2}\sigma^2}{2\lambda_* N^{3/2}}. \tag{23}$$

Proof. The first statement follows from Theorem 3 and Lemma 4.2. For the second statement, notice that for every $\beta \in (0,1)$ there is an $\epsilon \in (0,1)$ such that $\beta = \sqrt{\epsilon/(2-\epsilon)}$. The assumption on the batch size N guarantees that ϵ is a valid upper bound on the KL divergence. □

Corollary 1. *Let $\widetilde{V} := 18Z_k^2 - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2$ denote the residual variance left by the BPO process. Under the assumptions of Theorem 4, a total number of trajectories*

$$NK \leqslant \left\lceil 12(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0)) \left(\frac{3C_1\widetilde{V}}{\epsilon^4} + \frac{C_1 + 3C_2}{\epsilon^{10/3}} \right) \right\rceil$$

is sufficient for Algorithm 1 to obtain $\mathbb{E}[\|\nabla J(\boldsymbol{\theta}_{out})\|_2] \leqslant \epsilon$*, where* $\boldsymbol{\theta}_{out}$ *is chosen uniformly at random from the iterates* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ *of the algorithm, where* $C_1 = \frac{R_{\max}\sigma^2}{(1-\gamma)^2}$ *and* $C_2 = \frac{R_{\max}^4 \sigma^5 \|\boldsymbol{\varphi}\|_\infty (\sqrt{T}\sigma + 2T\|\boldsymbol{\varphi}\|_\infty)T^3}{2\lambda_*(1-\gamma)^5}$.

Proof. By Lemma B.4:

$$\begin{aligned}
\mathop{\mathbb{E}}_k[J(\boldsymbol{\theta}_{k+1} - J(\boldsymbol{\theta}_k))] &\geqslant \mathop{\mathbb{E}}_k\left[\langle \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k, \nabla J(\boldsymbol{\theta}_k)\rangle - \frac{R_{\max}\sigma^2}{(1-\gamma)^2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2\right] && (96)\\
&= \mathop{\mathbb{E}}_k\left[\alpha\langle \boldsymbol{v}_k, \nabla J(\boldsymbol{\theta}_k)\rangle - \frac{\alpha^2 R_{\max}\sigma^2}{(1-\gamma)^2}\|\boldsymbol{v}_k\|_2^2\right] && (97)\\
&= \alpha\|\nabla J(\boldsymbol{\theta}_k)\|_2^2 - \frac{\alpha^2 R_{\max}\sigma^2}{(1-\gamma)^2}E_k[\|\boldsymbol{v}_k\|_2^2] && (98)\\
&= \alpha\left(1 - \frac{\alpha R_{\max}\sigma^2}{(1-\gamma)^2}\right)\|\nabla J(\boldsymbol{\theta}_k)\|_2^2 - \frac{\alpha^2 R_{\max}\sigma^2}{(1-\gamma^2)}\mathop{\mathbb{V}\mathrm{ar}}_k[\boldsymbol{v}_k] && (99)\\
&\geqslant \alpha\left(1 - \frac{\alpha R_{\max}\sigma^2}{(1-\gamma)^2}\right)\|\nabla J(\boldsymbol{\theta}_k)\|_2^2 - \frac{\alpha^2 R_{\max}\sigma^2(18Z_k^2 - \|\nabla J(\boldsymbol{\theta}_k)\|_2^2)}{(1-\gamma)^2 N} && (100)\\
&\quad - \frac{\alpha^2 R_{\max}\sigma^4 Z_k(Z_k + 2G)GT^{3/2}}{2\lambda_*(1-\gamma)^2 N^{3/2}} && (101)\\
&\geqslant \alpha\left(1 - \frac{\alpha R_{\max}\sigma^2}{(1-\gamma)^2}\right)\|\nabla J(\boldsymbol{\theta}_k)\|_2^2 - \frac{\alpha^2 R_{\max}\sigma^2\widetilde{V}}{(1-\gamma)^2 N} && (102)\\
&\quad - \frac{\alpha^2 R_{\max}\sigma^4 Z_k(Z_k + 2G)GT^{3/2}}{2\lambda_*(1-\gamma)^2 N^{3/2}} && (103)\\
&\geqslant \alpha\left(1 - \frac{\alpha R_{\max}\sigma^2}{(1-\gamma)^2}\right)\|\nabla J(\boldsymbol{\theta}_k)\|_2^2 - \frac{\alpha^2 R_{\max}\sigma^2\widetilde{V}}{(1-\gamma)^2 N} && (104)\\
&\quad - \frac{\alpha^2 R_{\max}\sigma^4 Z_k(Z_k + 2G)GT^{3/2}}{2\lambda_*(1-\gamma)^2 N^{3/2}} && (105)\\
&\geqslant \alpha\left(1 - \frac{\alpha R_{\max}\sigma^2}{(1-\gamma)^2}\right)\|\nabla J(\boldsymbol{\theta}_k)\|_2^2 - \frac{\alpha^2 R_{\max}\sigma^2\widetilde{V}}{(1-\gamma)^2 N} && (106)\\
&\quad - \frac{\alpha^2 R_{\max}^4\sigma^5\|\boldsymbol{\varphi}\|_\infty(\sqrt{T}\sigma + 2T\|\boldsymbol{\varphi}\|_\infty)T^3}{2\lambda_*(1-\gamma)^5 N^{3/2}}. && (107)
\end{aligned}$$

Summing both sides for $k = 0, \ldots, K-1$, by the tower rule of expectation, the sum on the LHS telescopes:

$$\mathbb{E}[J(\boldsymbol{\theta}_K)] - J(\boldsymbol{\theta}_0) \geqslant \alpha(1 - \alpha C_1)\mathbb{E}\left[\sum_{k=0}^{K-1}\|\nabla J(\boldsymbol{\theta}_k)\|_2^2\right] - \frac{K\alpha^2 C_1\widetilde{V}}{N} - \frac{K\alpha^2 C_2}{N^{3/2}}. \quad (108)$$

Rearranging and dividing by K, by definition of $\boldsymbol{\theta}_{\mathrm{OUT}}$, provided $\alpha < 1/C_1$:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_{\mathrm{OUT}})\|_2^2\right] &\leqslant \frac{\mathbb{E}[J(\boldsymbol{\theta}_K)] - J(\boldsymbol{\theta}_0)}{\alpha(1-\alpha C_1)K} + \frac{\alpha C_1\widetilde{V}}{(1-\alpha C_1)N} + \frac{\alpha C_2}{(1-\alpha C_1)N^{3/2}} && (109)\\
&\leqslant \frac{J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0)}{\alpha(1-\alpha C_1)K} + \frac{\alpha C_1\widetilde{V}}{(1-\alpha C_1)N} + \frac{\alpha C_2}{(1-\alpha C_1)N^{3/2}}. && (110)
\end{aligned}$$

Now let $N = \epsilon^{-4/3}$ and $\alpha = \min\left\{\frac{1}{2C_1}, \frac{\epsilon^{2/3}}{6C_1\widetilde{V}}, \frac{1}{6C_2}\right\}$. Then:

$$\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_{\mathrm{OUT}})\|_2^2\right] \leqslant \frac{2(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{\alpha K} + 2\alpha C_1\widetilde{V}\epsilon^{4/3} + 2\alpha C_2\epsilon^2. \quad (111)$$

We consider three cases, and call $\overline{K}$ the smallest integer K such that $\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_{\mathrm{OUT}})\|_2^2\right] \leqslant \epsilon^2$. Note that the latter implies $\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_{\mathrm{OUT}})\|_2\right] \leqslant \epsilon$ by Jensen's inequality.

Case 1. Suppose $\frac{1}{2C_1} \leqslant \min\left\{\frac{\epsilon^{2/3}}{6C_1\widetilde{V}}, \frac{1}{6C_2}\right\}$. Then $\alpha = \frac{1}{2C_1}$ and

$$\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_{\mathrm{OUT}})\|_2^2\right] \leqslant \frac{4C_1(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{K} + \widetilde{V}\epsilon^{4/3} + \frac{C_2\epsilon^2}{C_1} \tag{112}$$

$$\leqslant \frac{4C_1(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{K} + \frac{\epsilon^2}{3} + \frac{\epsilon^2}{3}, \tag{113}$$

so $\overline{K} \leqslant \frac{12C_1(J(\boldsymbol{\theta}^*)-J(\boldsymbol{\theta}_0))}{\epsilon^2}$ in this case.

Case 2. Suppose $\frac{\epsilon^{2/3}}{6C_1\widetilde{V}} \leqslant \min\left\{\frac{1}{2C_1}, \frac{1}{6C_2}\right\}$. Then $\alpha = \frac{\epsilon^{2/3}}{6C_1\widetilde{V}}$ and

$$\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_{\mathrm{OUT}})\|_2^2\right] \leqslant \frac{12C_1\widetilde{V}(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{\epsilon^{2/3}K} + \frac{\epsilon^2}{3} + \frac{C_2\epsilon^{8/3}}{3C_1\widetilde{V}} \tag{114}$$

$$\leqslant \frac{12C_1\widetilde{V}(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{\epsilon^{2/3}K} + \frac{\epsilon^2}{3} + \frac{\epsilon^2}{3}, \tag{115}$$

so $\overline{K} \leqslant \frac{36C_1\widetilde{V}(J(\boldsymbol{\theta}^*)-J(\boldsymbol{\theta}_0))}{\epsilon^{8/3}}$ in this case.

Case 3. Suppose $\frac{1}{6C_2} \leqslant \min\left\{\frac{1}{2C_1}, \frac{\epsilon^{2/3}}{6C_1\widetilde{V}}\right\}$. Then $\alpha = \frac{1}{6C_2}$ and

$$\mathbb{E}\left[\|\nabla J(\boldsymbol{\theta}_{\mathrm{OUT}})\|_2^2\right] \leqslant \frac{12C_2(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{K} + \frac{C_1\widetilde{V}\epsilon^{4/3}}{3C_2} + \frac{\epsilon^2}{3} \tag{116}$$

$$\leqslant \frac{12C_2(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))}{K} + \frac{\epsilon^2}{3} + \frac{\epsilon^2}{3}, \tag{117}$$

so $\overline{K} \leqslant \frac{36C_2(J(\boldsymbol{\theta}^*)-J(\boldsymbol{\theta}_0))}{\epsilon^2}$ in this case.

Considering the three cases, we know for sure that

$$\overline{K} \leqslant 12(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))\left(\frac{3C_1\widetilde{V}}{\epsilon^{8/3}} + \frac{C_1 + 3C_2}{\epsilon^2}\right). \tag{118}$$

So the total number of trajectories is at most

$$N\overline{K} = \epsilon^{-4/3}\overline{K} \leqslant 12(J(\boldsymbol{\theta}^*) - J(\boldsymbol{\theta}_0))\left(\frac{3C_1\widetilde{V}}{\epsilon^4} + \frac{C_1 + 3C_2}{\epsilon^{10/3}}\right). \tag{119}$$

□

C Auxiliary Results

Proposition 1. *Let $\boldsymbol{X}$ be a zero-mean σ-subgaussian random vector in $\mathbb{R}^d$ in the sense of Assumption 4. Then*

$$\mathbb{E}\left[\|\boldsymbol{X}\|_2^2\right] \leqslant \sigma^2.$$

Proof. For any $\lambda > 0$ and $t \in \mathbb{R}^d$ with $\|\boldsymbol{t}\|_2 = 1$, by hypothesis, $\mathbb{E}[\exp(\lambda\boldsymbol{t}^\top\boldsymbol{X})] \leqslant \exp(\lambda^2\sigma^2/2)$. Then

$$1 + \lambda\boldsymbol{t}^\top\,\mathbb{E}[\boldsymbol{X}] + \frac{\lambda^2}{2}\,\mathbb{E}[(\boldsymbol{t}^\top\boldsymbol{X})^2] + o(\lambda^2) \leqslant 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2), \tag{120}$$

so $\mathbb{E}[(\boldsymbol{t}^\top\boldsymbol{X})^2] \leqslant \sigma^2$. The proof is concluded by noting that $\|\boldsymbol{X}\|_2 = \sup_{\boldsymbol{t}\in\mathbb{R}^d:\|\boldsymbol{t}\|_2=1}\{\boldsymbol{t}^\top\boldsymbol{X}\}$. □

Proposition 2 (Lemma 4.4 from Yuan et al. (2022))**.** *If there are constants $L_1, L_2 > 0$ such that the following holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $s \in \mathcal{S}$ (E-LS, Assumption 4.1 inYuan et al. (2022)):*

$$\mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\|\nabla \log \pi_{\boldsymbol{\theta}}(a|s)\|_2^2 \right] \leqslant L_1^2, \tag{121}$$

$$\mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\left\|\nabla^2 \log \pi_{\boldsymbol{\theta}}(a|s)\right\|_2 \right] \leqslant L_2, \tag{122}$$

then $\left\|\nabla^2 J(\boldsymbol{\theta})\right\|_2 \leqslant \frac{R_{\max}(L_1^2 + L_2)}{(1-\gamma)^2}$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

D Additional numerical results

In this section, we report the full experimental results of Section 6, with different target policy parameters (Table 2), standard deviations (Table 3), LQ horizons (Table 4) and state dimensions (Table 5). Each experiment was repeated 100 times and run with different hyper-parameters of our off-policy method, i.e., the defensive coefficient β, the biased off-policy practical gradient calculation (the offline estimation of the KL divergence here is not possible), and the batch sizes N_{BPO} and N_{PG}.

Table 2: LQ environment, with horizon = 2 and state dimension = 1, and target policy with $\log \sigma = 0$. Variance reduction in off-policy gradient, expressed as $\Delta\mathbb{V}\text{ar}$ and its 95% Gaussian confidence interval $(\Delta\mathbb{V}\text{ar}^-, \Delta\mathbb{V}\text{ar}^+)$, with different hyper-parameters and values of $\boldsymbol{\theta}$.

$\Delta\mathbb{V}\text{ar}$	$\Delta\mathbb{V}\text{ar}^-$	$\Delta\mathbb{V}\text{ar}^+$	biased	β	N_{BPO}	N_{PG}	$\boldsymbol{\theta}$
0.311 033	−0.068 349	0.690 415	False	0.0	10	90	−1.0
0.209 282	−0.216 109	0.634 673	False	0.0	50	50	−1.0
0.321 055	−0.169 663	0.811 773	False	0.4	10	90	−1.0
0.306 358	−0.159 384	0.772 100	False	0.4	50	50	−1.0
0.290 852	−0.136 284	0.717 988	False	0.8	10	90	−1.0
0.029 209	−0.385 136	0.443 554	False	0.8	50	50	−1.0
0.508 645	0.032 941	0.984 350	True	0.0	10	90	−1.0
0.703 738	0.253 894	1.153 583	True	0.0	30	70	−1.0
0.398 966	0.183 075	0.614 856	True	0.0	50	50	−1.0
0.270 046	−0.144 759	0.684 851	True	0.4	10	90	−1.0
0.469 772	0.258 537	0.681 006	True	0.4	50	50	−1.0
0.235 018	−0.137 044	0.607 080	True	0.8	10	90	−1.0
0.561 355	0.302 557	0.820 153	True	0.8	50	50	−1.0
0.140 721	0.006 513	0.274 928	False	0.0	10	90	−0.5
0.106 241	−0.011 837	0.224 319	False	0.0	30	70	−0.5
0.004 764	−0.112 903	0.122 432	False	0.0	50	50	−0.5
0.111 122	−0.034 218	0.256 462	False	0.4	10	90	−0.5
−0.027 326	−0.127 503	0.072 851	False	0.4	50	50	−0.5
0.037 222	−0.083 851	0.158 295	False	0.8	10	90	−0.5
−0.050 209	−0.168 186	0.067 768	False	0.8	50	50	−0.5
0.047 626	−0.069 773	0.165 025	True	0.0	10	90	−0.5
0.220 818	0.068 769	0.372 868	True	0.0	50	50	−0.5
−0.016 716	−0.179 981	0.146 548	True	0.4	10	90	−0.5
0.222 632	0.064 101	0.381 162	True	0.4	50	50	−0.5
0.078 851	−0.041 082	0.198 785	True	0.8	10	90	−0.5
0.195 087	0.059 070	0.331 105	True	0.8	50	50	−0.5
0.055 112	−0.037 841	0.148 065	False	0.0	10	90	0.0
−0.025 057	−0.207 522	0.157 408	False	0.0	30	70	0.0
−0.093 295	−0.238 400	0.051 810	False	0.0	50	50	0.0

−0.055 413	−0.187 053	0.076 227	False	0.4	10	90	0.0
−0.134 235	−0.228 541	−0.039 929	False	0.4	50	50	0.0
−0.044 929	−0.146 132	0.056 273	False	0.8	10	90	0.0
−0.031 705	−0.121 440	0.058 030	False	0.8	30	70	0.0
−0.144 370	−0.240 908	−0.047 833	False	0.8	50	50	0.0
0.063 952	−0.017 625	0.145 529	True	0.0	10	90	0.0
0.120 536	0.057 602	0.183 469	True	0.0	50	50	0.0
0.044 606	−0.063 858	0.153 070	True	0.4	10	90	0.0
0.094 860	0.012 727	0.176 992	True	0.4	50	50	0.0
0.035 522	−0.039 497	0.110 541	True	0.8	10	90	0.0
0.120 686	0.060 903	0.180 469	True	0.8	50	50	0.0
0.392 953	−0.018 980	0.804 886	False	0.0	10	90	0.5
0.122 100	−0.185 945	0.430 145	False	0.0	50	50	0.5
−0.053 468	−0.408 500	0.301 563	False	0.4	10	90	0.5
−0.094 985	−0.448 332	0.258 362	False	0.4	50	50	0.5
0.058 454	−0.334 086	0.450 995	False	0.8	10	90	0.5
−0.233 754	−0.643 237	0.175 729	False	0.8	30	70	0.5
−0.285 911	−0.647 637	0.075 815	False	0.8	50	50	0.5
0.217 064	−0.100 778	0.534 905	True	0.0	10	90	0.5
0.324 804	0.159 038	0.490 571	True	0.0	30	70	0.5
0.204 845	−0.114 787	0.524 477	True	0.0	50	50	0.5
0.084 464	−0.244 899	0.413 827	True	0.4	10	90	0.5
0.408 988	0.144 608	0.673 367	True	0.4	50	50	0.5
0.177 405	−0.197 537	0.552 347	True	0.8	10	90	0.5
0.296 821	0.071 193	0.522 449	True	0.8	50	50	0.5
1.388 987	0.323 475	2.454 499	False	0.0	10	90	1.0
0.562 928	−0.714 201	1.840 058	False	0.0	50	50	1.0
0.006 273	−1.342 498	1.355 045	False	0.4	10	90	1.0
−1.602 914	−3.087 272	−0.118 555	False	0.4	50	50	1.0
0.163 557	−1.012 538	1.339 652	False	0.8	10	90	1.0
−1.083 920	−2.889 235	0.721 395	False	0.8	50	50	1.0
1.643 050	−0.103 186	3.389 286	True	0.0	10	90	1.0
1.260 688	0.628 243	1.893 133	True	0.0	50	50	1.0
−0.856 033	−2.625 640	0.913 575	True	0.4	10	90	1.0
1.503 771	0.775 425	2.232 117	True	0.4	50	50	1.0
1.148 023	−0.616 740	2.912 785	True	0.8	10	90	1.0
2.048 126	1.127 738	2.968 514	True	0.8	50	50	1.0

Table 3: LQ environment, with horizon = 2 and state dimension = 1, and target policy with $\boldsymbol{\theta} = 0$. Variance reduction in off-policy gradient, expressed as $\Delta\mathbb{V}\mathrm{ar}$ and its 95% Gaussian confidence interval $(\Delta\mathbb{V}\mathrm{ar}^-, \Delta\mathbb{V}\mathrm{ar}^+)$, with different hyper-parameters and values of $\log\sigma$.

$\Delta\mathbb{V}\mathrm{ar}$	$\Delta\mathbb{V}\mathrm{ar}^-$	$\Delta\mathbb{V}\mathrm{ar}^+$	biased	β	N_{BPO}	N_{PG}	$\log\sigma$
−0.005 930	−0.029 759	0.017 899	False	0.0	10	90	−1.0
0.005 660	−0.009 167	0.020 486	False	0.0	30	70	−1.0
−0.012 477	−0.029 328	0.004 374	False	0.0	50	50	−1.0
−0.019 162	−0.045 400	0.007 076	False	0.4	10	90	−1.0
−0.009 285	−0.022 311	0.003 742	False	0.4	30	70	−1.0
−0.031 216	−0.049 090	−0.013 342	False	0.4	50	50	−1.0
0.003 573	−0.007 268	0.014 413	False	0.8	10	90	−1.0
−0.001 659	−0.016 295	0.012 977	False	0.8	30	70	−1.0

−0.019 955	−0.042 961	0.003 050	False	0.8	50	50	−1.0
0.007 892	−0.012 674	0.028 458	True	0.0	10	90	−1.0
0.003 067	−0.012 134	0.018 267	True	0.0	30	70	−1.0
0.022 473	0.009 903	0.035 043	True	0.0	50	50	−1.0
0.009 122	−0.008 918	0.027 162	True	0.4	10	90	−1.0
0.011 459	−0.002 588	0.025 507	True	0.4	30	70	−1.0
0.024 397	0.013 399	0.035 394	True	0.4	50	50	−1.0
0.008 570	−0.006 337	0.023 477	True	0.8	10	90	−1.0
0.013 768	0.000 079	0.027 457	True	0.8	30	70	−1.0
0.021 262	0.006 614	0.035 910	True	0.8	50	50	−1.0
0.008 268	−0.029 189	0.045 725	False	0.0	10	90	−0.5
−0.005 033	−0.033 272	0.023 205	False	0.0	30	70	−0.5
−0.012 141	−0.053 830	0.029 549	False	0.0	50	50	−0.5
−0.010 019	−0.055 467	0.035 429	False	0.4	10	90	−0.5
−0.041 924	−0.082 961	−0.000 888	False	0.4	30	70	−0.5
−0.017 607	−0.063 668	0.028 453	False	0.4	50	50	−0.5
0.005 316	−0.024 194	0.034 827	False	0.8	10	90	−0.5
−0.013 986	−0.039 499	0.011 528	False	0.8	30	70	−0.5
−0.036 312	−0.075 203	0.002 580	False	0.8	50	50	−0.5
−0.020 111	−0.084 565	0.044 343	True	0.0	10	90	−0.5
0.015 060	−0.035 335	0.065 454	True	0.0	30	70	−0.5
0.043 786	0.024 192	0.063 380	True	0.0	50	50	−0.5
0.006 319	−0.025 938	0.038 576	True	0.4	10	90	−0.5
0.026 350	−0.001 615	0.054 315	True	0.4	30	70	−0.5
0.047 319	0.022 475	0.072 162	True	0.4	50	50	−0.5
−0.008 211	−0.035 239	0.018 817	True	0.8	10	90	−0.5
0.033 399	0.010 778	0.056 020	True	0.8	30	70	−0.5
0.039 081	0.017 615	0.060 547	True	0.8	50	50	−0.5
0.055 784	−0.044 796	0.156 364	False	0.0	10	90	0.0
0.041 080	−0.035 048	0.117 208	False	0.0	30	70	0.0
−0.005 922	−0.126 384	0.114 540	False	0.0	50	50	0.0
−0.108 877	−0.277 607	0.059 853	False	0.4	10	90	0.0
−0.068 676	−0.192 034	0.054 683	False	0.4	30	70	0.0
−0.014 881	−0.095 332	0.065 571	False	0.4	50	50	0.0
−0.023 581	−0.100 187	0.053 025	False	0.8	10	90	0.0
−0.019 248	−0.106 283	0.067 788	False	0.8	30	70	0.0
−0.110 193	−0.239 846	0.019 460	False	0.8	50	50	0.0
−0.012 017	−0.095 599	0.071 565	True	0.0	10	90	0.0
0.065 525	−0.001 559	0.132 608	True	0.0	30	70	0.0
0.103 235	0.046 921	0.159 549	True	0.0	50	50	0.0
0.010 584	−0.077 404	0.098 572	True	0.4	10	90	0.0
0.043 050	−0.036 176	0.122 275	True	0.4	30	70	0.0
0.129 326	0.022 730	0.235 923	True	0.4	50	50	0.0
0.033 573	−0.057 323	0.124 468	True	0.8	10	90	0.0
0.042 062	−0.015 675	0.099 800	True	0.8	30	70	0.0
0.124 324	0.048 137	0.200 510	True	0.8	50	50	0.0
−0.033 518	−0.456 779	0.389 743	False	0.0	10	90	0.5
0.151 897	−0.199 452	0.503 245	False	0.0	30	70	0.5
0.157 868	−0.245 560	0.561 297	False	0.0	50	50	0.5
−0.261 459	−0.687 112	0.164 193	False	0.4	10	90	0.5
−0.044 700	−0.398 698	0.309 298	False	0.4	30	70	0.5
−0.136 862	−0.578 197	0.304 473	False	0.4	50	50	0.5
−0.160 264	−0.574 757	0.254 229	False	0.8	10	90	0.5
−0.293 061	−0.759 271	0.173 149	False	0.8	30	70	0.5

−0.288 713	−0.862 787	0.285 361	False	0.8	50	50	0.5
0.161 052	−0.278 498	0.600 601	True	0.0	10	90	0.5
0.148 964	−0.147 220	0.445 148	True	0.0	30	70	0.5
0.556 353	0.215 147	0.897 560	True	0.0	50	50	0.5
0.105 981	−0.300 877	0.512 839	True	0.4	10	90	0.5
0.227 993	−0.026 717	0.482 703	True	0.4	30	70	0.5
0.483 820	0.186 378	0.781 262	True	0.4	50	50	0.5
0.240 989	−0.039 873	0.521 851	True	0.8	10	90	0.5
0.419 434	0.145 579	0.693 288	True	0.8	30	70	0.5
0.590 495	0.244 142	0.936 848	True	0.8	50	50	0.5
1.535 046	−0.378 748	3.448 839	False	0.0	10	90	1.0
1.186 207	−0.690 749	3.063 163	False	0.0	30	70	1.0
0.581 094	−1.889 402	3.051 590	False	0.0	50	50	1.0
−0.436 245	−2.535 319	1.662 828	False	0.4	10	90	1.0
0.392 720	−1.539 439	2.324 879	False	0.4	30	70	1.0
−0.407 481	−2.796 212	1.981 250	False	0.4	50	50	1.0
−0.025 073	−2.070 740	2.020 595	False	0.8	10	90	1.0
0.604 685	−1.277 262	2.486 632	False	0.8	30	70	1.0
−2.374 359	−5.105 622	0.356 903	False	0.8	50	50	1.0
2.055 510	0.523 027	3.587 994	True	0.0	10	90	1.0
3.247 087	1.631 413	4.862 761	True	0.0	30	70	1.0
3.176 471	1.951 825	4.401 116	True	0.0	50	50	1.0
−0.638 350	−2.931 465	1.654 765	True	0.4	10	90	1.0
2.361 232	−0.389 329	5.111 792	True	0.4	30	70	1.0
3.773 828	2.399 339	5.148 317	True	0.4	50	50	1.0
−0.121 002	−2.025 865	1.783 861	True	0.8	10	90	1.0
2.700 678	0.718 395	4.682 962	True	0.8	30	70	1.0
4.041 229	2.016 358	6.066 101	True	0.8	50	50	1.0

Table 4: LQ environment, with state dimension = 1, and target policy with $\boldsymbol{\theta} = 0$ and $\log \sigma = 0$. Variance reduction in off-policy gradient, expressed as $\Delta\mathbb{V}\text{ar}$ and its 95% Gaussian confidence interval $(\Delta\mathbb{V}\text{ar}^-, \Delta\mathbb{V}\text{ar}^+)$, with different hyper-parameters and values of LQ horizon.

$\Delta\mathbb{V}\text{ar}$	$\Delta\mathbb{V}\text{ar}^-$	$\Delta\mathbb{V}\text{ar}^+$	biased	β	N_{BPO}	N_{PG}	horizon
0.069 930	−0.046 726	0.186 585	False	0.0	10	90	2
0.041 136	−0.072 254	0.154 527	False	0.0	30	70	2
−0.005 922	−0.126 384	0.114 540	False	0.0	50	50	2
−0.050 883	−0.162 004	0.060 239	False	0.4	10	90	2
0.010 338	−0.076 535	0.097 211	False	0.4	30	70	2
−0.090 330	−0.192 410	0.011 749	False	0.4	50	50	2
0.035 092	−0.055 714	0.125 898	False	0.8	10	90	2
−0.007 530	−0.102 390	0.087 330	False	0.8	30	70	2
−0.115 648	−0.213 301	−0.017 995	False	0.8	50	50	2
0.066 612	−0.001 504	0.134 728	True	0.0	10	90	2
0.085 898	0.031 732	0.140 063	True	0.0	30	70	2
0.103 235	0.046 921	0.159 549	True	0.0	50	50	2
0.112 833	0.030 839	0.194 826	True	0.4	10	90	2
0.095 228	−0.006 859	0.197 315	True	0.4	30	70	2
0.149 218	0.056 437	0.241 998	True	0.4	50	50	2
0.042 195	−0.048 001	0.132 391	True	0.8	10	90	2
0.093 129	0.009 514	0.176 744	True	0.8	30	70	2

0.105 378	0.035 148	0.175 607	True	0.8	50	50	2
10.687 620	−2.869 784	24.245 024	False	0.0	10	90	5
7.282 445	−6.616 917	21.181 807	False	0.0	30	70	5
2.874 308	−4.688 494	10.437 109	False	0.0	50	50	5
4.071 531	−5.723 477	13.866 538	False	0.4	10	90	5
0.956 628	−10.018 669	11.931 925	False	0.4	30	70	5
−5.491 321	−18.211 299	7.228 656	False	0.4	50	50	5
0.573 767	−7.492 679	8.640 214	False	0.8	10	90	5
−3.820 528	−12.886 054	5.244 998	False	0.8	30	70	5
−4.917 480	−15.161 070	5.326 109	False	0.8	50	50	5
10.507 537	0.036 861	20.978 213	True	0.0	10	90	5
12.273 186	3.825 430	20.720 942	True	0.0	30	70	5
18.397 351	11.233 154	25.561 549	True	0.0	50	50	5
1.784 933	−7.365 845	10.935 710	True	0.4	10	90	5
8.188 129	1.217 410	15.158 849	True	0.4	30	70	5
20.694 907	9.166 655	32.223 160	True	0.4	50	50	5
2.638 710	−9.021 860	14.299 280	True	0.8	10	90	5
10.948 408	3.223 581	18.673 235	True	0.8	30	70	5
17.933 160	9.614 722	26.251 598	True	0.8	50	50	5
309.723 170	48.773 653	570.672 686	False	0.0	10	90	10
264.708 738	8.706 979	520.710 497	False	0.0	30	70	10
−310.144 245	−633.900 151	13.611 661	False	0.0	50	50	10
−57.120 902	−253.024 398	138.782 594	False	0.4	10	90	10
−212.141 924	−498.899 103	74.615 254	False	0.4	30	70	10
−429.773 537	−786.701 764	−72.845 309	False	0.4	50	50	10
−133.179 844	−370.851 501	104.491 814	False	0.8	10	90	10
−182.821 632	−456.259 702	90.616 438	False	0.8	30	70	10
−435.518 703	−791.043 397	−79.994 010	False	0.8	50	50	10
220.182 609	11.927 906	428.437 312	True	0.0	10	90	10
287.629 645	102.303 168	472.956 122	True	0.0	30	70	10
397.739 142	159.122 421	636.355 863	True	0.0	50	50	10
31.267 834	−172.938 839	235.474 507	True	0.4	10	90	10
112.227 812	−64.333 427	288.789 050	True	0.4	30	70	10
229.049 254	78.704 906	379.393 601	True	0.4	50	50	10
75.251 773	−214.074 304	364.577 849	True	0.8	10	90	10
147.828 473	−45.398 299	341.055 245	True	0.8	30	70	10
223.758 261	63.647 799	383.868 723	True	0.8	50	50	10

Table 5: LQ environment, with horizon = 2, and target policy with $\boldsymbol{\theta} = 0$ and $\log \sigma = 0$. Variance reduction in off-policy gradient, expressed as $\Delta\mathbb{V}\text{ar}$ and its 95% Gaussian confidence interval $(\Delta\mathbb{V}\text{ar}^-, \Delta\mathbb{V}\text{ar}^+)$, with different hyper-parameters and values of LQ dimensions.

$\Delta\mathbb{V}\text{ar}$	$\Delta\mathbb{V}\text{ar}^-$	$\Delta\mathbb{V}\text{ar}^+$	biased	β	N_{BPO}	N_{PG}	horizon
−8.339 387	−24.727 999	8.049 225	False	0.0	10	90	2
0.015 846	−0.078 860	0.110 552	False	0.0	30	70	2
−0.084 267	−0.288 979	0.120 445	False	0.0	50	50	2
−0.061 526	−0.197 193	0.074 140	False	0.4	10	90	2
−0.057 192	−0.164 759	0.050 375	False	0.4	30	70	2
−0.104 342	−0.228 757	0.020 073	False	0.4	50	50	2
−0.036 944	−0.159 470	0.085 583	False	0.8	10	90	2
−0.086 518	−0.184 832	0.011 796	False	0.8	30	70	2

−0.203 195	−0.335 921	−0.070 469	False	0.8	50	50	2
−0.008 285	−0.214 530	0.197 959	True	0.0	10	90	2
0.104 098	0.032 116	0.176 080	True	0.0	30	70	2
0.238 017	0.131 980	0.344 053	True	0.0	50	50	2
0.011 235	−0.095 540	0.118 011	True	0.4	10	90	2
0.095 955	0.012 872	0.179 039	True	0.4	30	70	2
0.127 433	0.055 541	0.199 325	True	0.4	50	50	2
0.002 206	−0.080 722	0.085 135	True	0.8	10	90	2
0.079 307	0.000 681	0.157 932	True	0.8	30	70	2
0.125 603	0.058 244	0.192 963	True	0.8	50	50	2
0.194 184	−0.083 991	0.472 359	False	0.0	10	90	5
−0.146 855	−0.614 440	0.320 730	False	0.0	30	70	5
−0.177 411	−0.438 773	0.083 951	False	0.0	50	50	5
−0.289 803	−0.550 181	−0.029 424	False	0.4	10	90	5
−0.255 346	−0.520 408	0.009 716	False	0.4	30	70	5
−0.269 124	−0.526 112	−0.012 137	False	0.4	50	50	5
−0.129 578	−0.334 713	0.075 557	False	0.8	10	90	5
−0.123 404	−0.340 821	0.094 013	False	0.8	30	70	5
−0.437 729	−0.671 352	−0.204 107	False	0.8	50	50	5
−0.834 077	−2.383 864	0.715 710	True	0.0	10	90	5
0.182 321	−0.092 779	0.457 422	True	0.0	30	70	5
0.163 729	−0.067 104	0.394 563	True	0.0	50	50	5
−0.229 281	−0.510 593	0.052 031	True	0.4	10	90	5
0.088 913	−0.137 086	0.314 913	True	0.4	30	70	5
0.225 710	0.014 423	0.436 998	True	0.4	50	50	5
−0.046 998	−0.214 187	0.120 191	True	0.8	10	90	5
0.090 860	−0.086 864	0.268 584	True	0.8	30	70	5
0.229 097	0.034 306	0.423 888	True	0.8	50	50	5
1.044 491	0.832 316	1.256 666	False	0.0	10	90	10
0.040 743	−0.419 189	0.500 674	False	0.0	30	70	10
−0.638 193	−1.225 117	−0.051 268	False	0.0	50	50	10
−0.692 391	−1.118 963	−0.265 820	False	0.4	10	90	10
−0.385 588	−0.904 039	0.132 862	False	0.4	30	70	10
−0.746 861	−1.588 713	0.094 990	False	0.4	50	50	10
−0.007 001	−0.385 542	0.371 541	False	0.8	10	90	10
−0.372 875	−0.864 685	0.118 934	False	0.8	30	70	10
−0.936 066	−1.681 347	−0.190 786	False	0.8	50	50	10
−1.728 083	−5.132 161	1.675 995	True	0.0	30	70	10
0.268 508	−0.029 918	0.566 934	True	0.0	50	50	10
−0.118 744	−0.583 906	0.346 419	True	0.4	10	90	10
−0.272 643	−0.906 404	0.361 118	True	0.4	30	70	10
0.130 968	−0.137 975	0.399 911	True	0.4	50	50	10
0.194 654	−0.199 835	0.589 142	True	0.8	10	90	10
0.306 706	−0.081 120	0.694 532	True	0.8	30	70	10
0.601 394	0.316 451	0.886 338	True	0.8	50	50	10