Policy Optimization via Optimal Policy Evaluation

Alberto Maria Metelli* **Samuele Meta*** **Marcello Restelli***

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 as a what-if analysis tool, 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, showing the connections between the two objectives. We illustrate that variance minimization can be used as a performance improvement tool, with the advantage, compared with direct off-policy learning, of implicitly enforcing a trust region. We make use of these theoretical findings to build a PO algorithm, Policy Optimization via Optimal Policy Evaluation (PO^{2}PE), that employs variance minimization as an inner loop. Finally, we present empirical evaluations on continuous RL benchmarks, with a particular focus on the robustness to small batch sizes.

1 Introduction

Policy Optimization methods [PO, 7] have been widely exploited in Reinforcement Learning [RL, 39] with successful results in addressing, to name a few, continuous-control [e.g., 33, 24], robot manipulation [e.g., 12, 3], and locomotion [e.g., 22, 9]. Most of these algorithms employ the notion of *trust region* [5], introduced ante litteram in the RL literature by the *safe* RL approaches [21, 34], giving rise to a surge of effective algorithms, having TRPO [38] 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 [39]. 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 [35] 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, 30] and the phenomenon is particularly evident looking at the IS estimator variance, which grows exponentially with the Rényi divergence [37] between the behavioral and the target policy [27, 28]. In this off-policy learning (Off-PL) setting, IS is employed as a *what-if* analysis tool [30] 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, having the goal of controlling the uncertainty injected by the IS procedure.

However, IS originated in the Monte Carlo simulation community [17, 13] as an *active* tool for *variance minimization* (Off-VM). While in Off-PL, 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-VM, 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 [13, 19]. It might seem surprising, at first, that sampling from a policy, other than the target one, can lead to an estimator with less variance

*Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano - Milan - Italy. Corresponding author: Alberto Maria Metelli, email: `albertomaria.metelli@polimi.it`

Deep Reinforcement Learning Workshop 2021 @ NeurIPS 2021.

(even zero in some cases) w.r.t. the on-policy estimate. In this role, IS has been previously employed in RL, mainly to address rare events [10, 4] 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 [15] for evaluation and subsequently combined with policy gradient learning [14, 16]. However, in these works, the variance minimization (Off-VM) process and the off-policy learning (Off-PL) problem are treated separately.

The goal of this paper is to investigate the relation between variance minimization (Off-VM) and off-policy learning (Off-PL). The core question we address can be summarized as: "*Can Off-VM be employed as a tool for Off-PL, overcoming the need for an explicit trust region?*" 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. This provides a first hint about the connection between the minimum-variance sampling policy and the performance improvement, i.e., between Off-VM and Off-PL. Furthermore, it suggests that we could repeatedly apply the process of identifying the minimum-variance policy as a tool for policy improvement. The interesting aspect of such an approach is that, by minimizing the variance, it *implicitly* controls the divergence between two consecutive policies. In other words, it allows enforcing a trust region, without an explicit need for divergence constraints or penalizations.

Outline of the Contributions In this paper, we provide theoretical, algorithmic, and experimental contributions. After having introduced the necessary background (Section 2), we present the problem of finding the minimum-variance behavioral distribution (Section 3). Then, we study the properties of the Off-VM problem in two settings: unconstrained (Section 4) and constrained (Section 5). First, we assume that there are no restrictions in the choice of the behavioral distribution. We show that the minimum-variance behavioral distribution, besides leading to the well-known zero-variance estimator [19], is guaranteed to yield a performance 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 available distributions are constrained in a suitable space. In this setting, the zero-variance estimator could not be achievable. Nevertheless, we prove that such a procedure can lead to a performance improvement and preserves the ability to enforce a trust region. Based on these theoretical results, we propose *Policy Optimization via Optimal Policy Evaluation* (PO^{2}PE), a novel PO algorithm, that we particularize for parametric policy spaces (Section 6). Finally, we provide numerical simulations on continuous-control benchmarks, in comparison with POIS [27] and TRPO [38], with a particular focus on the robustness of PO^{2}PE to small batch sizes (Section 7). The proof of the results presented in the main paper are reported in Appendix A.

2 Preliminaries

In this section, we report 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 [37]: $D_\alpha(P\|Q)=\frac{1}{\alpha-1}\log\int_{\mathcal{X}}p(x)^\alpha q(x)^{1-\alpha}\mathrm{d}x$. In the limit of $\alpha\to1$, the Rényi divergence reduces to the KL-divergence $D_{\mathrm{KL}}(P\|Q)$, while for $\alpha\to\infty$, it reduces to $\operatorname{ess\,sup}_{x\sim Q}\{p(x)/q(x)\}$.

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, 30] 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, leading to the estimator:

$$\widehat{\mu}_{P/Q}=\frac{1}{n}\sum_{i\in[n]}\frac{p(x_i)}{q(x_i)}f(x_i).$$

The IS estimator is well-known to be unbiased [30], 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 [27]. The properties of $\widehat{\mu}_{P/Q}$ and several of its transformations have been extensively studied in the literature [e.g., 18, 40, 32, 23, 28, 26, 29].

Policy Optimization A Markov Decision Process [MDP, 36] 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, $\mathcal{P}:\mathcal{S}\times\mathcal{A}\to\mathscr{P}(\mathcal{S})$ 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* $\tau=(s_0,a_0,s_1,a_1,\dots,s_{H-1},a_{H-1},s_H)$ where $H\in\mathbb{N}$ is the trajectory length and $s_0\sim D_0$, $a_t\sim\pi_{\boldsymbol{\theta}}(\cdot|s_t)$, $s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)$ for all $t\in\{0,\dots,H-1\}$. 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: $p(\tau|\boldsymbol{\theta})=D_0(s_0)\prod_{t=0}^{H-1}\pi_{\boldsymbol{\theta}}(a_t|s_t)\mathcal{P}(s_{t+1}|s_t,a_t)$. An agent aims at finding a parametrization maximizing the expected return $J(\boldsymbol{\theta})$ [7]:

$$\boldsymbol{\theta}^*\in\operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta}\{J(\boldsymbol{\theta})\}\qquad\text{where}\qquad J(\boldsymbol{\theta})=\mathop{\mathbb{E}}_{\tau\sim p(\cdot|\boldsymbol{\theta})}[\mathcal{R}(\tau)].$$

In the remainder of the paper, we will 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 Off-VM, i.e., 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 $f\in\mathscr{B}(\mathcal{X},[0,\infty))$.[2] Furthermore, we do not enforce any restrictions 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 [20, 19]:

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

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 distribution P [30]:

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

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 (Off-VM) and the problem of finding a target distribution that maximizes the expectation $\mathbb{E}_{x\sim P}[f(x)]$ (Off-PL). In other words, we could employ Off-VM as a performance improvement tool, by repeatedly solving the problem in Equation (1).

In the following two sections, we will delve into the properties of the repeated construction of the minimum-variance distribution as a performance improvement tool 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). In both cases, we will address the following three questions:

- **(Q1)** Does this procedure always generate a distribution that is a *performance improvement*?
- **(Q2)** Does this procedure *converge* to a (global or local) maximum of f?
- **(Q3)** Can we quantify the divergence between two consecutive distributions, i.e., does this procedure enforce a *trust region*?

4 Unconstrained Probability Distribution Space

In Section 3, we have seen that Q^* is a performance improvement w.r.t. P. We now generalize of this construction, by composing function f with a non-negative monotonic strictly-increasing function

[2] 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.

$h:[0,\infty)\to[0,\infty)$. The rationale behind this choice is that if h is strictly-increasing, then $h\circ f$ has the same maxima as f.[3] We start defining the operator $\mathcal{I}_{h\circ f}:\mathscr{P}(\mathcal{X})\to\mathscr{P}(\mathcal{X})$:

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

Thus, $\mathcal{I}_{h\circ f}$ takes as input a target distribution $P\in\mathscr{P}(\mathcal{X})$, a function $h\circ f\in\mathscr{B}(\mathcal{X},[0,\infty))$, and outputs the minimum-variance behavioral distribution for the IS estimation of $\mathbb{E}_{x\sim P}[h(f(x))]$, i.e., $Q^*=\mathcal{I}_{h\circ f}[P]$. Intuitively, looking at Equation (3), by iterating the application of $\mathcal{I}_{h\circ f}$, we will obtain distributions tending to assign larger probability mass to points $x\in\mathcal{X}$ with high values of $f(x)$. Concerning **(Q1)**, the following result, due to [11], generalizes Equation (2) showing that whenever h is increasing, we can prove that $\mathcal{I}_{h\circ f}[P]$ is a performance improvement w.r.t. P.

Proposition 4.1 (Proposition 9 of [11])**.** *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:*

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

It is worth noting that, since h is a monotonic increasing function, we have that $\mathbb{C}\mathrm{ov}_{x\sim P}[h(f(x)),f(x)]\geqslant 0$ [6]. The following sections tackle questions **(Q2)** and **(Q3)**.

4.1 Convergence Properties

We now address question **(Q2)**, analyzing the effect of repeatedly applying operator $\mathcal{I}_{h\circ f}$. More formally, let us consider an initial distribution $P\in\mathscr{P}(\mathcal{X})$, and suppose to iterate the application of the operator $\mathcal{I}_{h\circ f}$, 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 $Q_k=\mathcal{I}_{h\circ f}[Q_{k-1}]=\left(\mathcal{I}_{h\circ f}\right)^k[P]$. The following result shows that, under certain conditions, the operator $\mathcal{I}_{h\circ f}$ admits fixed points and the sequence $(Q_k)_{k\in\mathbb{N}}$ converges to a distribution Q_∞ that assigns probability to the global maxima of f, restricted to the support of P.

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

(i) P is a fixed point of $\mathcal{I}_{h\circ f}$, i.e., $\mathcal{I}_{h\circ f}[P]=P$ 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 $\mathcal{I}_{h\circ f}$ converges to a distribution $Q_\infty=\lim_{k\to\infty}\left(\mathcal{I}_{h\circ f}\right)^k[P]$ with support $\mathcal{X}^*$. In particular:

$$\mathop{\mathbb{E}}_{x\sim Q_\infty}[f(x)]=\max_{x\in\mathrm{supp}(P)}\{f(x)\}.$$

Some remarks are in order. First, all three properties are independent of the function h as long as it is non-negative and monotonically increasing. This is expected since, under this condition, $h\circ f$ admits the same set of global optima of f. Second, as a corollary to point (i), any deterministic P is a fixed point of $\mathcal{I}_{h\circ f}$. Finally, 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 $\mathcal{I}_{h\circ f}$ 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.

4.2 Implicit Trust Region

The reader might wonder what are the advantages of casting the optimization of function f as such an iterative procedure. The reason lies in question **(Q3)**. We now prove that we are able to naturally control the divergence between two consecutive distributions Q_k and $Q_{k+1}=\mathcal{I}_{h\circ f}[Q_k]$, 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 $P\in\mathscr{P}(\mathcal{X})$, $f\in\mathscr{B}(\mathcal{X},[0,\infty))$, and $h:[0,\infty)\to[0,\infty)$ monotonic strictly-increasing. Then, for every $\alpha\in[0,\infty]$, it holds that:*

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

[3] As we shall see in the following sections, the different choices of h will be useful to control the trust region of the optimization process.

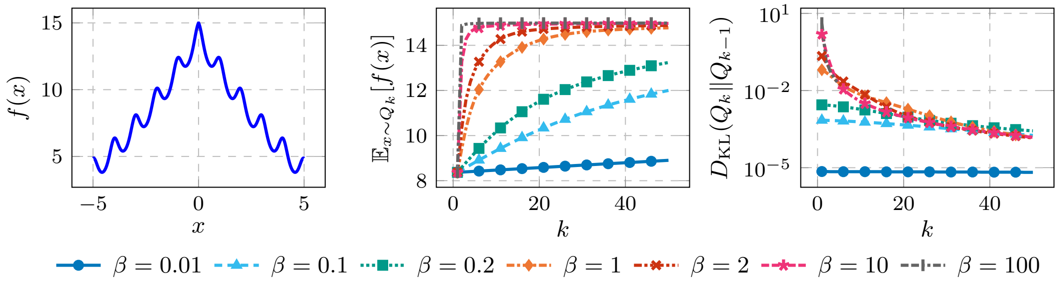

Figure 1: The Ackley function (left), the expectation of the distribution $Q_k = (\mathcal{I}_{h\circ f})^k[P]$ (center), and the KL-divergence (right) between two consecutive distributions Q_{k-1} and Q_k, with $h=(\cdot)^\beta$.

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

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

For $\alpha=2$, we obtain $D_2(I_{h\circ f}[P]\|P) = \log \frac{\mathbb{E}_{x\sim P}[h(f(x))^2]}{\mathbb{E}_{x\sim P}[h(f(x))]^2} \leqslant \frac{\mathbb{V}\mathrm{ar}_{x\sim P}[h(f(x))]}{\mathbb{E}_{x\sim P}[h(f(x))]^2}$. Thus, the divergence is large when the variance of $h(f(x))$ 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 [38], are unable to control higher-order divergences that can still be infinite. We can also appreciate the role of the increasing function h that works as a regularizer with the effect of controlling the width of the trust region. The following example shows that the faster h increases, the larger the induced trust region becomes.

Example 4.1. *We consider (a slight variation of) the one-dimensional Ackley function [1]: $f(x) = -5+20\exp(-0.1414|x|)+\exp(0.5(\cos(2\pi x)+1))+e$, 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 which, at the same time, lead to larger trust regions.*

5 Constrained Probability Distribution Space

The approach we have presented in Section 4 can be effectively 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 $\mathscr{P}(\mathcal{X})$. This is for instance the case of multi-armed bandit problems where any distribution over the arms can be played, but not the case of MDPs in which trajectory distributions are governed by the transition model and are, naturally, constrained. More formally, when consider a class of distributions $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$, even if $P \in \mathcal{Q}$, the distribution $\mathcal{I}_{h\circ f}[P]$ might not belong to $\mathcal{Q}$. Furthermore, while $\mathcal{I}_{h\circ f}[P]$ 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:

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

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 $P \in \mathscr{P}(\mathcal{X})$, $f \in \mathscr{B}(\mathcal{X},[0,\infty))$, and $h:[0,\infty)\to[0,\infty)$ monotonic strictly-increasing. Then, for any $\alpha \in (1,\infty)$, it holds that:*

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

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

$$Q^\dagger \in \arg\min_{Q\in\mathcal{Q}} \{D_\alpha(\mathcal{I}_{h\circ f}[P] \| Q)\}\,. \tag{5}$$

In the following sections, we shall address the questions **(Q1)**, **(Q2)**, and **(Q3)**.

5.1 Performance Improvement

In Proposition 4.1, we have seen that, whenever h is strictly-increasing, $\mathcal{I}_{h\circ f}[P]$ is a performance improvement w.r.t. P, evaluated under function f (and also under the composition between f and *any* strictly-increasing function). In this section, we address question **(Q1)**, showing that, when considering a subset of distributions $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$, the performance improvement cannot be in general guaranteed for f, but just for a *specific* monotonic transformation of f, depending on h and α.

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

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

In particular, for $\alpha = 1$, it holds that [11, Proposition 6]:

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

The result shows that by minimizing the α-moment of the transformed function $h\circ f$, we are able to guarantee a performance improvement on the function $(\cdot)^\alpha \circ h \circ f$. The result holds provided that $D_\alpha(\mathcal{I}_{h\circ f}[P]\|Q) \leqslant D_\alpha(\mathcal{I}_{h\circ f}[P]\|P)$, 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. In particular, if we select $h = (\cdot)^{1/\alpha}$, the guarantee holds for the function f directly. For all other choices, the performance improvement can be guaranteed for a monotonic transformation of f only.[4]

5.2 Convergence Properties

We now turn to **(Q2)**. By using Equation (5) as an iterate $Q_{k+1} \in \arg\min_{Q\in\mathcal{Q}}\{D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q)\}$ 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 $P \in \mathscr{P}(\mathcal{X})$, $f \in \mathscr{B}(\mathcal{X}, [0,\infty))$, and $h : [0,\infty) \to [0,\infty)$ monotonic strictly-increasing. Let $\mathcal{Q} \subseteq \mathscr{P}(\mathcal{X})$ and suppose that $h\circ f$ is bounded from above, then, the iterate $Q_{k+1} \in \arg\min_{Q\in\mathcal{Q}}\{D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q)\}$ (where possible ties are broken arbitrarily) satisfies:*

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

The convergence of the sequences $D_\alpha(\mathcal{I}_{h\circ f}[Q_k]\|Q_k)$ and $\mathbb{E}_{x\sim Q_k}[h(f(x))^\alpha]$ is derived by the performance improvement result of Theorem 5.2. The important point of Theorem 4.2 is that we achieve convergence to a *stationary point* of $\mathbb{E}_{x\sim Q}[h(f(x))^\alpha]$. If $\mathcal{Q}$ is a parametric space $\mathcal{Q}_\Theta = \{Q_{\boldsymbol{\theta}} \in \mathscr{P}(\mathcal{X}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$, then we are guaranteed to stop when $\mathbb{E}_{x\sim Q_{\boldsymbol{\theta}}}[\nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(x) h(f(x))^\alpha] = 0$, like for a general policy gradient [31] method maximizing $h(f(x))^\alpha$. 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 $Q_\infty = \arg\min_{Q\in\mathcal{Q}}\{D_\alpha(\mathcal{I}_{h\circ f}[Q]\|Q)\}$.

5.3 Implicit Trust Region

In Theorem 4.3, we have proved that the α-Rényi divergence between $\mathcal{I}_{h\circ f}[P]$ and P is bounded. In this section, we answer **(Q3)**, wondering whether similar properties hold when we consider a

[4] In Appendix B, we discuss the effects of optimizing a power of f instead of f itself, i.e., when $h = (\cdot)^\beta$.

limited set of distributions $\mathcal{Q}\subseteq\mathscr{P}(\mathcal{X})$. The following result shows that, under a particular form of convexity [42] of $\mathcal{Q}$, we are able to control the trust region as well.

Theorem 5.4. *Let $f\in\mathscr{B}(\mathcal{X},[0,\infty))$, and $h:[0,\infty)\to[0,\infty)$ monotonic strictly-increasing. Let $\mathcal{Q}\subseteq\mathscr{P}(\mathcal{X})$ be a $(1-\alpha)$-convex set [42, Definition 4], $P\in\mathcal{Q}$, $Q^\dagger\in\arg\min_{Q\in\mathcal{Q}}\{D_\alpha(\mathcal{I}_{h\circ f}[P]\|Q)\}$, and $\alpha\in[0,\infty]$, then it holds that:*

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

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

6 Policy Optimization via Optimal Policy Evaluation

In the previous sections, we have discussed the properties of the distributions that minimize the absolute central α-moments of the IS estimator, when the sampling distributions is chosen without restrictions (Section 4) or within a set of distributions (Section 5). In this section, we employ these results to build a sample-based Off-PL algorithm, which uses Off-VM as an inner loop. The pseudocode of the algorithm, named *Policy Optimization via Optimal Policy Evaluation* (PO^{2}PE), is reported in Algorithm 1. For generality of presentation, we consider a parametric distribution space $\mathcal{Q}_\Theta=\{Q_{\boldsymbol{\theta}}\in\mathscr{P}(\mathcal{X}):\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d\}$, that is a common setting encountered in PO.

The basic structure of PO^{2}PE consists of two nested loops. Given a target distribution $q_{\boldsymbol{\theta}_i}$, the inner loop aims at performing the `Evaluation` of the performance of $q_{\boldsymbol{\theta}_i}$. At each inner iteration $j\in[J]$, it collects samples $\mathcal{D}_{i,j}$ with the current behavioral distribution $q_{\overline{\boldsymbol{\theta}}_{i,j}}$ and employs them, together with all the samples collected so far $(\mathcal{D}_{i,k})_{k\in[j]}$, to compute the next behavioral distribution $q_{\overline{\boldsymbol{\theta}}_{i,j+1}}$, with the goal of minimizing the absolute central α-moment. This process is governed by two hyperparameters: h the transformation function and α the moment order. The outer loop, instead, aims to perform the `Optimization` of the target distribution $q_{\boldsymbol{\theta}_i}$. At the end of each outer iteration $i\in[I]$, the target distribution $q_{\boldsymbol{\theta}_{i+1}}$ is updated with the last behavioral distribution produced by the inner loop $q_{\overline{\boldsymbol{\theta}}_{i,J+1}}$. To get a usable algorithm, we need to further characterize how the samples are collected (Line 4), particularizing for the PO setting, and how to perform the optimization from samples (Line 5).

Algorithm 1: PO^{2}PE.

```
input  : $\alpha$ divergence order, $h$ function, $f$ function, $\mathcal{Q}_\Theta$
         distribution space, $\boldsymbol{\theta}_1\in\Theta$ initial parameter, $n$ batch size
output : final parameter $\boldsymbol{\theta}_{I+1}\in\Theta$
1 for $i=1,\dots,I$ do                                              Optimization
2     $\overline{\boldsymbol{\theta}}_{i,1}=\boldsymbol{\theta}_i$
3     for $j=1,\dots,J$ do                                          Evaluation
4         Collect $n$ samples $\mathcal{D}_{i,j}=\{(x_l,f(x_l))\}_{l\in[n]}$ with $Q_{\overline{\boldsymbol{\theta}}_{i,j}}$
5         Find $\overline{\boldsymbol{\theta}}_{i,j+1}$ by minimizing $D_\alpha\left(\mathcal{I}_{h\circ f}[Q_{\boldsymbol{\theta}_i}]\|Q_{\boldsymbol{\theta}}\right)$ using
            $(\mathcal{D}_{i,k})_{k\in[j]}$
6     end
7     $\boldsymbol{\theta}_{i+1}=\overline{\boldsymbol{\theta}}_{i,J+1}$
8 end
```

Sample-based Optimization The problem of finding the next behavioral distribution parameter $\overline{\boldsymbol{\theta}}_{i,j+1}$ using the samples collected so far $\left(\mathcal{D}_{i,k}\right)_{k\in[j]}$ is in all regards an off-policy learning problem. Let us define $\Phi_{i,j}=\frac{1}{j}\sum_{k\in[j]}q_{\overline{\boldsymbol{\theta}}_{i,k}}$ as the mixture of the j behavioral distributions experienced so far in the inner loop. Instead of directly estimating $D_\alpha\left(\mathcal{I}_{h\circ f}[Q_{\boldsymbol{\theta}_i}]\|Q_{\boldsymbol{\theta}}\right)$, we refer to the (non-central) α-moment, which is connected to the original objective through Proposition 5.1. Since we have samples coming from different behavioral distributions, we can use a *multiple* IS estimator [43]:

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

The (a) factor takes into account that we are using samples collected with the mixture $\Phi_{i,j}$ to estimate an expectation under $q_{\boldsymbol{\theta}}$, whereas the factor (b) 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 [32]. To perform the minimization of Equation (6), we employ a variance correction to mitigate the effect of finite samples [27], theoretically grounded in the following result.

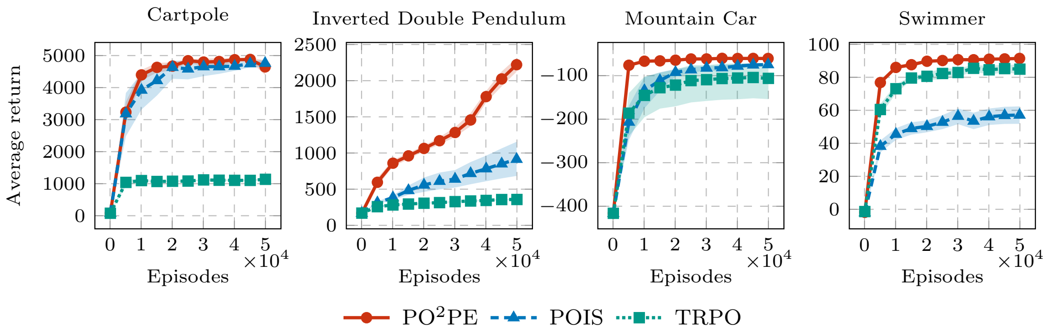

Figure 2: Average return as a function of the number of episodes for different environments and algorithms with batch size $n=100$, $\alpha=2$, $h=\mathrm{Id}$, and $J=1$ (20 runs ± 95% bootstrapped c.i.).

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

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

Some remarks are in order. First, the integral within the square root is an upper bound to the variance of the α-moment estimator $\widehat{d}_{\alpha}(\mathcal{I}_{h\circ f}[Q_{\boldsymbol{\theta}_i}]\|Q_{\boldsymbol{\theta}};\Phi_{i,j})$. In particular, when $\boldsymbol{\theta}=\boldsymbol{\theta}_i$, we obtain the exponentiated Rényi divergence, as illustrated in [28]. When all involved distributions are Guassians, it is possible to provide a closed-form tight bound on this quantity (Appendix C). Second, unlike the results available in the literature about concentration of IS estimator, without correction or transformation, we are able to provide an exponential concentration inequality (dependence on delta 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 [2].

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 7, the need for a penalization or constraint is significantly less relevant than in existing approaches, like TRPO [38], or POIS [27].

Sample Collection The sample collection (Line 4) depend on the kind of problem we are dealing with. Specifically, for the PO setting, $q_{\boldsymbol{\theta}}=p(\cdot|\boldsymbol{\theta})$ is the trajectory distribution induced by policy $\pi_{\boldsymbol{\theta}}$, and function f corresponds to the trajectory return $\mathcal{R}(\tau)$. At each inner iteration $j\in[J]$, we sample n trajectories $\{\tau_l\}_{l\in[n]}$ independently with the policy $\pi_{\overline{\boldsymbol{\theta}}_{i,j}}$ and we build the dataset $\mathcal{D}_{i,j}=\{(\tau_l,\mathcal{R}(\tau_l))\}_{l\in[n]}$. The correction term in Theorem 6.1 has to be estimated from samples as well, as done for the Rényi divergence in [27], since it involves integrals between trajectory distributions.

7 Experimental Evaluation

In this section, we provide the experimental evaluation of PO^{2}PE on continuous control tasks. We first compare the learning performance of PO^{2}PE with POIS [27] and TRPO [38] on four benchmarks. Then, we dive into two relevant aspects of PO^{2}PE: its robustness to small batch sizes and the effect of the transformation function h. All experiments are conducted with Gaussian policies, linear in the state variables, with fixed variance. The experimental details are reported in Appendix D.

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 $\alpha=2$, $h=\mathrm{Id}$, and one inner iteration ($J=1$). In the Cartpole environment, we observe that the performance of PO^{2}PE is slightly above that of 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 PO^{2}PE and the baselines is more evident, whereas in the Mountain Car domain, while POIS and TRPO display a similar convergence speed, PO^{2}PE reaches the optimal performance faster. Finally, in the Mujoco

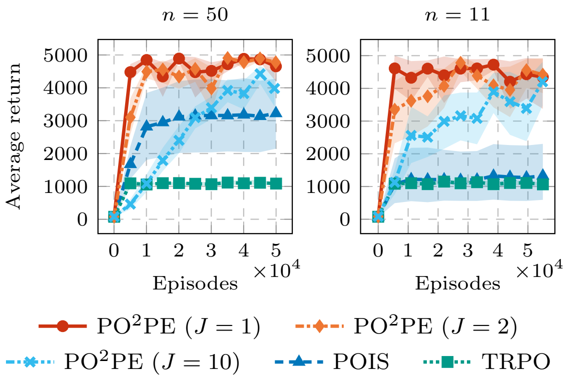

Figure 3: Average return as a function of the number of episodes in the Cartpole environment for different algorithms, batch-size n and inner iterations J (10 runs ± 95% bootstrapped c.i.).

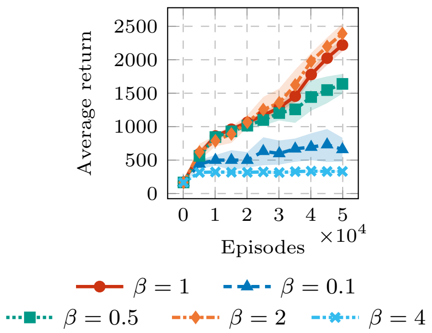

Figure 4: Average return as a function of the number of episodes in the Inverted Double Pendulum for different choices of $h = (\cdot)^\beta$ (5 runs ± 95% bootstrapped c.i.).

Swimmer domain [41], PO^{2}PE and TRPO clearly outperform POIS. In all three experiments, we appreciate the small variance of PO^{2}PE across the different runs.

Robustness to Small Batch Sizes Based on the previous results, we further investigate the properties of PO^{2}PE 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, PO^{2}PE 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 PO^{2}PE 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. Concerning the number of inner iterations J, although all considered cases approach the optimal performance, a small number of inner iterations seem to be beneficial for the stability.

Effect of the Function h While previous experiments we consider h to be the identity function, we now investigate the effects of using $h = (\cdot)^\beta$, i.e., a power function. In Figure 4, we show the learning curves of the Inverted Double Pendulum for different 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.

8 Discussion and Conclusions

In this paper, we have deepened the study of importance sampling beyond its usage as a passive tool for off-policy evaluation and learning. We imported the role of IS as a variance reduction active tool, typical of the Monte Carlo simulation field, to the off-policy learning setting. We have illustrated that by minimizing the absolute central α-moment of the IS estimator we are able to guarantee the performance improvement for a monotonic transformation of the original objective function and eventually converge, at least, to a stationary point. Interestingly, this approach is able to naturally induce a trust region, mitigating the need for an explicit penalization or constraint. The experimental evaluation confirmed our theoretical findings. PO^{2}PE is able to outperform POIS and TRPO on several continuous control tasks. Remarkably, our algorithm has proved to be robust to the reduction of the batch size and this represents a beneficial effect of the implicit trust region enforcement. 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 case in which samples are involved and an experimentation of PO^{2}PE coupled with more complex policy architectures.

References

[1] David Ackley. *A connectionist machine for genetic hillclimbing*, volume 28. Springer Science & Business Media, 2012.

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

[3] 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.

[4] Kamil Andrzej Ciosek and Shimon Whiteson. OFFER: off-environment reinforcement learning. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1819–1825. AAAI Press, 2017.

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

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

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

[8] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. `https://github.com/openai/baselines`, 2017.

[9] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1329–1338. JMLR.org, 2016.

[10] Jordan Frank, Shie Mannor, and Doina Precup. Reinforcement learning in the presence of rare events. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 336–343. ACM, 2008.

[11] Dibya Ghosh, Marlos C. Machado, and Nicolas Le Roux. An operator view of policy gradient methods. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[12] 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 2017, Singapore, Singapore, May 29 - June 3, 2017*, pages 3389–3396. IEEE, 2017.

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

[14] Josiah P. Hanna and Peter Stone. Towards a data efficient off-policy policy gradient. In *2018 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 26-28, 2018*. AAAI Press, 2018.

[15] Josiah P. Hanna, Philip S. Thomas, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. In Doina Precup and Yee Whye Teh, editors, *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*, pages 1394–1403. PMLR, 2017.

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

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

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

[19] H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Oper. Res.*, 1(5):263–278, 1953.

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

[21] Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In Claude Sammut and Achim G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pages 267–274. Morgan Kaufmann, 2002.

[22] 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 2004, April 26 - May 1, 2004, New Orleans, LA, USA*, pages 2619–2624. IEEE, 2004.

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

[24] 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 Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[25] Andreas Maurer et al. A bound on the deviation probability for sums of non-negative random variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1):15, 2003.

[26] 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 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 8958–8966. AAAI Press, 2021.

[27] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5447–5459, 2018.

[28] 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.

[29] Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian importance sampling for off-policy evaluation and learning. *ICML-21 Workshop on Reinforcement Learning Theory*, 2021.

[30] Art B Owen. Monte carlo theory, methods and examples, 2013.

[31] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4023–4032. PMLR, 2018.

[32] Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999. PMLR, 2019.

[33] Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008.

[34] Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 307–315. JMLR.org, 2013.

[35] Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In Pat Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 759–766. Morgan Kaufmann, 2000.

[36] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

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

[38] John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015.

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

[40] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3000–3006. AAAI Press, 2015.

[41] 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 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pages 5026–5033. IEEE, 2012.

[42] Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014.

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