# Policy Optimization as Online Learning with Mediator Feedback

**Alberto Maria Metelli*, Matteo Papini*, Pierluca D'Oro, Marcello Restelli**
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Piazza Leonardo da Vinci, 32, 20133, Milano, Italy
{albertomaria.metelli, matteo.papini, marcello.restelli}@polimi.it, pierluca.doro@mail.polimi.it

## Abstract

Policy Optimization (PO) is a widely used approach to address continuous control tasks. In this paper, we introduce the notion of *mediator feedback* that frames PO as an online learning problem over the policy space. The additional available information, compared to the standard bandit feedback, allows reusing samples generated by one policy to estimate the performance of other policies. Based on this observation, we propose an algorithm, *RANDomized-exploration policy Optimization via Multiple Importance Sampling with Truncation* (RANDOMIST), for regret minimization in PO, that employs a randomized exploration strategy, differently from the existing optimistic approaches. When the policy space is finite, we show that under certain circumstances, it is possible to achieve constant regret, while always enjoying logarithmic regret. We also derive problem-dependent regret lower bounds. Then, we extend RANDOMIST to compact policy spaces. Finally, we provide numerical simulations on finite and compact policy spaces, in comparison with PO and bandit baselines.

## 1 Introduction

Policy Optimization (PO, Deisenroth, Neumann, and Peters 2013) is a family of Reinforcement Learning (RL, Sutton and Barto 2018) algorithms based on the explicit optimization of the policy parameters. It represents the most promising approach for learning large-scale continuous control tasks and has already achieved marvelous results in video games (e.g., Vinyals et al. 2019) and robotics (e.g., Peng et al. 2020). These achievements, however, rely on massive amounts of simulation rollouts. The efficient use of experience data is essential both to reduce computational costs and to make learning online from real interaction possible. This is still largely an open problem and calls for better theoretical understanding. Any online-learning agent must face the exploration-exploitation dilemma: whether to leverage on its current knowledge to maximize performance or consider new alternatives. Fortunately, the Multi-Armed Bandit (MAB) literature (Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2018) provides a theoretical framework for the problem of efficient exploration under *bandit feedback*, i.e., observing the effects of the chosen actions. The dilemma is addressed by minimizing the cumulative *regret* of the online performance w.r.t. the optimal one. The most popular exploration strategies are based on the Optimism in the Face of Uncertainty (OFU, Lai and Robbins 1985), of which UCB1 (Auer, Cesa-Bianchi, and Fischer 2002) is the prototypical algorithm, and on Thompson Sampling (TS, Thompson 1933). Both suffer only sublinear regret (Auer, Cesa-Bianchi, and Fischer 2002; Agrawal and Goyal 2012; Kaufmann, Korda, and Munos 2012). TS typically performs better in practice (Chapelle and Li 2011), but it is only computationally efficient in artificial settings (Kveton et al. 2019b). More recent randomized algorithms such as PHE (Perturbed History Exploration) (Kveton et al. 2019a) are able to match the theoretical and practical advantages of TS without the computational burden, and with no assumptions on the payoff distribution.

The OFU principle has been applied to RL (Jaksch, Ortner, and Auer 2010) and recently also to PO (Chowdhury and Gopalan 2019; Efroni et al. 2020), at the level of action selection. These methods are promising but limited to finite actions. A different perspective is proposed by Papini et al. (2019), where the decision problem is not defined over the agent's actions but over the policy parameters. This change of viewpoint allows exploiting the special structure of the PO problem: for each policy, a sequence of states and actions performed by the agent is collected, constituting, alongside the rewards, a vastly richer signal than the simple bandit feedback. In this paper, we call it *mediator feedback* since this extra information acts as a mediator variable between the policy parameters and the return. OPTIMIST (Papini et al. 2019) is an OFU algorithm that uses Multiple Importance Sampling (MIS, Veach and Guibas 1995) to exploit the mediator feedback, so that the results of one policy provide information on all the others. This allows, in principle, to optimize over an infinite policy space with only finite samples and no regularity assumptions on the underlying process. There are two important limitations in Papini et al. (2019). First, the advantages of the mediator feedback over the bandit feedback are not clear from a theoretical perspective since the regret of OPTIMIST is comparable with that of UCB1 with finite policy space. Second, the policy selection of OPTIMIST requires maximizing a non-convex and non-differentiable index. In the continuous setting, this

[*] Equal contribution.

is addressed via discretization, with clear scalability issues.

In this work, we provide two major advancements. From the theoretical side, we provide regret lower bounds for the policy optimization problem with finite policy space, and we show that OPTIMIST actually enjoys *constant* regret under the assumptions made in (Papini et al. 2019). In fact, mediator feedback is so special that, under strong-enough assumptions, a greedy algorithm enjoys the same guarantees. We also devise a PHE-inspired randomized algorithm, called RANDOMIST (RANDomized-exploration policy Optimization via Multiple Importance Sampling with Truncation), with similar regret guarantees as OPTIMIST. From the practical side, this allows replacing the unfeasible index maximization of OPTIMIST with a sampling procedure. Although our regret guarantees apply to the finite setting only, we propose a heuristic version of RANDOMIST for continuous problems, using a Markov Chain Monte Carlo (MCMC, Owen 2013). We show the advantages of this algorithm over continuous OPTIMIST in terms of computational complexity and performance.

The structure of the paper is as follows. We start in Section 2 with the basic background. In Section 3, we formalize the mediator feedback in PO and derive two regret lower bounds. We illustrate, in Section 4, a way to exploit mediator feedback, based on importance sampling. Section 5 is devoted to the discussion of deterministic algorithms, providing the improved regret guarantees for OPTIMIST. In Section 6, we present RANDOMIST with its regret guarantees and the heuristic extension to the continuous case. In Section 7, we compare empirically RANDOMIST with relevant baselines on both illustrative examples and continuous-control problems. In Section 8, we discuss relationships with similar approaches from the bandit and RL literature. We conclude in Section 9, discussing the obtained results and proposing future research directions. The extended paper can be found at https://arxiv.org/abs/2012.08225.

## 2 Preliminaries

In this section, we introduce some notation, the background on Markov decision processes and policy optimization.

**Mathematical Background** Let $(\mathcal{X}, \mathscr{F})$ be a measurable space, we denote with $\mathscr{P}(\mathcal{X})$ the set of probability measures over $\mathcal{X}$. Let $P, Q \in \mathscr{P}(\mathcal{X})$ such that $P \ll Q$,[1] for any $\beta \in [0, \infty]$ the $\beta$-Rényi divergence (Rényi 1961) is defined as:[2]

$$D_\beta(P \| Q) = \frac{1}{\beta - 1} \log \int_{\mathcal{X}} \left( \frac{\mathrm{d}P}{\mathrm{d}Q} \right)^\beta \mathrm{d}Q.$$

We denote with $d_\beta(P\|Q) = \exp\left[D_\beta(P\|Q)\right]$ the exponentiated Rényi divergence (Cortes, Mansour, and Mohri 2010).

**Markov Decision Processes and Policy Optimization** A discrete-time Markov Decision Process (MDP, Puterman 1994) is a 6-tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the transition model that for each $(s, a) \in \mathcal{S} \times \mathcal{A}$ provides the probability distribution of the next state $\mathcal{P}(\cdot|s, a) \in \mathscr{P}(\mathcal{S})$, $\mathcal{R}(s, a) \in \mathbb{R}$ is the reward function, $\gamma \in [0, 1]$ is the discount factor, and $\mu \in \mathscr{P}(\mathcal{S})$ is the initial-state distribution. In Policy Optimization (PO, Peters and Schaal 2008), we model the agent's behavior by means of a policy $\pi_{\boldsymbol{\theta}}(\cdot|s) \in \mathscr{P}(\mathcal{A})$ belonging to a space of parametric policies $\Pi_\Theta = \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. The interaction between an agent and an MDP generates a sequence of state-action pairs, named *trajectory*: $\tau = (s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1})$ where $s_0 \sim \mu$, for all $h \in \{0, \ldots, H-1\}$ we have $a_h \sim \pi_{\boldsymbol{\theta}}(\cdot|s_h)$, $s_{h+1} \sim \mathcal{P}(\cdot|s_h, a_h)$ and $H \in \mathbb{N}$ is the trajectory length. Each parameter $\boldsymbol{\theta} \in \Theta$ determines a policy $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ which, in turn, induces a probability measure $p_{\boldsymbol{\theta}} \in \mathscr{P}(\mathcal{T})$ over the trajectory space $\mathcal{T}$. To every trajectory $\tau \in \mathcal{T}$, we associate an index of performance $\mathcal{R}(\tau) = \sum_{h=0}^{H-1} \gamma^h \mathcal{R}(s_h, a_h)$, called *return*. Without loss of generality we assume that $\mathcal{R}(\tau) \in [0, 1]$. Thus, we can evaluate the performance of a policy $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ by means of its *expected return*: $J(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}}\left[\mathcal{R}(\tau)\right]$. The goal of the agent consists in finding an optimal parameter, i.e., any $\boldsymbol{\theta}^*$ maximizing $J(\boldsymbol{\theta})$.[3]

## 3 Online Policy Optimization and Mediator Feedback

The online PO protocol works as follows. At each round $t \in [n]$, we evaluate a parameter vector $\boldsymbol{\theta}_t \in \Theta$ by running policy $\pi_{\boldsymbol{\theta}_t}$, collecting one (or more) trajectory $\tau_t \in \mathcal{T}$ and observing the corresponding return $\mathcal{R}(\tau_t)$. Then, based on the history $\mathcal{H}_t = \{(\boldsymbol{\theta}_i, \tau_i, \mathcal{R}(\tau_i))\}_{i=1}^t$, we update $\boldsymbol{\theta}_t$ to get $\boldsymbol{\theta}_{t+1}$. From an *online learning* perspective, the goal of the agent consists in maximizing the sum of the expected returns over $n$ rounds or, equivalently, minimizing the cumulative regret $R(n)$:

$$\max_{\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_n \in \Theta} \sum_{t=1}^n J(\boldsymbol{\theta}_t) \;\Leftrightarrow\; \min_{\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_n \in \Theta} R(n) = \sum_{t=1}^n \Delta(\boldsymbol{\theta}_t),$$

where $\Delta(\boldsymbol{\theta}) = J^* - J(\boldsymbol{\theta})$ is the optimality gap of $\boldsymbol{\theta} \in \Theta$ and $J^* = \sup_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta})$. Thus, whenever policy $\pi_{\boldsymbol{\theta}_t}$ is executed the agent receives the trajectory-return pair $(\tau_t, \mathcal{R}(\tau_t))$, that we name *mediator feedback* (MF). The term "mediator" refers to the side information, the trajectory $\tau_t$, that *mediates* between the parameter choice $\boldsymbol{\theta}_t$ and the return $\mathcal{R}(\tau_t)$. By naïvely approaching PO as an online-learning problem over policy space, we would only consider *bandit feedback*, in which just the return $\mathcal{R}(\tau_t)$ is observable. In comparison, the MF allows to better exploit the *structure* underlying the PO problem (Figure 1).[4] Indeed, while the return function $\mathcal{R}$

[1] $P$ is absolutely continuous w.r.t. $Q$, i.e., for every measurable set $\mathcal{Y} \subseteq \mathcal{X}$ we have $Q(\mathcal{Y}) = 0 \Rightarrow P(\mathcal{Y}) = 0$.

[2] In the limit, for $\beta \to 1$ we have $D_1(P\|Q) = D_{\mathrm{KL}}(P\|Q)$ and for $\beta \to \infty$ we have $D_\infty(P\|Q) = \operatorname{ess\,sup}_{\mathcal{X}} \frac{\mathrm{d}P}{\mathrm{d}Q}$.

[3] To simplify the presentation, we frame our results for the usual *action-based* PO. Our findings directly extend to *parameter-based* exploration (Sehnke et al. 2008), in which policies are indirectly optimized by learning a hyperpolicy that outputs the policy parameters. Coherently with Papini et al. (2019), the empirical evaluation of Section 7 is carried out in the parameter-based framework.

[4] In this paper, we employ the wording "bandit feedback" with a different meaning compared to some provably efficient approaches to PO (e.g., Efroni et al. 2020). See also Section 8.

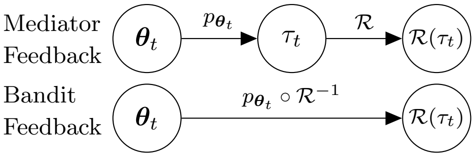

Figure 1: Graphical models comparing mediator and bandit feedbacks.

is unknown, the trajectory distribution $p_{\boldsymbol{\theta}}$ is *partially* known:

$$p_{\boldsymbol{\theta}}(\tau) = \mu(s_0) \prod_{h=0}^{H-1} \pi_{\boldsymbol{\theta}}(a_h|s_h) P(s_{h+1}|s_h, a_h). \quad (1)$$

The policy factors $\pi_{\boldsymbol{\theta}}$, that depend on $\boldsymbol{\theta}$, are known to the agent, whereas the factors due to the environment ($\mu$ and $P$) are unknown but do not depend on $\boldsymbol{\theta}$. Intuitively, if two policies $\pi_{\boldsymbol{\theta}}$ and $\pi_{\boldsymbol{\theta}'}$ are sufficiently "similar", given a trajectory $\tau$ from policy $\pi_{\boldsymbol{\theta}}$, the return $\mathcal{R}(\tau)$ provides information on the expected return of policy $\pi_{\boldsymbol{\theta}'}$ too.

**Regret Lower Bounds for Finite Policy Space** We focus on the intrinsic complexity of PO with finite policy space, deriving two lower bounds to the regret. The results are phrased, for simplicity, for the case of two policies, i.e., $|\Theta| = 2$, and the proof techniques are inspired to (Bubeck, Perchet, and Rigollet 2013). We start showing that, with enough structure between the policies, i.e., when the KL-divergence between the trajectory distributions is bounded, the best achievable regret is constant.

**Theorem 3.1.** *There exist an MDP and a parameter space* $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ *with* $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_1} \| p_{\boldsymbol{\theta}_2}) < \infty$, $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_2} \| p_{\boldsymbol{\theta}_1}) < \infty$ *and* $J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) = \Delta$ *such that, for sufficiently large* $n$*, all algorithms suffer regret* $\mathbb{E}\, R(n) \geqslant \frac{1}{32\Delta}$.

Instead, the presence of policies that are uninformative of one another, i.e., with infinite KL-divergence between the trajectory distributions, leads to a logarithmic regret.

**Theorem 3.2.** *There exist an MDP and a parameter space* $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ *with* $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_1} \| p_{\boldsymbol{\theta}_2}) = \infty$ *or* $D_{\mathrm{KL}}(p_{\boldsymbol{\theta}_2} \| p_{\boldsymbol{\theta}_1}) = \infty$*, and* $J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) = \Delta$ *such that, for any* $n \geqslant 1$*, all algorithms suffer regret* $\mathbb{E}\, R(n) \geqslant \frac{1}{8\Delta} \log(\Delta^2 n)$.

## 4 Exploiting Mediator Feedback with Importance Sampling

In this section, we illustrate how Importance Sampling techniques (IS, Cochran 1977; Owen 2013) can be employed to effectively exploit the mediator feedback in PO.[5]

**Monte Carlo Estimation** With the bandit feedback at each round $t \in [n]$, the agent has access to the history of parameter-return pairs $\mathcal{H}_t = \{(\boldsymbol{\theta}_i, \mathcal{R}(\tau_i)\}_{i=1}^{t-1}$. Let $T_t(\boldsymbol{\theta}) = \sum_{i=1}^{t-1} \mathbb{1}\{\boldsymbol{\theta}_i = \boldsymbol{\theta}\}$ be the number of trajectories collected with policy $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ up to round $t-1$. To estimate the expected return $J(\boldsymbol{\theta})$, if no additional structure is available, we can only use the samples collected when executing $\pi_{\boldsymbol{\theta}}$, leading to the Monte Carlo (MC) estimator:

$$\widehat{J}_t^{\mathrm{MC}}(\boldsymbol{\theta}) = \frac{1}{T_t(\boldsymbol{\theta})} \sum_{i=1}^{t-1} \mathcal{R}(\tau_i)\, \mathbb{1}\{\boldsymbol{\theta}_i = \boldsymbol{\theta}\}. \quad (2)$$

$\widehat{J}_t^{\mathrm{MC}}$ is unbiased for $J(\boldsymbol{\theta})$ and its variance scales with $\mathbb{V}\mathrm{ar}[\widehat{J}_t^{\mathrm{MC}}(\boldsymbol{\theta})] \leqslant 1/T_t(\boldsymbol{\theta})$. Clearly, $\widehat{J}_t^{\mathrm{MC}}(\boldsymbol{\theta})$ can be computed only for the policies that have been executed at least once.

**Multiple Importance Sampling Estimation** With the mediator feedback, at each round $t \in [n]$ we have access to additional information, i.e., the history of parameter-trajectory-return triples $\mathcal{H}_t = \{(\boldsymbol{\theta}_i, \tau_i, \mathcal{R}(\tau_i))\}_{i=1}^{t-1}$. Thanks to the factorization in Equation (1), we can compute the trajectory distribution ratios without knowing $P$ and $\mu$:

$$\frac{p_{\boldsymbol{\theta}}(\tau)}{p_{\boldsymbol{\theta}'}(\tau)} = \prod_{h=0}^{H-1} \frac{\pi_{\boldsymbol{\theta}}(a_h|s_h)}{\pi_{\boldsymbol{\theta}'}(a_h|s_h)}.$$

Thus, we can use *all* the samples to estimate the expected return of *any* policy. Let $\Phi_t = \sum_{j=1}^{t-1} \frac{1}{t-1} p_{\boldsymbol{\theta}_j}$ be the mixture induced by the policies executed up to time $t-1$: if $p_{\boldsymbol{\theta}} \ll \Phi_t$, we can employ a Multiple Importance Sampling (MIS, Veach and Guibas 1995) estimator (with balance heuristic):[6]

$$\widehat{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1} \sum_{i=1}^{t-1} \omega_{\boldsymbol{\theta},t}(\tau_i) \mathcal{R}(\tau_i), \quad (3)$$

where $\omega_{\boldsymbol{\theta},t}(\tau_i) = p_{\boldsymbol{\theta}}(\tau_i)/\Phi_t(\tau_i)$ is the *importance weight*. Thus, for estimating the expected return $J(\boldsymbol{\theta})$ of policy $\pi_{\boldsymbol{\theta}}$ we do not need to execute $\pi_{\boldsymbol{\theta}}$, but just require the absolute continuity $p_{\boldsymbol{\theta}} \ll \Phi_t$ (surely fulfilled if $T_t(\boldsymbol{\theta}) \geqslant 1$). The statistical properties of the MIS estimator can be phrased in terms of the Rényi divergence. We can prove that $0 \leqslant \widehat{J}_t(\boldsymbol{\theta}) \leqslant d_\infty(p_{\boldsymbol{\theta}} \| \Phi_t)$ and the variance can be bounded as $\mathbb{V}\mathrm{ar}[\widehat{J}_t(\boldsymbol{\theta})] \leqslant d_2(p_{\boldsymbol{\theta}} \| \Phi_t)/(t-1)$ (Metelli et al. 2018; Papini et al. 2019; Metelli et al. 2020). Since the variance of $\widehat{J}_t(\boldsymbol{\theta})$ scales with $d_2(p_{\boldsymbol{\theta}} \| \Phi_t)/(t-1)$ instead of $1/T_t(\boldsymbol{\theta})$, as for $\widehat{J}_t^{\mathrm{MC}}(\boldsymbol{\theta})$, we refer to $\eta_t(\boldsymbol{\theta}) := (t-1)/d_2(p_{\boldsymbol{\theta}} \| \Phi_t)$ as the *effective number of trajectories*. It is worth noting that $\eta_t(\boldsymbol{\theta}) \geqslant T_t(\boldsymbol{\theta})$ (Lemma C.4); thus, thanks to the structure introduced by the mediator feedback, the MIS estimator variance is always smaller than the MC estimator variance.[7]

**Truncated Multiple Importance Sampling Estimation** The main limitation of the MIS estimator is that the importance weight $\omega_{\boldsymbol{\theta},t}$ displays a *heavy-tail* behavior, preventing exponential concentration, unless $d_\infty(p_{\boldsymbol{\theta}} \| \Phi_t)$ is finite (Metelli et al. 2018). A common solution consists in

[5] We stress that IS is just *one* method, and not necessarily the best one, to exploit the structure of the PO problem.

[6] For an extensive discussion of importance sampling and heuristics (e.g., balance heuristic) refer to (Owen 2013).

[7] The effective number of trajectories $\eta_t(\boldsymbol{\theta})$ is, in fact, the *effective sample size* of $\widehat{J}_t(\boldsymbol{\theta})$ (Martino, Elvira, and Louzada 2017).

*truncating* the estimator (Ionides 2008) at the cost of introducing a negative bias. Given a (time-variant and policy-dependent) truncation threshold $M_t(\boldsymbol{\theta}) < \infty$, the Truncated MIS (TMIS) was introduced by Papini et al. (2019):

$$\check{J}_t(\boldsymbol{\theta}) = \frac{1}{t-1}\sum_{i=1}^{t-1} \check{\omega}_{\boldsymbol{\theta},t}(\tau_i)\mathcal{R}(\tau_i), \tag{4}$$

where $\check{\omega}_{\boldsymbol{\theta},t}(\tau_i) = \min\{M_t(\boldsymbol{\theta}), \omega_{\boldsymbol{\theta},t}(\tau_i)\}$. TMIS enjoys more desirable theoretical properties than plain MIS. While its variance scales similarly to $\widehat{J}_t(\boldsymbol{\theta})$ since $\mathbb{V}\mathrm{ar}[\check{J}_t(\boldsymbol{\theta})] \leqslant d_2(p_{\boldsymbol{\theta}}\|\Phi_t)/(t-1)$, the range can be bounded as $0 \leqslant \check{J}_t(\boldsymbol{\theta}) \leqslant M_t(\boldsymbol{\theta})$. Thus, the range is controlled by $M_t(\boldsymbol{\theta})$ and no longer by the divergence $d_\infty(p_{\boldsymbol{\theta}}\|\Phi_t)$, which may be infinite. Similarly, the bias can be bounded as $J(\boldsymbol{\theta}) - \mathbb{E}_{\tau_i \sim p_{\boldsymbol{\theta}_i}}[\check{J}_t(\boldsymbol{\theta})] \leqslant d_2(p_{\boldsymbol{\theta}}\|\Phi_t)/M_t(\boldsymbol{\theta})$ (see Papini et al. (2019) and Lemma C.1 for details). If we are interested in minimizing the joint contribution of bias and variance, this suggests to increase $M_t(\boldsymbol{\theta})$ progressively over the rounds.

## 5 Deterministic Algorithms

In this section, we consider finite policy spaces ($|\Theta| < \infty$) and discuss algorithms for PO that select policies deterministically, i.e., $\boldsymbol{\theta}_t$ is a deterministic function of history $\mathcal{H}_{t-1}$.

**Follow The Leader** The simplest algorithm accounting for the mediator feedback is Follow The Leader (FTL). It maintains a TMIS estimator $\check{J}_t(\boldsymbol{\theta})$ and selects the policy with the highest estimated expected return, i.e., $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta}\in\Theta} \check{J}_t(\boldsymbol{\theta})$. This is a pure-exploitation algorithm, unsuited for bandit feedback. Surprisingly, under a strong form of mediator feedback, FTL enjoys *constant* regret.

**Theorem 5.1.** *Let $\Theta = [K]$, $v(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}'\in\Theta} d_2(p_{\boldsymbol{\theta}}\|p_{\boldsymbol{\theta}'})$ for all $\boldsymbol{\theta} \in \Theta$ and $v^*(\boldsymbol{\theta}) = \max\{v(\boldsymbol{\theta}), v(\boldsymbol{\theta}^*)\}$, where $\pi_{\boldsymbol{\theta}^*}$ is an optimal policy. If $v := \max_{\boldsymbol{\theta}\in\Theta} v(\boldsymbol{\theta}) < \infty$, then, for any $\alpha > 1$, the expected regret of FTL using TMIS with truncation $M_t(\boldsymbol{\theta}) = \sqrt{\frac{t d_2(p_{\boldsymbol{\theta}}\|\Phi_t)}{\alpha \log t}}$ is bounded as:*

$$\begin{aligned}\mathbb{E}\, R(n) \leqslant &\sum_{\boldsymbol{\theta}\in\Theta:\Delta(\boldsymbol{\theta})>0} \frac{48\alpha v^*(\boldsymbol{\theta})}{\Delta(\boldsymbol{\theta})} \log \frac{24\alpha v^*(\boldsymbol{\theta})}{\Delta(\boldsymbol{\theta})^2} \\ &+ \Delta(\boldsymbol{\theta}_1) + \frac{2K}{\alpha - 1}\min\left\{1, \sqrt{2\log v}\right\}.\end{aligned} \tag{5}$$

We refer to the condition when all pairwise Rényi divergences are finite (i.e., $v < \infty$) as *perfect mediator feedback*. In such case, we have the remarkable property that running *any* policy in $\Pi_\Theta$ provides information for *all* the others. Indeed, the effective number of trajectories satisfies $\eta_t(\boldsymbol{\theta}) \geqslant (t-1)/v$ (Lemma C.4). Unfortunately, when $v = \infty$, FTL degenerates to *linear* regret (Fact D.1).

**UCB1** We can always apply an algorithm for standard bandit feedback, like UCB1 (Lai and Robbins 1985; Auer, Cesa-Bianchi, and Fischer 2002), to PO with finite policy space, ignoring the mediator feedback. UCB1 maintains the sample mean $\widehat{J}^{\mathrm{MC}}_t(\boldsymbol{\theta})$ of the observed returns for each $\boldsymbol{\theta} \in \Theta$ and selects the one that maximizes $\widehat{J}^{\mathrm{MC}}_t(\boldsymbol{\theta}) + \sqrt{(\alpha \log t)/T_t(\boldsymbol{\theta})}$. The optimistic bonus favors policies that have been selected less often, in accordance with the OFU principle. Being designed for bandit feedback, UCB1 guarantees $\mathcal{O}(\Delta^{-1}\log n)$ regret (Auer, Cesa-Bianchi, and Fischer 2002) even if $v = \infty$, but it cannot exploit mediator feedback when actually present.

**Algorithm 1** OPTIMIST

```
Input: initial parameter $\boldsymbol{\theta}_1$, $\alpha > 1$
  Execute $\pi_{\boldsymbol{\theta}_1}$, observe $\tau_1 \sim p_{\boldsymbol{\theta}_1}$ and $\mathcal{R}(\tau_1)$
  for $t = 2, \dots, n$ do
      Compute expected return estimate $\check{J}_t(\boldsymbol{\theta})$
      Compute index:
          $B_t(\boldsymbol{\theta}) = \check{J}_t(\boldsymbol{\theta}) + (1+\sqrt{2})\sqrt{\frac{\alpha \log t}{\eta_t(\boldsymbol{\theta})}}$
      Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta}\in\Theta} B_t(\boldsymbol{\theta})$
      Execute $\pi_{\boldsymbol{\theta}_t}$, observe $\tau_t \sim p_{\boldsymbol{\theta}_t}$ and $\mathcal{R}(\tau_t)$
  end for
```

In principle, we could employ FTL or UCB1 based on whether $v$ is finite or infinite. There are two reasons why this approach might be inappropriate. First, we would disregard the possibility to share information among pairs of policies with finite divergence, losing possible practical benefits (not captured by the current regret analysis). Second, even when $v < \infty$, the regret of FTL is $\mathcal{O}(v\Delta^{-1}\log(v\Delta^{-2}))$ that, at finite time, might be worse than $\mathcal{O}(\Delta^{-1}\log n)$, especially for large $v$. Note that deriving the conditions on $v$ so that the regret of UCB1 is smaller than that of FLT is not practical since it would require the knowledge of the gap $\Delta$.

**OPTIMIST** The difficulty in combining the advantages of FTL and UCB1 is overcome by OPTIMIST (Algorithm 1), an OFU-based algorithm introduced by Papini et al. (2019).[8] It selects policies as to maximize an *optimistic* TMIS expected return estimate that favors policies with a lower effective number of trajectories. In the original paper (Papini et al. 2019), OPTIMIST is only shown to enjoy sublinear regret in high probability under perfect mediator feedback ($v < \infty$). We show here that OPTIMIST actually enjoys constant regret under perfect mediator feedback (like FTL) without ever degenerating into linear regret (like UCB1).

**Theorem 5.2.** *Let $\Theta = [K]$ and $v(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}'\in\Theta} d_2(p_{\boldsymbol{\theta}}\|p_{\boldsymbol{\theta}'})$ for all $\boldsymbol{\theta} \in \Theta$ ($v(\boldsymbol{\theta})$ can be infinite). For any $\alpha > 1$, the expected regret of OPTIMIST with truncation $M_t(\boldsymbol{\theta}) = \sqrt{\frac{t d_2(p_{\boldsymbol{\theta}}\|\Phi_t)}{\alpha \log t}}$ is bounded as:*
*(a) if $v := \max_{\boldsymbol{\theta}\in\Theta} v(\boldsymbol{\theta}) < \infty$:*

$$\mathbb{E}\, R(n) \leqslant \sum_{\boldsymbol{\theta}\in\Theta:\Delta(\boldsymbol{\theta})>0} \frac{48\alpha v(\boldsymbol{\theta})}{\Delta(\boldsymbol{\theta})} \log \frac{24\alpha v(\boldsymbol{\theta})}{\Delta(\boldsymbol{\theta})^2}$$

[8] We consider here a slight variant of OPTIMIST with an explicit exploration parameter $\alpha$ in place of the original confidence parameter $\delta$ from (Papini et al. 2019), since we focus on expected regret rather than high-probability regret. $\alpha$ controls the repartition of the regret between the constant and logarithmic parts.

$$+ \Delta(\boldsymbol{\theta}_1) + \frac{2K}{\alpha - 1} \min\left\{1, \sqrt{2 \log v}\right\};$$

*(b) in any case:*

$$\mathbb{E}\, R(n) \leqslant \sum_{\boldsymbol{\theta} \in \Theta : \Delta(\boldsymbol{\theta}) > 0} \frac{24\alpha}{\Delta(\boldsymbol{\theta})} \log n + \frac{\alpha + 1}{\alpha - 1} K,$$

*with an instance-independent expected regret of* $\mathbb{E}\, R(n) \leqslant 4\sqrt{6\alpha K n \log n} + (\alpha + 1)K/(\alpha - 1)$.

Note also that the regret correctly goes to zero with the divergence (when $v = 1$, all the policies are equivalent). It is an interesting open problem whether better regret guarantees can be provided for the intermediate case, i.e., when some (but not all) the Rényi divergences are finite.

**Algorithm 2** RANDOMIST

```
Input: initial parameter $\boldsymbol{\theta}_1$, scale $a \geqslant 0$, translation $b \geqslant 0$, $\alpha > 1$
  Execute $\pi_{\boldsymbol{\theta}_1}$, observe $\tau_1 \sim p_{\boldsymbol{\theta}_1}$ and $\mathcal{R}(\tau_1)$
  for $t = 2, \dots, n$ do
      Compute expected return estimate $\check{J}_t(\boldsymbol{\theta})$
      Generate perturbation:
          $U_t(\boldsymbol{\theta}) = \frac{1}{\eta_t(\boldsymbol{\theta})} \sum_{l=1}^{a\eta_t(\boldsymbol{\theta})} \zeta_l + b$, with $\zeta_l \sim \mathrm{Ber}(1/2)$
      Select $\boldsymbol{\theta}_t \in \arg\max_{\boldsymbol{\theta} \in \Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$
      Execute $\pi_{\boldsymbol{\theta}_t}$, observe $\tau_t \sim p_{\boldsymbol{\theta}_t}$ and $\mathcal{R}(\tau_t)$
  end for
```

## 6 Randomized Algorithms

In this section, we propose a novel algorithm for regret minimization in PO that selects the policies with a randomized strategy. RANDOMIST (RANDomized-exploration policy Optimization via Multiple Importance Sampling with Truncation, Algorithm 2) is based on PHE (Kveton et al. 2019a) and employs additional samples to *perturb* the TMIS expected return estimate $\check{J}_t(\boldsymbol{\theta})$, enforcing exploration.[9] Clearly, RANDOMIST shares the randomized nature of exploration with the Bayesian approaches for bandits (e.g., Thompson Sampling (Thompson 1933)) although no prior-posterior mechanism is explicitly implemented and no assumption (apart for boundedness) on the return distribution is needed. At each round $t = 2, \dots, n$, we update the TMIS expected return estimate for each policy $\check{J}_t(\boldsymbol{\theta})$ and we generate the perturbation $U_t(\boldsymbol{\theta})$ that is obtained through $a\eta_t(\boldsymbol{\theta})$ *pseudo-rewards* sampled from a Bernoulli distribution $\mathrm{Ber}(1/2)$. Then, we play the policy maximizing the *perturbed estimated expected return*, i.e., the sum of the estimated expected return $\check{J}_t(\boldsymbol{\theta})$ and the perturbation $U_t(\boldsymbol{\theta})$. The two hyperparameters are the *perturbation scale* $a > 0$ and the *perturbation translation* $b > 0$. Informally, $a$ and $b$ are responsible for the amount of exploration: $a$ governs the variance of the perturbation, while $b$ (which is absent in PHE) accounts for the negative bias introduced by the TMIS estimator. We now present the properties of RANDOMIST with finite parameter space and propose an extension to deal with compact parameter spaces.

**Finite Parameter Space** If the policy space is finite, we can show that RANDOMIST enjoys guarantees similar to those of OPTIMIST on the expected regret.

**Theorem 6.1.** *Let* $\Theta = [K]$, $v(\boldsymbol{\theta}) = \max_{x' \in \Theta} d_2(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}'})$ *for all* $\boldsymbol{\theta} \in \Theta$ *(*$v(\boldsymbol{\theta})$ *can be infinite) and* $v^*(\boldsymbol{\theta}) = \max\{v(\boldsymbol{\theta}), v(\boldsymbol{\theta}^*)\}$ *where* $\pi_{\boldsymbol{\theta}*}$ *is an optimal policy. For any* $\alpha > 1$, *the expected regret of RANDOMIST with truncation* $M_t(\boldsymbol{\theta}) = \sqrt{\frac{t d_2(p_{\boldsymbol{\theta}} \| \Phi_t)}{\alpha \log t}}$ *is bounded as follows:*

*(a) if* $v := \max_{\boldsymbol{\theta} \in \Theta} v(\boldsymbol{\theta}) < \infty$, $b \leqslant \sqrt{(\alpha \log t)/\eta_t(\boldsymbol{\theta})}$ *and* $a \geqslant 0$*:*

$$\mathbb{E}\, R(n) \leqslant \Delta(\boldsymbol{\theta}_1) + \frac{\alpha + 3}{\alpha - 1} \min\left\{1, \sqrt{2 \log v}\right\} K$$
$$+ \sum_{\boldsymbol{\theta} \in \Theta : \Delta(\boldsymbol{\theta}) > 0} \frac{(188 + 32a)\alpha v^*(\boldsymbol{\theta})}{\Delta(\boldsymbol{\theta})} \log \frac{(94 + 16a)\alpha v^*(\boldsymbol{\theta})}{\Delta(\boldsymbol{\theta})^2};$$

*(b) no matter the value of* $v$, *if* $a > 8$ *and* $J(\boldsymbol{\theta}) - \mathbb{E}[\check{J}_t(\boldsymbol{\theta})] \leqslant b \leqslant \sqrt{(\alpha \log t)/\eta_t(\boldsymbol{\theta})}$*:*

$$\mathbb{E}\, R(n) \leqslant \sum_{\boldsymbol{\theta} \in \Theta : \Delta(\boldsymbol{\theta}) > 0} \frac{(52 + 110a)c\alpha}{\Delta(\boldsymbol{\theta})} \log n + 2\frac{\alpha + 1}{\alpha - 1} K,$$

*where* $c = 2 + \frac{e^2 \sqrt{a}}{\sqrt{2\pi}} \exp\left[\frac{16}{a-8}\right]\left(1 + \sqrt{\frac{\pi a}{a-8}}\right)$, *with an instance-independent expected regret of* $\mathbb{E}\, R(n) \leqslant 2\sqrt{(52 + 110a)c\alpha K n \log n} + 2\frac{\alpha+1}{\alpha-1} K$.

Under perfect mediator feedback RANDOMIST enjoys constant regret, like OPTIMIST, although with a dependence on $v^*(\boldsymbol{\theta})$, which involves the divergence w.r.t. an optimal policy. Moreover, in such case, since exploration is not needed, we could even set $a = b = 0$ reducing RANDOMIST to FTL. Similarly to OPTIMIST, when we allow $v = \infty$, the regret becomes logarithmic and the hyperparameters $a$ and $b$ must be carefully set to enforce exploration.

**Compact Parameter Space** When the parameter space is a compact set, i.e., $\Theta = [-M, M]^d$, the $\arg\max$ in Algorithm 2 cannot be explicitly computed. However, the random variable $\boldsymbol{\theta} \in \arg\max_{\boldsymbol{\theta}' \in \Theta} \check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$ can be seen as sampled from the distribution for $\boldsymbol{\theta}$ of being the parameter in $\Theta$ with the largest perturbed estimated expected return, whose p.d.f. is given by (D'Eramo et al. 2017):

$$\mathfrak{g}_t^*(\boldsymbol{\theta}) = g\Big(\check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \Theta} \check{J}_t(\boldsymbol{\theta}') + U_t(\boldsymbol{\theta}') | \mathcal{H}_{t-1}\Big)$$
$$= \int_{\mathbb{R}} \frac{g_{\boldsymbol{\theta}}(y)}{G_{\boldsymbol{\theta}}(y)} \prod_{\Theta} G_{\boldsymbol{\theta}'}(y) \mathrm{d}\boldsymbol{\theta}' \mathrm{d}y, \qquad (6)$$

where $\prod_{\Theta} G_{\boldsymbol{\theta}}(y) \mathrm{d}\boldsymbol{\theta} = \exp\left(\int_{\Theta} \log G_{\boldsymbol{\theta}}(y) \mathrm{d}\boldsymbol{\theta}\right)$ is the *product integral* (Davis and Chatfield 1970), $g_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\theta}}$ are the p.d.f. and the c.d.f. of the random variable $\check{J}_t(\boldsymbol{\theta}) + U_t(\boldsymbol{\theta})$ conditioned to the history $\mathcal{H}_{t-1}$. The *computation* of $\mathfrak{g}_t^*$ (even up to a constant) is challenging as the product integral requires a numerical integration over the parameter space $\Theta$. Provided that an approximation (up to a constant) $\mathfrak{g}_t^\dagger$ of $\mathfrak{g}_t^*$ is available, we can use a Monte Carlo Markov Chain method (Owen

[9] In this sense, RANDOMIST, as well as PHE, resembles the Follow the Perturbed Leader (Hannan 1957) strategy.

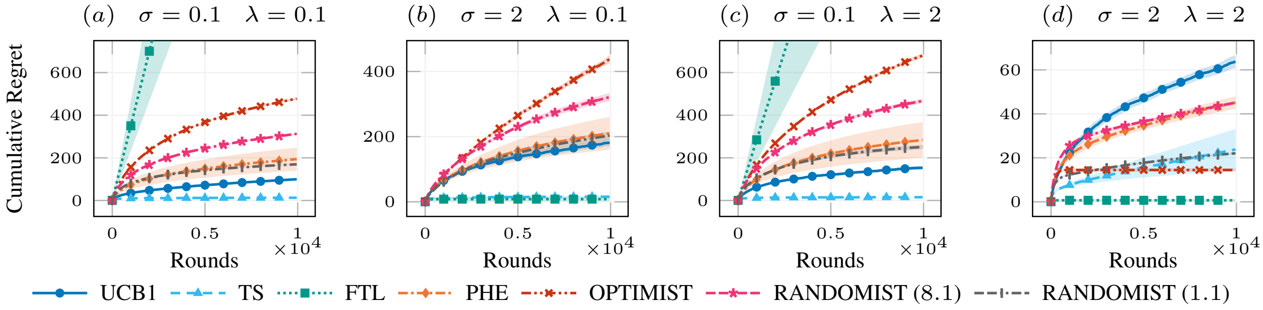

Figure 2: Cumulative regret on the illustrative PO for four values of $\sigma$ and $\lambda$. 20 runs, 95% c.i.

2013) to generate a sample $\boldsymbol{\theta}_t \sim \mathfrak{g}_t^\dagger$. As a practical approximation, we consider the p.d.f. for $\boldsymbol{\theta}$ of having a perturbed estimated expected return larger than that of the previously executed policies:[10] $\mathfrak{g}_t^\dagger(\boldsymbol{\theta}) \propto \int_{\mathbb{R}} g_{\boldsymbol{\theta}}(y) \prod_{i=1}^{t-1} G_{\boldsymbol{\theta}_i}(y) \mathrm{d}y$.

Since $\mathcal{O}(d)$ iterations of MCMC are sufficient to generate a sample (Beskos and Stuart 2009), where $d$ is the dimensionality of $\Theta$, and one evaluation of $\mathfrak{g}_t^\dagger$ can be performed in time $\mathcal{O}(t^3)$, the per-round complexity of RANDOMIST is $\mathcal{O}(dt^3)$. This can be further reduced to $\mathcal{O}(dt^2)$ via clever caching (see Appendix F). OPTIMIST (Papini et al. 2019) can also be applied to continuous parameter spaces, with an $\widetilde{\mathcal{O}}(\sqrt{vdn})$ high-probability regret bound. However, it is not clear how to perform the maximization step of OPTIMIST efficiently in this setting, since the optimistic index is non-differentiable and non-convex in the parameter variable. Discretization is adopted in (Papini et al. 2019), leading to $\mathcal{O}(t^{1+d/2})$ time complexity, that is exponential in $d$. The RANDOMIST variant proposed here, although heuristic, has only polynomial dependence on $d$, thus scaling more favorably to high-dimensional problems.

## 7 Numerical Simulations

We present the numerical simulations, starting with an illustrative example and then moving to RL benchmarks. For the RL experiments, similarly to Papini et al. (2019), the evaluation is carried out in the parameter-based PO setting (Sehnke et al. 2008), where the policy parameters $\boldsymbol{\theta}$ are sampled from a *hyperpolicy* $\nu_{\boldsymbol{\xi}}$ and the optimization is performed in the space of *hyperparameters* $\Xi$ (Appendix A). This setting is particularly convenient since the Rényi divergence between hyperpolicies can be computed exactly (at least for Gaussians). Details and an additional experiment on the Cartpole domain are reported in Appendix F.

**Illustrative Problems** The goal of this experiment is to show the advantages of the additional structure offered by the mediator feedback over the bandit feedback. We design a class of 5-policy PO problems, isomorphic to bandit problems, in which trajectories are collapsed to a single real action $\mathcal{T} = \mathbb{R}$ and $\mathcal{R}(\tau) = \max\{0, \min\{1, \tau/4\}\}$. The policies are Gaussians ($\mathcal{N}(0,\sigma^2)$, $\mathcal{N}(1,\sigma^2)$, $\mathcal{N}(2,\sigma^2)$, $\mathcal{N}(2.95,\lambda^2)$, $\mathcal{N}(3,\sigma^2)$) defined in terms of the two values $\sigma, \lambda > 0$. The optimal policy is the fifth one and we have a near-optimal parameter, the fourth, with a different variance. Intuitively, we can tune the parameters $\sigma$ and $\lambda$ to vary the Rényi divergences. We compare RANDOMIST with $a = 8.1$ (as prescribed in Theorem 6.1) and $a = 1.1$, and $b = \sqrt{(\alpha \log t)/\eta_t(\boldsymbol{\theta})}$ for both cases, with OPTIMIST (Papini et al. 2019), FTL, UCB1 (Auer, Cesa-Bianchi, and Fischer 2002), PHE (Kveton et al. 2019a), and TS with Gaussian prior (Agrawal and Goyal 2013). The cumulative regret is shown in Figure 2 for four combinations of $\sigma$ and $\lambda$. In (a) and (d) we are in a perfect mediator feedback, but in (a) $\log v \simeq 2.25$ and (d) $\log v \simeq 900$. Instead, in (b) or (c), we have $v = \infty$. We notice that FTL displays a (near-)linear regret in (a) as expected since $v = \infty$ but also in (c) where $v$ is finite but very large. RANDOMIST with theoretical value of $a = 8.1$ always displays a good behavior and better than OPTIMIST, except in (d) where the latter shows a remarkable constant regret. We also note that when the amount of information shared among parameters is small, UCB1 performs better than OPTIMIST as well as PHE over RANDOMIST. Furthermore, TS with Gaussian prior performs very well across the tasks, although it considers the bandit feedback. This can be explained since TS assumes the correct return distribution. It also suggests that RANDOMIST could be improved when coped with other perturbation distributions (e.g., Gaussian). Finally, we observe that RANDOMIST with $a = 1.1$, although violating the conditions of Theorem 6.1, keeps showing a sublinear regret even in (b) and (c) when $v = \infty$.

**Linear Quadratic Gaussian Regulator** The Linear Quadratic Gaussian Regulator (LQG, Curtain 1997) is a benchmark for continuous control. We consider the monodimensional case and a Gaussian hyperpolicy $\nu_\xi = \mathcal{N}(\xi, 0.15^2)$ where $\xi$ is the learned parameter. From $\nu_\xi$, we sample the gain $\theta$ of a deterministic linear policy: $a_h = \theta s_h$. This experiment aims at comparing RANDOMIST with UCB1 (Auer, Cesa-Bianchi, and Fischer 2002), GPUCB (Srinivas et al. 2010), and OPTIMIST (Papini et al. 2019) in a finite policy space by discretizing $[-1, 1]$ in $K = 100$ parameters. In Figure 3, we notice that OPTIMIST and RANDOMIST outperform UCB1. While RAN-

[10] $\mathfrak{g}_t^\dagger$ can be seen as obtained from $\mathfrak{g}_t^*$ applying a quadrature with $\{\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_{t-1}\}$ as nodes for the inner integral.

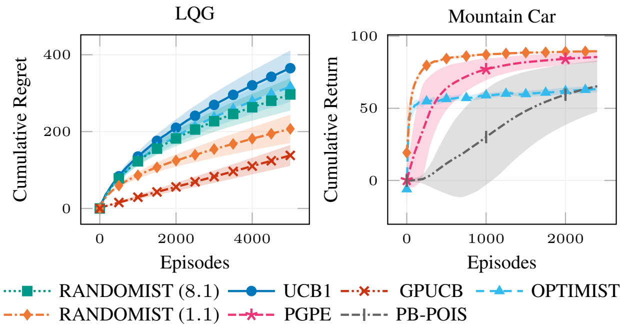

Figure 3: Cumulative regret in the LQG (30 runs, 95% c.i.) and cumulative return in the Mountain Car (5 runs, 95% c.i.).

DOMIST with $a = 8.1$ and OPTIMIST have similar performance, RANDOMIST improves significantly when setting $a$ to 1.1. As in (Papini et al. 2019), the good performance of GPUCB is paired with a lack of theoretical guarantees due to the arbitrary choice of the GP kernel.

**Mountain Car** To test RANDOMIST in a continuous parameter space, we employ the approximation described above in the Mountain Car environment (Sutton and Barto 2018). We consider the setting of (Papini et al. 2019), employing PGPE (Sehnke et al. 2008) and PB-POIS (Metelli et al. 2018) as baselines. We use a Gaussian hyperpolicy $\nu_{\boldsymbol{\xi}} = \mathcal{N}(\boldsymbol{\xi}, \mathrm{diag}(0.15, 3)^2)$ with learned mean $\boldsymbol{\xi}$, from which we sample the parameters of a deterministic policy, linear in position and velocity. The exploration phase is performed by sampling from the approximate density $\mathfrak{g}_t^\dagger$, taking 10 steps of the Metropolis-Hastings algorithm (Owen 2013) with Gaussian proposal $q_m = \mathcal{N}(\boldsymbol{\theta}_m, \mathrm{diag}(0.15, 3)^2)$. Figure 3 shows that RANDOMIST outperforms both policy gradient baselines and OPTIMIST, in terms of learning speed and final performance.

## 8 Related Works

In this section, we revise the related literature, with attention to bandits with expert advice and to provably efficient PO. Additional comparisons are reported in Appendix B.

**Mediator Feedback and Expert Advice** A related formulation are the *Bandits with Expert Advice* (BEA, Bubeck and Cesa-Bianchi 2012, Section 4.2), introduced as an approach to adversarial contextual bandits. To draw a parallelism with PO, let $\mathcal{T}$ be the set of arms and $\Theta = [K]$ the finite set of experts. At each step $t$, the agent receives *advice* $p^t_{\boldsymbol{\theta}} \in \mathscr{P}(\mathcal{T})$ from each expert $\boldsymbol{\theta} \in \Theta$, selects one expert $\boldsymbol{\theta}_t$, and pulls arm $\tau_t \sim p^t_{\boldsymbol{\theta}_t}$. The goal is to minimize the *in-class* regret, competing with the best expert in hindsight. Differently from the trajectory distributions of PO, expert advice can change with time. A major concern of BEA, also relevant to PO, is the dependency of the regret on the number $K$ of experts (resp. policies). A naïve application of Exp3 (Auer et al. 2002) yields $\mathcal{O}(\sqrt{nK \log K})$ regret. Like our PO algorithms, this is impractical when the experts are exponentially many. Exp4 (Auer et al. 2002) achieves $\mathcal{O}(\sqrt{n|\mathcal{T}| \log K})$ regret, which scales well with $K$, but is vacuous in the case of infinite arms. McMahan and Streeter (2009) replace $|\mathcal{T}|$ with the *degree of agreement* of the experts, which has interesting similarities with our distributional-divergence approach. *Meta-bandit* approaches (Agarwal et al. 2017; Pacchiano et al. 2020) are so general that could be applied both to continuous-arm BEA and PO, but also exhibit a superlogarithmic dependence on $K$. Beygelzimer et al. (2011) obtain $\widetilde{\mathcal{O}}(\sqrt{dn})$ regret competing with an infinite set of experts of VC-dimension $d$, mirrored in PO by OPTIMIST on compact spaces of dimension $d$ (Papini et al. 2019, Theorem 3).

**Provably Efficient PO** Recently, a surge of approaches to deal with PO in a theoretically sound way, with both stochastic or adversarial environments, has emerged. These works consider either *full-information*, i.e., the agent observes the whole reward function $\{\mathcal{R}(s_h, a)\}_{a \in \mathcal{A}}$ regardless the played action (e.g., Rosenberg and Mansour 2019; Cai et al. 2019), or the *bandit feedback* (with a different meaning compared to the use we have made in this paper), in which only the reward of the chosen action is observed $\mathcal{R}(s_h, a_h)$ (e.g., Jin et al. 2019; Efroni et al. 2020). These methods are not directly comparable with the mediator feedback, although both settings exploit the structure of the PO problem. While with MF we explicitly model the policy space $\Pi_\Theta$, these methods search in the space of all Markovian stationary policies. Furthermore, they are limited to tabular MDPs, while MF can deal natively with continuous state-action spaces.

## 9 Discussion and Conclusions

We have deepened the understanding of policy optimization as an online learning problem with additional feedback. We believe that mediator feedback has potential applications even beyond PO. Indeed, the problem of optimizing over probability distributions also encompasses GANs and variational inference (Chu, Blanchet, and Glynn 2019) and, more generally, MF emerges in any Bayesian network in which we control the conditional distributions on some vertexes, via parameters $\boldsymbol{\theta}$, while the other are fixed and independent from $\boldsymbol{\theta}$. Furthermore, we have introduced a novel randomized algorithm, RANDOMIST, and we have shown its advantages both in terms of computational complexity and performance. The algorithm could be improved by adopting a different perturbation, e.g., Gaussian, as already hinted in (Kveton et al. 2019b). Further work is needed to match the theoretical regret lower bounds. Currently, a major discrepancy is the use of the KL-divergence in the lower bounds instead of the larger Rényi divergence required by algorithms based on IS. Moreover, the algorithm employs the ratio importance weight and, thus, it might suffer from the curse of horizon (Liu et al. 2018). Finally, the case of non-perfect mediator feedback could be related to graphical bandits (Alon et al. 2017), where finite Rényi divergences are the edges of a directed feedback graph, in order to capture the actual difficulty of this intermediate case.

## Acknowledgments

This work has been partially supported by the Italian MIUR PRIN 2017 Project ALGADIMAR ”Algorithms, Games, and Digital Markets”.

## References

Agarwal, A.; Luo, H.; Neyshabur, B.; and Schapire, R. E. 2017. Corralling a Band of Bandit Algorithms. In *COLT*.

Agrawal, S.; and Goyal, N. 2012. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *COLT*.

Agrawal, S.; and Goyal, N. 2013. Further Optimal Regret Bounds for Thompson Sampling. In *AISTATS*.

Alon, N.; Cesa-Bianchi, N.; Gentile, C.; Mannor, S.; Mansour, Y.; and Shamir, O. 2017. Nonstochastic Multi-Armed Bandits with Graph-Structured Feedback. *SIAM J. Comput.* 46(6): 1785–1826.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* 47(2-3): 235–256.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.* 32(1): 48–77.

Beskos, A.; and Stuart, A. 2009. Computational complexity of Metropolis-Hastings methods in high dimensions. In *Monte Carlo and Quasi-Monte Carlo Methods 2008*, 61–71.

Beygelzimer, A.; Langford, J.; Li, L.; Reyzin, L.; and Schapire, R. E. 2011. Contextual Bandit Algorithms with Supervised Learning Guarantees. In *AISTATS*.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning* 5(1): 1–122.

Bubeck, S.; Perchet, V.; and Rigollet, P. 2013. Bounded regret in stochastic multi-armed bandits. In *COLT*.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2019. Provably Efficient Exploration in Policy Optimization. *arXiv preprint arXiv:1912.05830* .

Chapelle, O.; and Li, L. 2011. An Empirical Evaluation of Thompson Sampling. In *NeurIPS*.

Chowdhury, S. R.; and Gopalan, A. 2019. Online Learning in Kernelized Markov Decision Processes. In *AISTATS*.

Chu, C.; Blanchet, J. H.; and Glynn, P. W. 2019. Probability Functional Descent: A Unifying Perspective on GANs, Variational Inference, and Reinforcement Learning. In Chaudhuri, K.; and Salakhutdinov, R., eds., *ICML*.

Cochran, W. G. 1977. *Sampling Techniques, 3rd Edition*. John Wiley. ISBN 0-471-16240-X.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning Bounds for Importance Weighting. In *NeurIPS*.

Curtain, R. F. 1997. Linear-quadratic control: An introduction. *Autom.* 33(5): 1004.

Davis, W.; and Chatfield, J. 1970. Concerning product integrals and exponentials. *AMS* .

Deisenroth, M. P.; Neumann, G.; and Peters, J. 2013. A Survey on Policy Search for Robotics. *Foundations and Trends in Robotics* 2(1-2): 1–142.

D'Eramo, C.; Nuara, A.; Pirotta, M.; and Restelli, M. 2017. Estimating the maximum expected value in continuous reinforcement learning problems. In *AAAI*.

Efroni, Y.; Shani, L.; Rosenberg, A.; and Mannor, S. 2020. Optimistic Policy Optimization with Bandit Feedback. *arXiv preprint arXiv:2002.08243* .

Hannan, J. 1957. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games* 3: 97–139.

Ionides, E. L. 2008. Truncated importance sampling. *JCGS* 17(2): 295–311.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *J. Mach. Learn. Res.* 11: 1563–1600.

Jin, C.; Jin, T.; Luo, H.; Sra, S.; and Yu, T. 2019. Learning adversarial markov decision processes with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192* .

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In *ALT*.

Kveton, B.; Szepesvári, C.; Ghavamzadeh, M.; and Boutilier, C. 2019a. Perturbed-History Exploration in Stochastic Multi-Armed Bandits. In *IJCAI*.

Kveton, B.; Szepesvári, C.; Vaswani, S.; Wen, Z.; Lattimore, T.; and Ghavamzadeh, M. 2019b. Garbage In, Reward Out: Bootstrapping Exploration in Multi-Armed Bandits. In *ICML*.

Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1): 4–22.

Lattimore, T.; and Szepesvári, C. 2018. Bandit algorithms .

Liu, Q.; Li, L.; Tang, Z.; and Zhou, D. 2018. Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation. In *NeurIPS*, 5361–5371.

Martino, L.; Elvira, V.; and Louzada, F. 2017. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing* 131: 386–401.

McMahan, H. B.; and Streeter, M. J. 2009. Tighter Bounds for Multi-Armed Bandits with Expert Advice. In *COLT*.

Metelli, A. M.; Papini, M.; Faccio, F.; and Restelli, M. 2018. Policy Optimization via Importance Sampling. In *NeurIPS*.

Metelli, A. M.; Papini, M.; Montali, N.; and Restelli, M. 2020. Importance Sampling Techniques for Policy Optimization. *JMLR* 21(141): 1–75.

Owen, A. B. 2013. Monte Carlo theory, methods and examples. *Monte Carlo Theory, Methods and Examples* .

Pacchiano, A.; Phan, M.; Abbasi-Yadkori, Y.; Rao, A.; Zimmert, J.; Lattimore, T.; and Szepesvári, C. 2020. Model Selection in Contextual Stochastic Bandit Problems. *CoRR* abs/2003.01704.

Papini, M.; Metelli, A. M.; Lupo, L.; and Restelli, M. 2019. Optimistic Policy Optimization via Multiple Importance Sampling. In *ICML*.

Peng, X. B.; Coumans, E.; Zhang, T.; Lee, T.-W.; Tan, J.; and Levine, S. 2020. Learning Agile Robotic Locomotion Skills by Imitating Animals. *arXiv preprint arXiv:2004.00784* .

Peters, J.; and Schaal, S. 2008. Reinforcement learning of motor skills with policy gradients. *Neural Networks* 21(4): 682–697.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley.

Rényi, A. 1961. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*.

Rosenberg, A.; and Mansour, Y. 2019. Online Convex Optimization in Adversarial Markov Decision Processes. In *ICML*.

Sehnke, F.; Osendorfer, C.; Rückstieß, T.; Graves, A.; Peters, J.; and Schmidhuber, J. 2008. Policy Gradients with Parameter-Based Exploration for Control. In *ICANN*.

Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M. W. 2010. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *ICML*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4): 285–294.

Veach, E.; and Guibas, L. J. 1995. Optimally combining sampling techniques for Monte Carlo rendering. In Mair, S. G.; and Cook, R., eds., *SIGGRAPH*, 419–428. ACM.

Vinyals, O.; Babuschkin, I.; Czarnecki, W.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D.; Powell, R.; Ewalds, T.; Georgiev, P.; Oh, J.; Horgan, D.; Kroiss, M.; Danihelka, I.; Huang, A.; Sifre, L.; Cai, T.; Agapiou, J.; Jaderberg, M.; and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575(7782): 350–354.