On the Relation between Policy Improvement and Off-Policy Minimum-Variance Policy Evaluation

Alberto Maria Metelli[1] **Samuele Meta**[1] **Marcello Restelli**[1]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

Abstract

Off-policy methods are the basis of a large number of effective Policy Optimization (PO) algorithms. In this setting, Importance Sampling (IS) is typically employed for off-policy evaluation, with the goal of estimating the performance of a target policy, given samples collected with a different behavioral policy. However, in Monte Carlo simulation, IS represents a variance minimization approach. In this field, a suitable behavioral distribution is employed for sampling, allowing diminishing the variance of the estimator below the one achievable when sampling from the target distribution. In this paper, we analyze IS in these two guises in the context of PO. We provide a novel view of off-policy PO, showing a connection between the policy improvement and variance minimization objectives. Then, we illustrate how minimizing the off-policy variance can, in some circumstances, lead to a policy improvement, with the advantage, compared with direct off-policy learning, of implicitly enforcing a trust region. Finally, we present numerical simulations on continuous RL benchmarks, with a particular focus on the robustness to small batch sizes.

1 INTRODUCTION

Policy Optimization methods [PO, Deisenroth et al., 2013] have been widely exploited in Reinforcement Learning [RL, Sutton and Barto, 2018] with successful results in addressing, to name a few, continuous-control [e.g., Peters and Schaal, 2008, Lillicrap et al., 2016], robot manipulation [e.g., Gu et al., 2017, Chatzilygeroudis et al., 2020], and locomotion [e.g., Kohl and Stone, 2004, Duan et al., 2016]. Most of these algorithms employ the notion of *trust region* [Conn et al., 2000], introduced ante litteram in the RL literature by the *safe* RL approaches [Kakade and Langford, 2002, Pirotta et al., 2013], giving rise to a surge of effective algorithms, having TRPO [Schulman et al., 2015] as the progenitor. The core of any RL algorithm, being value-based or policy-based, lies in the ability to employ the samples collected with the current (or *behavioral*) policy to evaluate the performance of a candidate (or *target*) policy [Sutton and Barto, 2018]. The skeleton rationale behind the usage of a trust region is to control the set of candidate policies whose performance can be accurately evaluated. Intuition suggests that if the candidate policy is "sufficiently close" to the current one, this *off-policy* evaluation problem [Precup et al., 2000] will provide a good estimate for the performance of the candidate policy. Formally, this idea has been studied in the field of Importance Sampling [IS, Owen, 2013] and the phenomenon is particularly apparent looking at the IS estimator variance, which grows exponentially with the Rényi divergence [Rényi, 1961] between the behavioral and the target policy [Metelli et al., 2018, 2020, Liotet et al., 2022]. In this off-policy learning setting, IS is employed as a *what-if* analysis tool [Owen, 2013] and its role is *passive*, as samples have been already collected with the current behavioral policy. In this sense, the trust region is an *a-posteriori* remedy for the limitations of off-policy evaluation, for controlling the uncertainty injected by the IS procedure.

However, IS originated in the Monte Carlo simulation community [Hesterberg, 1988, Hammersley, 2013] as an *active* tool for *variance minimization*. While in off-policy learning, the behavioral policy is fixed and we look for the best target policy, whose performance we aim to estimate, here the roles are reversed. Indeed, in off-policy minimum-variance evaluation, the target policy is fixed and we search for the behavioral policy (from which to collect samples) that yields an IS estimate with the minimum possible variance [Hammersley, 2013, Kahn and Marshall, 1953]. It might seem surprising, at first, that sampling from a policy, other than the target one, can lead to an estimator with less variance (even zero in some cases) w.r.t. the on-policy estimate. In

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:1423–1433.

this role, IS has been previously employed in RL, mainly to address rare events [Frank et al., 2008, Ciosek and Whiteson, 2017] which naturally lead to high-variance estimates, when tackled on-policy. The idea of explicitly using IS as a variance reduction technique, with the goal of finding an optimal behavioral policy, was proposed by [Hanna et al., 2017] for evaluation and subsequently combined with policy gradient learning [Hanna and Stone, 2018, Hanna et al., 2019].

Contributions The goal of this paper is to investigate the relation between *policy improvement* and *off-policy minimum variance policy evaluation*. Intuitively, given a target policy, when the reward function is positive, one way to reduce the variance of the IS estimator is to assign larger probability to the trajectories that have a large impact on the mean, i.e., those with high returns. Thus, in some circumstances, reducing the variance of the IS estimator moves the policy towards a policy improvement direction. After having introduced the background (Section 2), we present the problem of finding the minimum-variance behavioral distribution (Section 3). Then, we study the properties of such a distribution in relation with policy improvement in two settings: unconstrained (Section 4) and constrained (Section 5). First, we assume that there are no restrictions for choosing the behavioral distribution. We show that the minimum-variance behavioral distribution, besides leading to the zero-variance estimator [Kahn and Marshall, 1953], is guaranteed to yield a policy improvement, requiring the non-negativity of the reward only. Furthermore, we prove that this approach allows controlling the divergence between two consecutive distributions, thus enforcing an implicit *trust region*. Although this provides a valuable starting point, the minimum-variance distribution might be unrealizable given the environment transition model, i.e., there might be no policy inducing it. For this reason, we move to the scenario in which the distributions are constrained in a suitable space. In this setting, the zero-variance estimator could not be achievable. Furthermore, the presence of a constrained space introduces a bias in terms of policy improvement, still preserving the trust region enforcement. Finally, we provide numerical simulations on both *action-based* and *parameter-based* paradigms of policy optimization Metelli et al. [2018] to test the effects of minimum-variance policy evaluation in comparison with policy optimization. The simulation are conducted on continuous-control benchmarks, in comparison with POIS [Metelli et al., 2018] and TRPO [Schulman et al., 2015], with a particular focus on the robustness of to small batch sizes (Section 6). The proof of the results presented in the main paper are reported in Appendix 1.

2 PRELIMINARIES

In this section, we provide the necessary background that will be employed in the paper.

Mathematical Notation Let $\mathcal{X}$ be a set, and let $\mathfrak{F}_{\mathcal{X}}$ be a σ-algebra over $\mathcal{X}$. We denote with $\mathscr{P}(\mathcal{X})$ the space of probability measures over $(\mathcal{X}, \mathfrak{F}_{\mathcal{X}})$. Let $P \in \mathscr{P}(\mathcal{X})$, whenever needed, we assume that P admits a density function p. For a subset $\mathcal{Y} \subseteq \mathbb{R}$, we denote with $\mathscr{B}(\mathcal{X}, \mathcal{Y})$ the space of measurable functions $f : \mathcal{X} \to \mathcal{Y}$. Let $P, Q \in \mathscr{P}(\mathcal{X})$ be two probability measures such that $P \ll Q$, i.e., P is absolutely continuous w.r.t. Q, for every $\alpha \in [0, \infty]$, we define the *α-Rényi divergence* as [Rényi, 1961]: $D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}x$. In the limit of $\alpha \to 1$, the Rényi divergence reduces to the KL-divergence $D_{\mathrm{KL}}(P\|Q)$, while for $\alpha \to \infty$, it corresponds to $\operatorname{ess\,sup}_{x \sim Q} \{p(x)/q(x)\}$. Let $\alpha \in (0, +\infty)$, a set of probability measures $\mathcal{Q}$ is *α-convex* [van Erven and Harremoës, 2014] if for every $P, Q \in \mathcal{Q}$ and $\lambda \in [0, 1]$ it holds that the probability measure $Q_\lambda := Z_\lambda^{-1} \left(\lambda P^\alpha + (1-\lambda) Q^\alpha\right)^{1/\alpha} \in \mathcal{Q}$, where Z_λ is a normalization constant.

Importance Sampling Let $P, Q \in \mathscr{P}(\mathcal{X})$ with $P \ll Q$ and let $f \in \mathscr{B}(\mathcal{X}, \mathbb{R})$. Importance Sampling [IS, Owen, 2013] allows estimating the expectation of f under a *target* distribution P, i.e., $\mathbb{E}_{x \sim P}[f(x)]$ having samples $\{x_i\}_{i \in [n]}$ collected with a *behavioral* distribution Q: $\widehat{\mu}_{P/Q} = \frac{1}{n} \sum_{i \in [n]} \frac{p(x_i)}{q(x_i)} f(x_i)$. The IS estimator is unbiased [Owen, 2013], i.e., $\mathbb{E}_{x_i \sim Q}[\widehat{\mu}_{P/Q}] = \mathbb{E}_{x \sim P}[f(x)]$, but it might suffer from large variance, due to the heavy-tailed behavior [Metelli et al., 2018, 2021b]. The properties of $\widehat{\mu}_{P/Q}$ and its transformations have been extensively studied in the literature [e.g., Ionides, 2008, Thomas et al., 2015, Papini et al., 2019, Metelli et al., 2020, Kuzborskij et al., 2021, Metelli et al., 2021a].

Policy Optimization A Markov Decision Process [MDP, Puterman, 1994] is a 6-tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, D_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the transition model, $R : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the reward function, $\gamma \in [0, 1]$ is the discount factor, and $D_0 \in \mathscr{P}(\mathcal{S})$ is the initial state distribution. The agent's behavior is modeled by a *parametric* policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \mathscr{P}(\mathcal{A})$ belonging to a parametric policy space $\Pi_\Theta = \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$. The interaction between an agent and the MDP generates a *trajectory* $\tau = (s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}, s_H)$ where $H \in \mathbb{N}$ is the trajectory length and $s_0 \sim D_0$, $a_t \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$ for all $t \in \{0, \ldots, H-1\}$. Given a trajectory τ, the *return* is the discounted sum of the rewards $\mathcal{R}(\tau) = \sum_{t=0}^{H-1} \gamma^t R(s_t, a_t)$. For a policy $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$, we denote with $p(\cdot|\boldsymbol{\theta})$ the induced trajectory distribution: $p(\tau|\boldsymbol{\theta}) = D_0(s_0) \prod_{t=0}^{H-1} \pi_{\boldsymbol{\theta}}(a_t|s_t) P(s_{t+1}|s_t, a_t)$. In the *action-based* (AB) setting, an agent aims at finding a parametrization fulfilling: $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta} \in \Theta} \{J(\boldsymbol{\theta})\}$, where:

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p(\cdot|\boldsymbol{\theta})} \left[\mathcal{R}(\tau)\right]$$

is the *expected return*. $\pi_{\boldsymbol{\theta}}$ must be stochastic to ensure exploration. Instead, in the *parameter-based* (PB) setting, we

consider a *hyperpolicy* $\nu_{\boldsymbol{\rho}} \in \mathscr{P}(\Theta)$, belonging to a parametric hyperpolicy space $\mathcal{N}_{\mathcal{P}} = \{\nu_{\boldsymbol{\rho}} : \boldsymbol{\rho} \in \mathcal{P} \subseteq \mathbb{R}^l\}$, from which we sample the parameters $\boldsymbol{\theta}$ of the policy. In this case, the policy $\pi_{\boldsymbol{\theta}}$ can be deterministic since exploration is managed at the hyperpolicy level and the agent goal becomes to learn a hyperpolicy parametrization maximizing the expected return: $\boldsymbol{\rho}^* \in \arg\max_{\boldsymbol{\rho} \in \mathcal{P}} \{J(\boldsymbol{\rho})\}$, where:

$$J(\boldsymbol{\rho}) = \mathbb{E}_{\boldsymbol{\theta} \sim \nu_{\boldsymbol{\rho}}}[J(\boldsymbol{\theta})].$$

In the paper, we keep the presentation as general as possible, introducing the results for arbitrary distributions. Then, we will particularize for the parametric PO setting.

3 MINIMUM–VARIANCE BEHAVIORAL DISTRIBUTION

In this section, we revise the problem of finding a behavioral distribution $Q \in \mathscr{P}(\mathcal{X})$ that induces an IS estimate $\widehat{\mu}_{P/Q}$ with minimum variance, knowing the (fixed) target distribution $P \in \mathscr{P}(\mathcal{X})$ and function $f \in \mathscr{B}(\mathcal{X}, [0, \infty))$.[1] Furthermore, we do not enforce any restriction on the possible forms of the behavioral distribution $Q \in \mathscr{P}(\mathcal{X})$. The problem and the corresponding well-known *minimum-variance behavioral distribution* Q^* are stated in the following for all $x \in \mathcal{X}$ [Kahn, 1950]:

$$\min_{Q \in \mathscr{P}(\mathcal{X})} \underset{x \sim Q}{\mathbb{V}\mathrm{ar}}\left[\frac{p(x)}{q(x)} f(x)\right] \Longrightarrow q^*(x) = \frac{p(x) f(x)}{\mathbb{E}_{x \sim P}[f(x)]}. \quad (1)$$

We observe that the IS estimator $\widehat{\mu}_{P/Q^*}$ is non-stochastic, equal to the quantity we aim to estimate, i.e., $\widehat{\mu}_{P/Q^*} = \mathbb{E}_{x \sim P}[f(x)]$. This suggests that the construction of Q^* is infeasible as it requires knowledge of $\mathbb{E}_{x \sim P}[f(x)]$. Since Q^* generates a non-stochastic estimator, it not only leads to zero-variance but, clearly, simultaneously minimizes the absolute central moments of any order. A second, and most remarkable property, is that Q^* is a *performance improvement* w.r.t. P, i.e., the expectation of f under Q^* is larger than the expectation of f under the target P [Owen, 2013]:

$$\mathbb{E}_{x \sim Q^*}[f(x)] - \mathbb{E}_{x \sim P}[f(x)] = \frac{\mathbb{V}\mathrm{ar}_{x \sim P}[f(x)]}{\mathbb{E}_{x \sim P}[f(x)]} \geqslant 0. \quad (2)$$

It is worth noting that the magnitude of the improvement is directly related to the reduction in variance $\mathbb{V}\mathrm{ar}_{x \sim P}[f(x)]$. Equation (2) suggests an appealing connection between the problem of finding the minimum-variance behavioral distribution and the problem of finding a target distribution that maximizes the expectation $\mathbb{E}_{x \sim P}[f(x)]$, i.e., policy optimization.

[1] We restrict our attention to non-negative functions. From the RL perspective, this choice is w.l.o.g. since we can always define an equivalent non-negative reward function, by means of a translation of the original one.

Before proceeding, let us map this general setting to PO. In the action-based (AB) setting, x is the trajectory τ, P and Q are trajectory distributions $p(\tau|\boldsymbol{\theta})$ induced by policies $\pi_{\boldsymbol{\theta}}$. Instead, in the parameter-based (PB) setting, x is the pair $(\boldsymbol{\theta}, \tau)$, P and Q are joint distributions $\nu_{\boldsymbol{\rho}}(\boldsymbol{\theta}) p(\tau|\boldsymbol{\theta})$ induced by hyperpolicies $\nu_{\boldsymbol{\rho}}$. In both cases, function f is the trajectory return $\mathcal{R}(\tau)$.

In the following two sections, we will delve into the properties of the minimum-variance distribution under two assumptions: (i) there are no restrictions in the choice of the behavioral distribution $Q \in \mathscr{P}(\mathcal{X})$ (Section 4); (ii) the behavioral distribution must be chosen within a subset $Q \in \mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$ (Section 5).

4 UNCONSTRAINED PROBABILITY DISTRIBUTION SPACE

In Section 3, we have seen that Q^* is a performance improvement w.r.t. P. We now formalize this construction by defining the operator $\mathcal{I}_f : \mathscr{P}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$:

$$(\mathcal{I}_f[P])(x) = \frac{p(x) f(x)}{\mathbb{E}_{x \sim P}[f(x)]}, \quad \forall x \in \mathcal{X}. \quad (3)$$

Thus, $\mathcal{I}_f$ takes as input a target distribution $P \in \mathscr{P}(\mathcal{X})$, a function $f \in \mathscr{B}(\mathcal{X}, [0, \infty))$, and outputs the minimum-variance behavioral distribution for the IS estimation of $\mathbb{E}_{x \sim P}[f(x)]$, i.e., $Q^* = \mathcal{I}_f[P]$. Intuitively, looking at Equation (3), by iterating the application of $\mathcal{I}_f$, we will obtain distributions tending to assign larger probability mass to points $x \in \mathcal{X}$ with high values of $f(x)$. The following result, due to Ghosh et al. [2020], generalizes Equation (2), showing that not only $\mathcal{I}_f[P]$ is a performance improvement w.r.t. P, even when considering a composition between a monotonic increasing function h and f, i.e., using the operator $\mathcal{I}_{h \circ f}$.

Proposition 4.1 (Proposition 9 of Ghosh et al. [2020])**.** *Let $P \in \mathscr{P}(\mathcal{X})$, $f \in \mathscr{B}(\mathcal{X}, [0, \infty))$, and $h : [0, \infty) \to [0, \infty)$ monotonic increasing. Then, $\mathcal{I}_{h \circ f}[P]$ is a performance improvement w.r.t. P:*

$$\underset{x \sim \mathcal{I}_{h \circ f}[P]}{\mathbb{E}}[f(x)] - \underset{x \sim P}{\mathbb{E}}[f(x)] = \frac{\mathbb{C}\mathrm{ov}_{x \sim P}[h(f(x)), f(x)]}{\mathbb{E}_{x \sim P}[h(f(x))]} \geqslant 0.$$

Note that, since h is a monotonic increasing function, we have that $\mathbb{C}\mathrm{ov}_{x \sim P}[h(f(x)), f(x)] \geqslant 0$ [Cuadras, 2002].

4.1 CONVERGENCE PROPERTIES

We now analyze the effect of repeatedly applying operator $\mathcal{I}_f$. More formally, let us consider an initial distribution $P \in \mathscr{P}(\mathcal{X})$, and suppose to iterate the application of

the operator $\mathcal{I}_f$, generating the sequence of distributions $(Q_k)_{k\in\mathbb{N}}$, where $Q_0 = P$ and for every $k \in \mathbb{N}_{\geqslant 0}$ we have $Q_k = \mathcal{I}_f[Q_{k-1}] = (\mathcal{I}_f)^k[P]$. The following result shows that, under certain conditions, the operator $\mathcal{I}_f$ admits fixed points and the sequence $(Q_k)_{k\in\mathbb{N}}$ converges to a distribution Q_∞ that assigns probability only to the global maxima of f, restricted to the support of P, i.e., $\mathrm{supp}(P)$.

Theorem 4.2. *Let $P \in \mathscr{P}(\mathcal{X})$ and $f \in \mathscr{B}(\mathcal{X}, [0,\infty))$. Then, the following statements hold:*

(i) P is a fixed point of $\mathcal{I}_f$, i.e., $\mathcal{I}_f[P] = P$ a.s., if and only if $\mathbb{V}\mathrm{ar}_{x\sim P}[f(x)] = 0$;

(ii) let $\mathcal{X}^* = \arg\max_{x\in\mathrm{supp}(P)}\{f(x)\}$ be the set of maxima of f restricted to the support of P. If $\mathcal{X}^*$ is non-empty and measurable then, the repeated application of $\mathcal{I}_f$ converges to a distribution $Q_\infty = \lim_{k\to\infty}(\mathcal{I}_f)^k[P]$ with support $\mathcal{X}^*$. In particular $\mathbb{E}_{x\sim Q_\infty}[f(x)] = \max_{x\in\mathrm{supp}(P)}\{f(x)\}$.

As a corollary to point (i), any deterministic P is a fixed point of $\mathcal{I}_f$. Furthermore, from point (ii), we deduce that if we select P that assigns non-zero probability to all points in $\mathcal{X}$, i.e., $\mathrm{supp}(P) = \mathcal{X}$, the iterated application of $\mathcal{I}_f$ converges to the distribution Q_∞ such that $\mathbb{E}_{x\sim Q_\infty}[f(x)] = \max_{x\in\mathcal{X}}\{f(x)\}$, i.e., we are performing a global optimization of f. It is worth noting that the reasoning above can be generalized by performing an application of a strictly-increasing function $h : [0,\infty) \to [0,\infty)$ leading to the operator $\mathcal{I}_{h\circ f}$ preserving the same properties.

4.2 IMPLICIT TRUST REGION

We now prove that we are able to naturally control the divergence between two consecutive distributions Q_k and $Q_{k+1} = \mathcal{I}_f[Q_k]$ with $k \in \mathbb{N}$, with the effect of enforcing an *implicit* trust region. The following result shows how it is possible to obtain a bound on the α-Rényi divergence between two consecutive distributions.

Theorem 4.3. *Let $P \in \mathscr{P}(\mathcal{X})$ and $f \in \mathscr{B}(\mathcal{X}, [0,\infty))$. Then, for every $\alpha \in [0,\infty]$, it holds that:*

$$D_\alpha\left(\mathcal{I}_f[P]\|P\right) = \frac{1}{\alpha-1}\log\frac{\mathbb{E}_{x\sim P}[f(x)^\alpha]}{\mathbb{E}_{x\sim P}[f(x)]^\alpha}.$$

In particular, for $\alpha = 1$ it holds that:

$$D_{KL}(\mathcal{I}_f[P]\|P) = \frac{\mathbb{C}\mathrm{ov}_{x\sim P}[f(x), \log f(x)]}{\mathbb{E}_{x\sim P}[f(x)]}.$$

For $\alpha = 2$, we obtain $D_2(I_f[P]\|P) = \log\frac{\mathbb{E}_{x\sim P}[f(x)^2]}{\mathbb{E}_{x\sim P}[f(x)]^2} \leqslant \frac{\mathbb{V}\mathrm{ar}_{x\sim P}[f(x)]}{\mathbb{E}_{x\sim P}[f(x)]^2}$. Thus, the divergence is large when the variance of $f(x)$ is. The result is particularly remarkable as we are able to control the Rényi divergences of *any* order $\alpha \in [0,\infty]$. This is a relevant achievement since the trust regions commonly used, like KL-divergence [Schulman et al., 2015], are unable to control higher-order divergences that can still be infinite.

Example 4.1. *We consider (a slight variation of) the one-dimensional Ackley function [Ackley, 2012]: $f(x) = -5 + 20\exp(-0.1414|x|) + \exp(0.5(\cos(2\pi x)+1)) + e$, shown in Figure 1 (left) and the class of increasing functions $(h\circ f)(x) = f(x)^\beta$ where $\beta \geqslant 0$. We consider an initial uniform distribution $P = \mathrm{Uni}([-5,5])$. In Figure 1, we plot the expectation of distribution $Q_k = (\mathcal{I}_{h\circ f})^k[P]$ (center) and the KL-divergence between two consecutive distributions (right), as a function of the number of applications k, for the different β values. We observe that convergence to the global optimum ($x^* = 0$ and $f(x^*) = 15$) is faster for higher powers that also lead to larger trust regions. We can now appreciate the role of the increasing function h that works as a regularizer with the effect of controlling the size of the trust region.*

5 CONSTRAINED PROBABILITY DISTRIBUTION SPACE

The approach we have presented in Section 4 can be applied when there are *no* restrictions on the class of distributions that can be played, i.e., we can select Q in the whole space $\mathscr{P}(\mathcal{X})$. However, in the action-based PO, we can intervene on the policy $\pi_{\boldsymbol{\theta}}$ factors only of the distribution $p(\tau|\boldsymbol{\theta}) = D_0(s_0)\prod_{t=0}^{H-1}\pi_{\boldsymbol{\theta}}(a_t|s_t)P(s_{t+1}|s_t,a_t)$, leading to a constrained setting. Similarly, in the parameter-based PO, we can act on the hyperpolicy $\nu_{\boldsymbol{\rho}}$ while keeping the trajectory distribution $p(\tau|\boldsymbol{\theta})$ fixed.

More in general, when considering a class of distributions $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$, even if $P \in \mathcal{Q}$, the distribution $\mathcal{I}_f[P]$ might not belong to $\mathcal{Q}$. Furthermore, while $\mathcal{I}_f[P]$ minimizes *all* absolute central α-moments of the IS estimator, as it leads to a non-stochastic estimator (Section 3), there may exist different distributions in $\mathcal{Q}$ minimizing the different absolute central α-moments:

$$\min_{Q\in\mathcal{Q}}\mathbb{E}_{x\sim Q}\left[\left|\frac{p(x)}{q(x)}f(x) - \mathbb{E}_{x\sim P}[f(x)]\right|^\alpha\right]. \qquad (4)$$

Apart from $\alpha = 2$, where the problem in Equation (4) reduces to Equation (1), for general value of $\alpha \in [0,\infty]$, the optimization is not straightforward (e.g., Equation (4) is not differentiable for $\alpha \in (0,2)$). The following result shows that performing a *moment projection* through the α-Rényi divergence is a reasonable surrogate for minimizing the absolute central α-moments of Equation (4).

Proposition 5.1. *Let $P \in \mathscr{P}(\mathcal{X})$ and $f \in \mathscr{B}(\mathcal{X}, [0,\infty))$. Then, for any $\alpha \in [2,\infty)$, it holds that:*

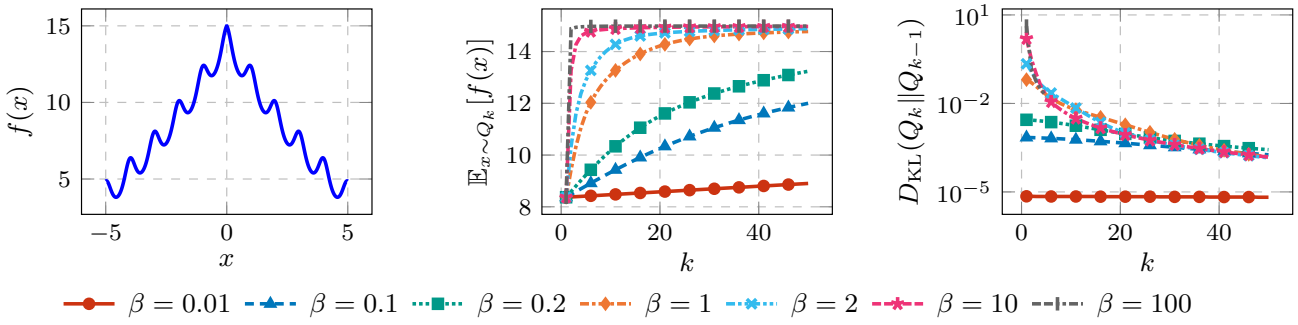

Figure 1: The Ackley function (left), the expectation of the distribution $Q_k = (\mathcal{I}_{h\circ f})^k[P]$ (center), and the KL-divergence (right) between two consecutive distributions Q_{k-1} and Q_k, with $h = (\cdot)^\beta$.

$$\underbrace{\mathbb{E}_{x\sim Q}\left[\left|\frac{p(x)}{q(x)}f(x) - \mathbb{E}_{x\sim P}[f(x)]\right|^\alpha\right]}_{\textit{absolute central } \alpha\textit{-moment}} \leqslant \underbrace{\mathbb{E}_{x\sim Q}\left[\left(\frac{p(x)}{q(x)}f(x)\right)^\alpha\right]}_{\textit{(non-central) } \alpha\textit{-moment}} = e^{(\alpha-1)D_\alpha(\mathcal{I}_f[P]\|Q)}\mathbb{E}_{x\sim P}[f(x)]^\alpha.$$

Thus, having considered the subset of distributions $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$, whenever $\mathcal{I}_f[P] \notin \mathcal{Q}$, we replace it with the corresponding moment projection performed through the α-Rényi divergence:

$$Q^\dagger \in \arg\min_{Q\in\mathcal{Q}} \{D_\alpha(\mathcal{I}_f[P]\|Q)\}. \tag{5}$$

5.1 PERFORMANCE IMPROVEMENT

In Proposition 4.1, we have seen that $\mathcal{I}_f[P]$ is a performance improvement w.r.t. P, evaluated under function f (and also under the composition between f and *any* strictly-increasing function h). Unfortunately, when we move to a constrained set of distributions $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$, the performance improvement cannot be in general guaranteed for function f. However, as we shall see, the performance improvement still holds for a monotonic transformation of f, depending on the choice of α.

Theorem 5.2. *Let $P \in \mathscr{P}(\mathcal{X})$ and $f \in \mathscr{B}(\mathcal{X}, [0,\infty))$. Let $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$, $Q \in \mathcal{Q}$, and $\alpha \in [0,\infty]$, then, it holds that:*

$$\mathop{\mathbb{E}}_{x\sim Q}[f(x))^\alpha] - \mathop{\mathbb{E}}_{x\sim P}[f(x)^\alpha] \geqslant \frac{\mathbb{E}_{x\sim P}[f(x)]^\alpha}{\alpha - 1} \times \left(e^{(\alpha-1)D_\alpha(\mathcal{I}_f[P]\|P)} - e^{(\alpha-1)D_\alpha(\mathcal{I}_f[P]\|Q)}\right)$$

In particular, for $\alpha = 1$, it holds that [Ghosh et al., 2020, Proposition 6]:

$$\mathop{\mathbb{E}}_{x\sim Q}[f(x)] - \mathop{\mathbb{E}}_{x\sim P}[f(x)] \geqslant \mathop{\mathbb{E}}_{x\sim P}[f(x)] \times \left(D_{KL}(\mathcal{I}_f[P]\|P) - D_{KL}(\mathcal{I}_f[P]\|Q)\right).$$

While the inequality holds in general, the performance improvement is obtained provided that $D_\alpha(\mathcal{I}_f[P]\|Q) \leqslant D_\alpha(\mathcal{I}_f[P]\|P)$, which is always guaranteed when $P \in \mathcal{Q}$ and $Q = Q^\dagger$, being $Q^\dagger$ defined in Equation (5) as the minimizer of the second divergence term. The theorem shows that by minimizing the α-moment of the function f, we are able to guarantee a performance improvement on the function $f(\cdot)^\alpha$. In particular, if we select $\alpha = 1$, we obtain a guarantee on the performance improvement of function f. From the RL perspective, therefore, moving in the direction of minimizing the α-moment provides an improvement for the expected α-power of the return $\mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}[\mathcal{R}(\tau)^\alpha]$.

5.2 CONVERGENCE PROPERTIES

By using Equation (5) as an iterate $Q_{k+1} \in \arg\min_{Q\in\mathcal{Q}} \{D_\alpha(\mathcal{I}_f[Q_k]\|Q)\}$ to generate a sequence of distributions $(Q_k)_{k\in\mathbb{N}}$, we are *not* guaranteed to converge to any fixed-point distribution Q_∞, differently form the unconstrained setting (Theorem 4.2). This is because the minimization might yield multiple solutions. Nevertheless, we are able to provide guarantees on the final divergence value and on the performance of the distributions Q_k.

Theorem 5.3. *Let $P \in \mathscr{P}(\mathcal{X})$ and $f \in \mathscr{B}(\mathcal{X}, [0,\infty))$. Let $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$ and suppose that f is bounded from above, then, the iterate $Q_{k+1} \in \arg\min_{Q\in\mathcal{Q}} \{D_\alpha(\mathcal{I}_f[Q_k]\|Q)\}$ (where possible ties are broken arbitrarily) satisfies:*

- *(i) the sequence of divergences $D_\alpha(\mathcal{I}_f[Q_k]\|Q_k)$ is convergent;*
- *(ii) the sequence of expectations $\mathbb{E}_{x\sim Q_k}[f(x)^\alpha]$ is non-decreasing in $k \in \mathbb{N}$ and converges to a stationary point of $\mathbb{E}_{x\sim Q}[f(x)^\alpha]$ w.r.t. $Q \in \mathcal{Q}$.*

The convergence of the sequences $D_\alpha(\mathcal{I}_f[Q_k]\|Q_k)$ and $\mathbb{E}_{x\sim Q_k}[f(x)^\alpha]$ is derived by the performance improvement of Theorem 5.2. Importantly, Theorem 5.3 shows the convergence to a *stationary point* of $\mathbb{E}_{x\sim Q}[f(x)^\alpha]$.

If $\mathcal{Q}$ is a parametric space $\mathcal{Q}_{\boldsymbol{\Xi}} = \{Q_{\boldsymbol{\xi}} \in \mathscr{P}(\mathcal{X}) : \boldsymbol{\xi} \in \boldsymbol{\Xi} \subseteq \mathbb{R}^d\}$,[2] then we are guaranteed to stop when $\mathbb{E}_{x\sim Q_{\boldsymbol{\xi}}}\left[\nabla_{\boldsymbol{\xi}} \log q_{\boldsymbol{\xi}}(x) f(x)^\alpha\right] = 0$, like for a general policy gradient method maximizing $f(x)^\alpha$ [Papini et al., 2018]. Compared to the result for the unconstrained distribution space (Theorem 4.2), we loose the convergence to a fixed point. This property can be recovered under the assumption that the iterate in Equation (5) admits a unique solution for every P. In such a case, we will converge to a distribution $Q_\infty = \arg\min_{Q\in\mathcal{Q}} \{D_\alpha(\mathcal{I}_f[Q]\|Q)\}$.

5.3 IMPLICIT TRUST REGION

In Theorem 4.3, we have proved that the α-Rényi divergence between $\mathcal{I}_f[P]$ and P is bounded. In this section, we study whether similar properties hold when we consider a limited set of distributions $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$. The following result shows that, under a particular form of convexity [van Erven and Harremoës, 2014] of $\mathcal{Q}$, we are able to control the trust region as well.

Theorem 5.4. *Let $\alpha \in [0,1)$ and $f \in \mathscr{B}(\mathcal{X},[0,\infty))$. Let $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$ be a $(1-\alpha)$-convex set [van Erven and Harremoës, 2014, Definition 4], $P \in \mathcal{Q}$, $Q^\dagger \in \arg\min_{Q\in\mathcal{Q}} \{D_\alpha(\mathcal{I}_f[P]\|Q)\}$, then it holds that:*

$$D_\alpha\left(Q^\dagger\|P\right) \leqslant D_\alpha\left(\mathcal{I}_f[P]\|P\right) - D_\alpha\left(\mathcal{I}_f[P]\|Q^\dagger\right).$$

Therefore, we are always guaranteed that the trust region induced by $Q^\dagger$ is tighter compared to the one induced by $Q^* = \mathcal{I}_f[P]$ computed in Theorem 4.3, i.e., $D_\alpha\left(Q^\dagger\|P\right) \leqslant D_\alpha\left(\mathcal{I}_f[P]\|P\right)$.

A summary of the properties of the unconstrained and constrained settings is reported in Table 1.

6 NUMERICAL SIMULATIONS

In this section, we numerically validate the theoretical findings presented in the previous sections. To this end we make use of a sample-based policy learning algorithm *Minimum-Variance Policy Evaluation for Policy Improvement* (MBPExPI). generality, we consider a parametric distribution space $\mathcal{Q}_{\boldsymbol{\Xi}} = \{Q_{\boldsymbol{\xi}} \in \mathscr{P}(\mathcal{X}) : \boldsymbol{\xi} \in \boldsymbol{\Xi} \subseteq \mathbb{R}^d\}$, a common setting met in PO.

6.1 ALGORITHM

The goal of MBPExPI consists in illustrating what are the effects of employing the minimization of the α-moment of the IS estimator to learn proficient policies. As we have seen in the previous section, this approach is guaranteed to yield performance improvement for $\alpha = 1$ only (in the constrained case). However, as we shall see, from an empirical perspective other choices of α would deliver surprisingly remarkable performances too.[3]

Algorithm 1: MBPExPI.

input : α divergence order, h function, f function, $\mathcal{Q}_{\boldsymbol{\Xi}}$ distribution space, $\boldsymbol{\xi}_1$ initial parameter, n batch size

output : final parameter $\boldsymbol{\xi}_{I+1} \in \boldsymbol{\Xi}$

for $i = 1, \dots, I$ **do** ; // Optimization

  $\bar{\boldsymbol{\xi}}_{i,1} = \boldsymbol{\xi}_i$

  for $j = 1, \dots, J$ **do** ; // Evaluation

    Collect $\mathcal{D}_{i,j} = \{(x_l, f(x_l))\}_{l\in[n]}$ with $Q_{\bar{\boldsymbol{\xi}}_{i,j}}$

    Using $(\mathcal{D}_{i,k})_{k\in[j]}$, perform M steps of gradient descent on the objective of Theorem 6.1

  end

  $\boldsymbol{\xi}_{i+1} = \bar{\boldsymbol{\xi}}_{i,J+1}$

end

The structure of MBPExPI consists of two nested loops. The outer loop (Optimization) acts on the target distribution $q_{\boldsymbol{\xi}_i}$. At the end of each outer iteration $i \in [I]$, the target distribution $q_{\boldsymbol{\xi}_{i+1}}$ is updated with the last behavioral distribution produced by the inner loop $q_{\bar{\boldsymbol{\xi}}_{i,J+1}}$. Instead, the inner loop (Evaluation) takes the target distribution provided by the outer loop $q_{\boldsymbol{\xi}_i}$ and provides a new behavioral distribution. At each inner iteration $j \in [J]$, it collects samples $\mathcal{D}_{i,j}$ with the current behavioral distribution $q_{\bar{\boldsymbol{\xi}}_{i,j}}$ and employs them, together with all the samples collected so far $(\mathcal{D}_{i,k})_{k\in[j]}$, to compute the next behavioral distribution $q_{\bar{\boldsymbol{\xi}}_{i,j+1}}$, with the goal of finding the behavioral distribution minimizing the absolute central α-moment of the IS estimator (Equation 4). As we shall see, the optimization is performed using samples and by resorting to a penalized objective.

Sample-based Optimization The problem of finding the next behavioral distribution parameter $\bar{\boldsymbol{\xi}}_{i,j+1}$ using the samples collected so far $(\mathcal{D}_{i,k})_{k\in[j]}$ is an off-policy learning problem. Let us define $\Phi_{i,j} = \frac{1}{j}\sum_{k\in[j]} q_{\bar{\boldsymbol{\xi}}_{i,k}}$ as the mixture of the j behavioral distributions experienced so far in the inner loop. Instead of directly estimating $D_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}]\|Q_{\boldsymbol{\xi}}\right))$, we refer to the (non-central) α-moment, which is connected to the original objective through Proposition 5.1. Since we

[2] In the action-based PO $\boldsymbol{\xi} = \boldsymbol{\theta}$ are the policy parameters and $\boldsymbol{\Xi} = \Theta$, while in the parameter-based PO $\boldsymbol{\xi} = \boldsymbol{\rho}$ are the hyperpolicy parameters and $\boldsymbol{\Xi} = \mathcal{P}$.

[3] While we limit our presentation to *actor-only* algorithms, our framework can be applied to *actor-critic* methods by setting, for instance, $f = Q_{\boldsymbol{w}}$ (i.e., the critic) and $q_{\boldsymbol{\xi}} = \pi_{\boldsymbol{\theta}}$ (i.e., the actor). Clearly, the convergence properties of such an approach would depend on the critic accuracy.

Setting	Iterate	Performance improvement	Convergence	In policy search?
Unconstrained	$Q_{k+1} = \mathcal{I}_f[Q_k]$	Yes, on $h \circ f$ (h any monotonic increasing)	Global optimum of f	Not realistic
Constrained	$Q_{k+1} \in \arg\min_{Q \in \mathcal{Q}} D_\alpha(\mathcal{I}_f[Q_k] \| Q)$	Yes, on $f(\cdot)^\alpha$	Stationary point of $\mathbb{E}_{x\sim Q}[f(x)^\alpha]$	Realistic

Table 1: Summary of the properties of the constrained and unconstrained settings.

have samples coming from different behavioral distributions, we can use a *multiple* IS estimator [Veach and Guibas, 1995]:[4]

$$\widehat{d}_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}] \| Q_{\boldsymbol{\xi}}; \Phi_{i,j}\right) = \frac{1}{nj} \sum_{k\in[j]} \sum_{l\in[n]} \underbrace{\frac{q_{\boldsymbol{\xi}}(x_{k,l})}{\Phi_{i,j}(x_{k,l})}}_{(a)} \times \underbrace{\frac{q_{\boldsymbol{\xi}_i}(x_{k,l})^\alpha}{q_{\boldsymbol{\xi}}(x_{k,l})^\alpha} f(x_{k,l})^\alpha}_{(b)}. \tag{6}$$

The (a) factor accounts that we are using samples collected with the mixture $\Phi_{i,j}$ to estimate an expectation under $q_{\boldsymbol{\xi}}$, whereas the (b) factor is the actual variable we want to compute the expectation of, i.e., the α-moment. It is simple to prove that the expectation of $\widehat{d}_\alpha$ is indeed the α-moment [Papini et al., 2019]. To minimize Equation (6), we employ a variance correction to mitigate the effect of finite samples [Metelli et al., 2018], theoretically grounded in the following result.

Theorem 6.1. *Let $\mathcal{Q}_{\boldsymbol{\Xi}} \subseteq \mathscr{P}(\mathcal{X})$ be a set of parametric distributions and let $\boldsymbol{\xi}, \boldsymbol{\xi}_i \in \boldsymbol{\Xi}$. If $\|f\|_\infty \leqslant \overline{m}$, then, if samples are independent, for every $\delta \in [0,1]$, with probability at least $1-\delta$ it holds that:*

$$\mathbb{E}_{x\sim\boldsymbol{\xi}}\left[\left(\frac{q_{\boldsymbol{\xi}_i}(x)}{q_{\boldsymbol{\xi}}(x)} f(x)\right)^\alpha\right] \leqslant \widehat{d}_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}] \| Q_{\boldsymbol{\xi}}; \Phi_{i,j}\right) + \overline{m}^\alpha \sqrt{\frac{2\log\frac{1}{\delta}}{nj} \int_{\mathcal{X}} \frac{q_{\boldsymbol{\xi}_i}(x)^{2\alpha}}{\Phi_{i,j}(x) q_{\boldsymbol{\xi}}(x)^{2(\alpha-1)}} \mathrm{d}x}.$$

Some remarks are in order. First, the integral within the square root is an upper bound to the variance of the α-moment estimator $\widehat{d}_\alpha\left(\mathcal{I}_f[Q_{\boldsymbol{\xi}_i}] \| Q_{\boldsymbol{\xi}}; \Phi_{i,j}\right)$. In particular, when $\boldsymbol{\xi} = \boldsymbol{\xi}_i$, we obtain the exponentiated Rényi divergence, as illustrated in [Metelli et al., 2020]. When all involved distributions are Guassians, it is possible to provide a closed-form tight bound on this quantity (Appendix 2). Second, unlike the results available in the literature about concentration of IS estimator, without corrections or transformations, we are able to provide an exponential concentration inequality (dependence on δ of the form $\log(1/\delta)$), instead of a polynomial concentration (dependence of the form $1/\delta$). This is due to the fact that we are dealing with random variables that are bounded to zero from below and they allow applying stronger unilateral Bernstein's concentration inequalities [Boucheron et al., 2009]. The reader might object that to optimize the proposed objective function, designed to enforce an implicit trust region, we are actually introducing an additional correction term. This is necessary for theoretical purposes, but, as we shall see in the Section 6.2, the need for a penalization or constraint is significantly less relevant than in existing approaches, like TRPO [Schulman et al., 2015], or POIS [Metelli et al., 2018]. The expression of the gradient of the right hand side of Theorem 6.1 is reported in Appendix 3.

Sample Collection In the action-based setting (AB-MBPExPI), we sample n trajectories $\{\tau_l\}_{l\in[n]}$ independently with the policy $\pi_{\overline{\boldsymbol{\theta}}_{i,j}}$ and we build the dataset $\mathcal{D}_{i,j} = \{(\tau_l, \mathcal{R}(\tau_l))\}_{l\in[n]}$. Instead, in the parameter-based setting (PB-MBPExPI), we sample independently n policy parameters $\{\boldsymbol{\theta}_l\}_{l\in[n]}$ and for each of them we run policy $\pi_{\boldsymbol{\theta}_l}$ once to generate trajectory τ_l. The corresponding dataset is given by $\mathcal{D}_{i,j} = \{((\boldsymbol{\theta}_l, \tau_l), \mathcal{R}(\tau_l))\}_{l\in[n]}$. For the AB case, the correction in Theorem 6.1 is estimated from samples, as done for the Rényi divergence in [Metelli et al., 2018], since it involves integrals between trajectory distributions, while the closed form exists for Gaussian distributions (Appendix 2).

6.2 RESULTS

In this section, we provide the simulation results on continuous control tasks. We first compare the learning performance of MBPExPI with POIS [Metelli et al., 2018] and TRPO [Schulman et al., 2015] on four benchmarks. Then, deepen two relevant aspects of MBPExPI: its robustness to small batch sizes and the effect of applying a monotonic increasing transformation h on function f. All experiments are conducted with Gaussian policies, linear in the state, with fixed variance. The experimental details are reported in Appendix 4. The code to reproduce the presented results is provided at: `https://github.com/albertometelli/uai2023`.

Comparison with POIS and TRPO In Figure 2, we show the average return as a function of the number of collected episodes, with a batch size $n = 100$, using $\alpha = 2$, and one

[4] Clearly, when $\alpha = 1$, the expression does not depend on the behavioral distribution. Thus, for the sake of the algorithm, it makes sense to consider $\alpha > 1$ only.

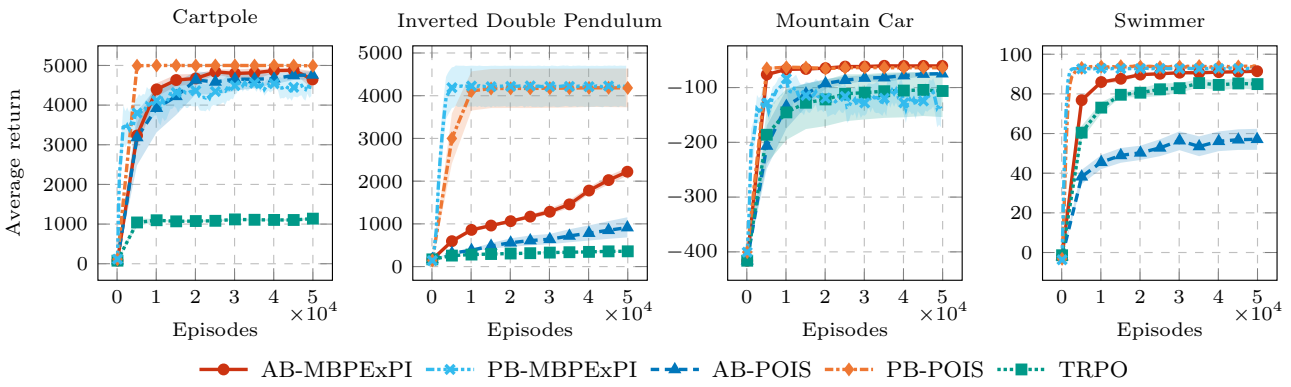

Figure 2: Average return as a function of the number of episodes for different environments and algorithms with batch size $n = 100$, $\alpha = 2$, and $J = 1$ (20 runs $\pm$ 95% bootstrapped c.i.).

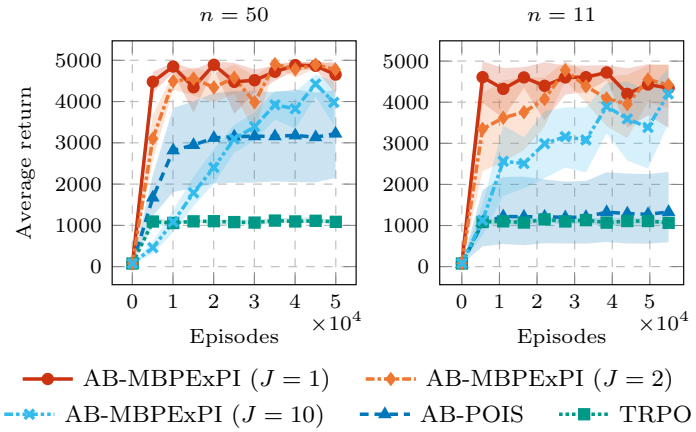

Figure 3: Average return as a function of the number of episodes in the Cartpole environment for different algorithms, batch-size n and inner iterations J (10 runs $\pm$ 95% bootstrapped c.i.).

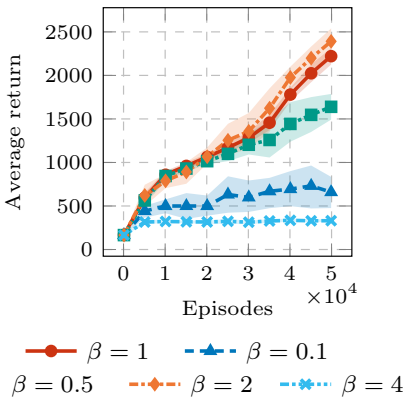

Figure 4: Average return as a function of the number of episodes in the Inverted Double Pendulum for different choices of $h = (\cdot)^\beta$ (5 runs $\pm$ 95% bootstrapped c.i.).

inner iteration ($J = 1$). In the Cartpole environment, we observe that the performance of AB-MBPExPI is slightly above that of AB-POIS and PB-MBPExPI; while the fastest learning curve is shown by PB-POIS. Instead, TRPO converges to a suboptimal policy that fails keeping the pole in the vertical position. In the Inverted Double Pendulum experiment, the gap between AB-MBPExPI and AB-POIS and TRPO is more evident. The PB versions outperform the AB ones with MBPExPI slightly faster than POIS. In the Mountain Car domain, while AB-POIS, TRPO, and PB-MBPExPI display a similar convergence speed, AB-MBPExPI and PB-POIS reach the optimal performance faster. Finally, in the Mujoco Swimmer domain [Todorov et al., 2012], AB-MBPExPI and TRPO clearly outperform AB-POIS, although the fastest learning curves are displayed by the PB versions of POIS and MBPExPI.

Robustness to Small Batch Sizes Based on the previous results, we further investigate the properties of MBPExPI in terms of variance control. In the Cartpole domain, we test the robustness to the reduction of the batch size. In Figure 3, we show the average return as a function of the number of collected episodes for batch sizes $n \in \{11, 50\}$ and different number of inner iterations J. Also considering the $n = 100$ case (Figure 2), we notice, as expected, that the variance of each setting increases overall as n decreases. Nevertheless, MBPExPI proves to be robust, always succeeding in reaching the optimal performance. Differently, POIS suffers the reduced batch size, while TRPO always converging to the same suboptimal policy. The desirable behavior of MBPExPI is indeed an effect of the kind of objective function we employ that explicitly accounts for the variance of the estimator, trying to minimize it, and, as we have shown in the previous sections, it allows enforcing an implicit trust region. Finally, a small number of inner iterations J is beneficial for the stability.

Effect of the Function h We now investigate the effects of using a transformation function $h = (\cdot)^\beta$. Thus, instead of optimizing the expected return, we will optimize the β-power of the expected return. In Figure 4, we show the learning curves of the Inverted Double Pendulum for dif-

ferent values of β. We notice that for β close to 1 (0.5, 1, 2) the curves are not very dissimilar, while for too extreme powers (0.1 and 4) the learning performance degrades. This example shows an interesting phenomenon, i.e., even if we optimize a power of return, within certain limits, we are still able to converge to a (near-)optimal policy.

7 DISCUSSION AND CONCLUSIONS

In this paper, we have studied the relation between policy improvement and off-policy minimum-variance policy evaluation. Specifically, we imported the role of IS as a variance reduction active tool, typical of the Monte Carlo simulation, to the off-policy learning setting. We have illustrated that by minimizing the absolute central α-moment of the IS estimator yields a performance improvement guaranteed on a power of the original objective function, i.e., the expected return in RL. Although the performance improvement is ensured for the case of $\alpha = 1$ only, we have empirically illustrated that even considering $\alpha > 1$, especially $\alpha = 2$ (i.e., minimizing the variance), delivers remarkable learning curves. This phenomenon is justified by the fact that minimizing the variance of IS estimator, as proved theoretically, naturally induces a trust region, mitigating the need for an explicit penalization or constraint. Thus, the bias due to the fact that we are not providing a performance improvement for the expected return (but just for the expected α-power of the return) is compensated by the reduced variance and enforced trust region. Furthermore, this method has proved to be remarkably robust to the reduction of the batch size. We believe that this work contributes to shed light on an appealing facet of off-policy learning with possible new research opportunities. Future works include an extension of the convergence analysis to the sample-based setting and an experimentation of with more complex policy architectures. Specifically, an interesting direction is to investigate the application to *actor-critic* architectures.

Acknowledgements

This paper is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

David Ackley. *A connectionist machine for genetic hill-climbing*, volume 28. Springer Science & Business Media, 2012.

Stéphane Boucheron, Gábor Lugosi, Pascal Massart, et al. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

Konstantinos I. Chatzilygeroudis, Vassilis Vassiliades, Freek Stulp, Sylvain Calinon, and Jean-Baptiste Mouret. A survey on policy search algorithms for learning robot controllers in a handful of trials. *IEEE Trans. Robotics*, 36(2):328–347, 2020. doi: 10.1109/TRO.2019.2958211.

Kamil Andrzej Ciosek and Shimon Whiteson. OFFER: off-environment reinforcement learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 1819–1825. AAAI Press, 2017.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Carles M Cuadras. On the covariance between functions. *Journal of Multivariate Analysis*, 81(1):19–27, 2002.

Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. A survey on policy search for robotics. *Found. Trends Robotics*, 2(1-2):1–142, 2013. doi: 10.1561/2300000021.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1329–1338. JMLR.org, 2016.

Jordan Frank, Shie Mannor, and Doina Precup. Reinforcement learning in the presence of rare events. In *Proceedings of the Twenty-Fifth International Conference (ICML)*, volume 307 of *ACM International Conference Proceeding Series*, pages 336–343. ACM, 2008. doi: 10.1145/1390156.1390199.

Dibya Ghosh, Marlos C. Machado, and Nicolas Le Roux. An operator view of policy gradient methods. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Shixiang Gu, Ethan Holly, Timothy P. Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3389–3396. IEEE, 2017. doi: 10.1109/ICRA.2017.7989385.

John Hammersley. *Monte carlo methods*. Springer Science & Business Media, 2013.

Josiah P. Hanna and Peter Stone. Towards a data efficient off-policy policy gradient. In *2018 AAAI Spring Symposia*. AAAI Press, 2018.

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)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1394–1403. PMLR, 2017.

Josiah Paul Hanna et al. *Data efficient reinforcement learning with off-policy and simulated data*. PhD thesis, 2019.

Timothy Classen Hesterberg. *Advances in importance sampling*. PhD thesis, Citeseer, 1988.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Oper. Res.*, 1(5):263–278, 1953. doi: 10.1287/opre.1.5.263.

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

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Machine Learning, Proceedings of the Nineteenth International Conference (ICML)*, pages 267–274. Morgan Kaufmann, 2002.

Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2619–2624. IEEE, 2004. doi: 10.1109/ROBOT.2004.1307456.

Ilja Kuzborskij, Claire Vernade, András György, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. 130:640–648, 2021.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations (ICLR)*, 2016.

Pierre Liotet, Francesco Vidaich, Alberto Maria Metelli, and Marcello Restelli. Lifelong hyper-policy optimization with multiple importance sampling regularization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, pages 7525–7533. AAAI Press, 2022.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 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, Matteo Papini, Pierluca D'Oro, and Marcello Restelli. Policy optimization as online learning with mediator feedback. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 8958–8966. AAAI Press, 2021a.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 8119–8132, 2021b.

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 *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 4023–4032. PMLR, 2018.

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999. PMLR, 2019.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21 (4):682–697, 2008. doi: 10.1016/j.neunet.2008.02.003.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 307–315. JMLR.org, 2013.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 759–766. Morgan Kaufmann, 2000.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994. ISBN 978-0-47161977-2. doi: 10.1002/9780470316887.

Alfréd Rényi. On measures of entropy and information. Technical report, Hungarian Academy of Sciences Budapest Hungary, 1961.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015.

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 (AAAI)*, pages 3000–3006. AAAI Press, 2015.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Trans. Inf. Theory*, 60 (7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

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)*, pages 419–428. ACM, 1995.