Behind the Myth of Exploration in Policy Gradients

Adrien Bolland [1] **Gaspard Lambrechts** [1] **Damien Ernst** [1 2]

Abstract

Policy-gradient algorithms are effective reinforcement learning methods for solving control problems with continuous state and action spaces. To compute near-optimal policies, it is essential in practice to include exploration terms in the learning objective. Although the effectiveness of these terms is usually justified by an intrinsic need to explore environments, we propose a novel analysis and distinguish two different implications of these techniques. First, they make it possible to smooth the learning objective and to eliminate local optima while preserving the global maximum. Second, they modify the gradient estimates, increasing the probability that the stochastic parameter update eventually provides an optimal policy. In light of these effects, we discuss and illustrate empirically exploration strategies based on entropy bonuses, highlighting their limitations and opening avenues for future works in the design and analysis of such strategies.

1. Introduction

Many practical problems require making sequential decisions in environments, based on state observations, in order to minimize a cost or maximize a reward. Reinforcement learning is a framework for solving such decision-making problems that has been successful on complex tasks, including playing games (Mnih et al., 2015; Silver et al., 2017), managing power systems (Aittahar et al., 2024), controlling robots (Kalashnikov et al., 2018), or interacting with electricity markets (Boukas et al., 2021).

Reinforcement learning can be divided into three families of algorithms, namely, model-based, value-based, and policy-based methods. Each method exhibits different learning dynamics and requirements for computing high-performing policies. On the one hand, the first two families of algorithms are subject to the exploration-exploitation dilemma during the learning procedure. In short, in order to learn statistical estimates of the environment or the value functions as fast as possible, from which a good policy can be computed, it is necessary to take actions that increase the quality of the estimates that are likely not optimal. This need for exploration to achieve high performance is theoretically well understood and has been the subject of many works (Dann et al., 2017; Azar et al., 2017; Neu & Pike-Burke, 2020). On the other hand, in policy-based methods, and especially for policy-gradient algorithms (Duan et al., 2016; Andrychowicz et al., 2020), the main theoretical requirement to converge towards globally (or even locally) optimal solutions is that policies remain sufficiently stochastic during the learning procedure (Bhandari & Russo, 2019; Bhatt et al., 2019; Agarwal et al., 2020; Zhang et al., 2021a; Bedi et al., 2022). Interestingly, stochastic policies have smoother returns (Ahmed et al., 2019; Bolland et al., 2023), but neither softmax nor Gaussian policies guarantee enough stochasticity for ensuring (fast) convergence (Mei et al., 2020; 2021; Bedi et al., 2022). This requirement of stochasticity in policy gradient is often abusively called exploration and often understood as the need to infinitely sample all states and actions.

Practitioners have tried to meet the theoretical requirement of sufficient randomness of policies in policy gradient via reward-shaping strategies, whereby a learning objective that promotes or hinders behaviors by providing reward bonuses for some states and actions is optimized as a surrogate to the return of the policy. These bonuses typically promote actions that reduce the uncertainty of the agent about its environment (Pathak et al., 2017; Burda et al., 2018; Zhang et al., 2021c), or that maximize the entropy of states and/or actions (Bellemare et al., 2016; Lee et al., 2019; Guo et al., 2021; Williams & Peng, 1991; Haarnoja et al., 2019). Optimizing a surrogate objective is particularly effective for solving tasks with complex dynamics and reward functions, or with sparse rewards (Islam et al., 2019; Lee et al., 2019; Liu & Abbeel, 2021; Zhang et al., 2021b; Guo et al., 2021).

The differences between theory and practical implementations of exploration has led to common folklore seeking to explain the intuition behind and the efficiency of policy-gradient methods. This work is part of the research line that studies the maximization of practical surrogate learning objective functions from a mathematical optimization

[1]Montefiore Institute, University of Liège, Belgium [2]Telecom Paris, Institut Polytechnique de Paris, France. Correspondence to: Adrien Bolland <adrien.bolland@uliege.be>.

perspective. Close to our work, studies of the learning objective with entropy regularization (an exploration-based reward shaping technique where the entropy of the policy is added in the learning objective) were conducted. It includes the study by (Ahmed et al., 2019) concluding that it helps to provide smooth learning objective functions. The same exploration strategy was reinterpreted as a robust optimization method by Husain et al. (2021) and equivalently as a two-player game by Brekelmans et al. (2022). Bolland et al. (2023) furthermore argued that optimizing an entropy regularized objective is equivalent to optimizing the return of another policy with larger variance. Chung et al. (2021) also studied the effect on the learning dynamics when including baselines in policy gradient, which is close to adding exploration terms in the learning objective. These studies are specific to some exploration methods and the literature lacks unified explanations and interpretations about exploration in policy-gradient methods.

Before delving into our contributions, we recall that the convergence of stochastic ascent methods is driven by the objective function and how the ascent directions are estimated. First, the objective function shall be (pseudo) concave to find its global maximum (Bottou, 1998). Second, the convergence rate is influenced by the distribution of the stochastic ascent estimates (Chen & Luss, 2018; Ajalloeian & Stich, 2020). In this paper, we rigorously study policy-gradient methods with exploration-based reward shaping through the lens of these two optimization theory aspects. More precisely, we first discuss the effect of exploration on the learning objective and the relationship between an optimal policy and a policy maximizing the learning objective. Second, we elaborate on the distribution of the gradient estimates of the learning objective and its likelihood of providing a direction in which the learning objective and the return increase. We furthermore illustrate how some common exploration strategies help improve the performance of policy-gradient algorithms with respect to these two aspects. In practice, finding good exploration strategies is known to be problem specific and we thus introduce a general framework for the study and interpretation of exploration in policy-gradient methods instead of trying to find the best exploration method for a given task.

The paper is organized as follows. In Section 2, we provide the background about policy gradients and about exploration. Section 3 focuses on the effect of exploration on the learning objective while Section 4 is dedicated to the effect on the gradient estimates used in the policy-gradient algorithms[1]. Finally, conclusions and future works are discussed in Section 5.

[1]Experimental details and implementations can be found at `https://github.com/adrienBolland/micro-rl-lib`.

2. Background

In this section, we introduce the reinforcement learning problem in Markov decision processes and discuss the policy-gradient optimization method with exploration.

2.1. Markov Decision Processes

We study problems in which an agent makes sequential decisions in a stochastic environment in order to maximize an expected sum of rewards (Sutton & Barto, 2018). The environment is modeled with an infinite-time Markov Decision Process (MDP) composed of a state space $\mathcal{S}$, an action space $\mathcal{A}$, an initial state distribution with density p_0, a transition distribution (modeling the dynamics) with conditional density p, a bounded reward function ρ, and a discount factor $\gamma \in [0, 1($. When an agent interacts with the MDP, first, an initial state $s_0 \sim p_0(\cdot)$ is sampled, then, the agent provides at each time step t an action $a_t \in \mathcal{A}$ leading to a new state $s_{t+1} \sim p(\cdot|s_t, a_t)$. Such a sequence of states and actions $h_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t) \in \mathcal{H}$ is called a history and $\mathcal{H}$ is the set of all histories of any arbitrary length. In addition, after an action a_t is executed, a reward $r_t = \rho(s_t, a_t) \in \mathbb{R}$ is observed.

A policy $\pi \in \Pi = \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is a mapping from the state space $\mathcal{S}$ to the set of probability measures on the action space $\mathcal{P}(\mathcal{A})$, where $\pi(a|s)$ is the associated conditional probability density of action a in state s. The function $J : \Pi \to \mathbb{R}$ is defined as the function mapping any policy π to the expected discounted sum of rewards gathered by an agent interacting in the MDP by sampling actions from the policy π. We call return of the policy π the value provided by that function

$$J(\pi) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi,\gamma}(\cdot) \\ a \sim \pi(\cdot|s)}} \left[\rho(s, a) \right] \ , \tag{1}$$

where $d^{\pi,\gamma}(\cdot)$ is the discounted state-visitation probability (Manne, 1960). In reinforcement learning, we seek to find an optimal policy π^* maximizing the expected discounted sum of rewards J.

2.2. Policy-Gradient Algorithms

Policy-gradient algorithms (locally) optimize a parameterized policy π_θ to find the optimal parameter θ^* for which the return of the policy $J(\pi_{\theta^*})$ is maximized. Naively optimizing the parameterized policy by solely maximizing its return may provide sub-optimal results. This problem is mitigated in practice by implementing exploration strategies. These techniques consist in optimizing a surrogate learning objective L that intrinsically encourages certain behaviors. In this work, we consider reward-shaping strategies where the expected discounted sum of rewards is extended by K additional reward terms ρ_i^{int}, called intrinsic motivation terms, and optimize the learning objective

$$L(\theta) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi_\theta, \gamma}(\cdot) \\ a \sim \pi_\theta(\cdot|s)}} \left[\rho(s,a) + \sum_{i=0}^{K-1} \lambda_i \rho_i^{int}(s,a) \right]$$
$$= J(\pi_\theta) + J^{int}(\pi_\theta) \ , \quad (2)$$

where λ_i are non-negative weights for each intrinsic reward and where $J^{int}(\pi_\theta)$ is the intrinsic return of the policy. The parameter maximizing the learning objective is denoted by $\theta^\dagger$, which we distinguish from the optimal policy parameter θ^*. Most of the intrinsic motivation terms can be classified in the two following groups.

Uncertainty-based motivations. It is common to provide bonuses for performing actions that reduce the uncertainty of the agent about its environment (Pathak et al., 2017; Burda et al., 2018; Zhang et al., 2021c). The intrinsic motivation terms are then proportional to the prediction errors of a model of the MDP dynamics. The latter model is usually learned.

Entropy-based motivations. It is also common to provide bonuses for visiting states and/or playing actions that are less likely in histories (Bellemare et al., 2016; Lee et al., 2019; Guo et al., 2021). In this work, we focus on two of these bonuses

$$\rho^s(s,a) = -\log d^{\pi_\theta,\gamma}(\phi(s)) \quad (3)$$
$$\rho^a(s,a) = -\log \pi_\theta(a|s) \ , \quad (4)$$

where $\phi(s)$ is a feature built from the state s. The corresponding intrinsic returns are maximized for policies that visit uniformly every feature, and for policies with uniformly distributed actions in each state, respectively. Note that these rewards require to estimate the distribution over the states and/or actions. Furthermore, they implicitly depend on the policy parameter θ. The second technique is usually referred to as entropy regularization (Williams & Peng, 1991; Haarnoja et al., 2019).

In this work, we consider on-policy policy-gradient algorithms, which were among others reviewed by (Duan et al., 2016) and (Andrychowicz et al., 2020). These algorithms optimize differentiable parameterized policies with gradient-based local optimization. They iteratively approximate an ascent direction $\hat{d}$ relying on histories sampled from the policy in the MDP and update the parameters in the ascent direction, or in a combination of the previous ascent directions (Hinton et al., 2012; Kingma & Ba, 2014). For the sake of simplicity and without loss of generality, we consider that the ascent direction $\hat{d}$ is composed of the sum of an estimate of the gradient of the return $\hat{g} \approx \nabla_\theta J(\pi_\theta)$ and an estimate of the gradient of the intrinsic return $\hat{i} \approx \nabla_\theta J^{int}(\pi_\theta)$. In practice, the first is usually unbiased while the second is computed neglecting some partial derivatives of θ and is thus biased, typically neglecting the influence of the policy on the intrinsic reward.

3. Study of the Learning Objective

In this section, we study the influence of the exploration terms on the learning objective defined in equation (2). We define two criteria under which the learning objective can be globally optimized by ascent methods, and such that the solution is close to an optimal policy. We then graphically illustrate how exploration modifies the learning objective to remove local extrema.

3.1. Policy-Gradient Learning Objective

Policy-gradient algorithms using exploration maximize the learning objective function L, as defined in equation (2). We introduce two criteria related to this learning objective for studying the performance of the policy-gradient algorithm. First, we say that a learning objective L is ϵ-coherent when its global maximum is in an ϵ-neighborhood of the return of an optimal policy. Second, we call learning objectives that have a unique maximum and no other stationary point pseudoconcave.

Coherence criterion. A learning objective L is ϵ-coherent if, and only if,

$$J(\pi_{\theta^*}) - J(\pi_{\theta^\dagger}) \leq \epsilon \ , \quad (5)$$

where $\theta^* \in \text{argmax}_\theta J(\pi_\theta)$ and where $\theta^\dagger \in \text{argmax}_\theta L(\theta)$.

Pseudoconcavity criterion. A learning objective L is pseudoconcave if, and only if,

$$\exists!\ \theta^\dagger : \nabla L(\theta^\dagger) = 0 \wedge L(\theta^\dagger) = \max_\theta L(\theta) \ . \quad (6)$$

If the pseudoconcavity criterion is respected, there is a single optimum, and it is thus possible to globally optimize the learning objective function by (stochastic) gradient ascent (Bottou, 2010)[2]. If the learning objective is furthermore ϵ-coherent, the latter solution is also a near-optimal policy, where ϵ is the bound on the suboptimality of its return.

3.2. Illustration of the Effect of Exploration on the Learning Objective

Exploration is of paramount importance when complex dynamics and reward functions are involved, where many locally optimal policies may exist (Lee et al., 2019; Liu & Abbeel, 2021; Zhang et al., 2021b). In the following, we first define an environment and a policy parameterization

[2]For the sake of keeping discussions simple, the definition of pseudoconcavity is simplified (Mangasarian, 1975), and assumptions discussed in Section 1 to ensure convergence when optimizing Markov chains are neglected.

introduced by Bolland et al. (2023) that will serve as an example where it is possible to graphically illustrate the effect of exploration on the optimization process. For the sake of the analysis, we then represent the learning objectives associated with different exploration strategies, and depict their global and local optima. Learning objectives with a single global optimum respect the pseudoconcavity criterion. In addition, we represent the neighborhood Ω of the optimal policy parameters, such that any learning objective with its global maximum within this region is coherent for a given ϵ. In light of the coherence and the pseudoconcavity criteria, we finally elaborate on the policy parameter computed by stochastic gradient ascent algorithms.

We consider the environment illustrated in Figure 1a where a car moves in a valley. We denote by x and v the position and speed of the car, both forming its state $s = (x, v)$. The valley contains two separate low points, positioned in $x_{initial} = -3$ and $x_{target} = 3$, separated by a peak. The car starts at rest $v_0 = 0$ at the highest low point $x_0 = x_{initial}$ and receives rewards proportional to the depth of the valley at its current position. The reward function is provided in Figure 1b. We consider a policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|\mu_K(s), \sigma)$, namely a normally disturbed proportional controller with $\mu_K(s) = K \times (x - x_{target})$, parameterized by the vector $\theta = (K, \sigma)$. Figure 1c illustrates the contour map of the return of the policy as a function of the parameters K and σ. The optimal parameters are represented by a black dot and correspond to a policy that drives the car so as to pass the peak and reach the lowest valley floor in x_{target}. The green area represents the set of parameters $\Omega = \{\theta' | \max_\theta J(\pi_\theta) - J(\pi_{\theta'}) \leq \epsilon\}$ for $\epsilon = 1$, and is used in the following discussion.

Figure 2 illustrates learning objectives combining the intrinsic rewards defined in equations (3) and (4) for different values of the weights λ_1 and λ_2. Here, the feature from equations (3) is composed of the position $\phi(s) = x$. First, we observe that for weights approaching zero, the parameter $\theta^\dagger$ maximizing the learning objective, represented by a black dot, corresponds to a policy with a high return. More precisely, it is in the green set Ω such that ϵ-coherence is guaranteed for a small value of $\epsilon = 1$. Larger weights require larger values of ϵ for guaranteeing the ϵ-coherence criterion. Nevertheless, when increasing the weights, we also observe that the learning objective eventually becomes pseudoconcave. There appears to be a trade-off between the two criteria. In Figure 2b, we observe that in this environment, it is possible to find a learning objective that respects the pseudoconcavity criterion and the ϵ-coherence criterion for $\epsilon = 1$. Indeed, there is a single global maximum in Figure 2b represented by a black dot that is furthermore part of the set Ω.

Shaping the reward function with an exploration strategy based on the state-visitation entropy appears to be a good solution for optimizing the policy. However, a notable drawback is that the reward depends on the policy and its (gradient) computation requires to estimate a complex probability measure. In this example, the intrinsic reward function itself was estimated by Monte-Carlo sampling for every parameter, which would not scale for complex problems and requires approximations and costly evaluation strategies (Islam et al., 2019). In Appendix A we present an alternative problem-dependent intrinsic reward, independent of the policy parameters and thus simple to compute efficiently, that still respects the pseudoconcavity and ϵ-coherence criteria, and in Appendix B we extend the study to more complex environments where the policy is a deep neural network and the state-visitation probability is approximated.

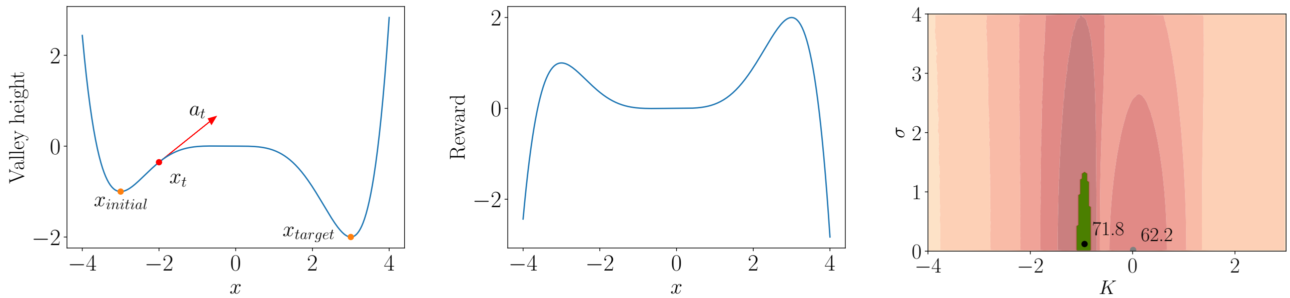
The observations suggest that well-chosen exploration strategies can lead to learning objective functions that satisfy the two criteria defined in the previous section, thereby guaranteeing that policies suboptimal by at most ϵ can be computed by local optimization. When designing exploration strategies, it is essential to keep in mind that we modify the learning objective for the algorithms to converge to optimal policy parameters, which can be achieved when both criteria are respected. While strategies such as enforcing entropy can be effective in some environments, they are only heuristic strategies and not to be relied upon exclusively. Furthermore, as illustrated, both criteria may be subject to a trade-off. In more complex environments, an efficient exploration strategy may require to balance both criteria, e.g., through a schedule on the learning objective weights.

4. Study of the Ascent Direction Distribution

Optimizing pseudoconcave functions with stochastic ascent methods are guaranteed to converge (at a certain rate) under assumptions on the distribution of the gradient estimates at hand (Bottou, 2010; Chen & Luss, 2018; Ajalloeian & Stich, 2020). In this section, we study the influence of the exploration terms on this distribution in the context of policy gradients. More precisely, we study the probability of improving the learning objective, which, intuitively, shall be sufficiently large for the algorithm to be efficient. We formalize this intuition and illustrate how exploration strategies can increase this probability, leading to more efficient policy-gradient methods.

4.1. Policy-Gradient Estimated Ascent Direction

In general, gradient ascent algorithms update parameters in a direction $\hat{d}$ in order to locally improve an objective function f. The quality of these algorithms can therefore be studied (for a small step size $\alpha \to 0$) through the random variable representing the quantity by which the objective



(a) Hill environment. (b) Reward function ρ. (c) Return $J(\pi_{K,\sigma})$.

Figure 1: Illustration of the *hill environment* in Figure 1a and its reward function in Figure 1b. In Figure 1c, the return of the policy $\pi_{K,\sigma}$ with the global and local maximum represented in black and grey, together with their respective return values.

increases

$$X = f(\theta + \alpha\hat{d}) - f(\theta) = \alpha \left\langle \hat{d}, \nabla_\theta f(\theta) \right\rangle , \tag{7}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product. This variable depends on the random event $\hat{d}$ estimated by Monte-Carlo simulations in practice. The (asymptotic) convergence of gradient ascent algorithms can be studied by deriving upper bounds on the expectation of X, which usually involves the parameters of the algorithms. Doing so when gradients are biased is an active research field, where most results do not fit to our study. We therefore instead elaborate directly on the distributions $\mathbb{P}(\|X\|)$ that quantifies the magnitude of the variation of the objective function f, and $\mathbb{P}(X > 0)$ that quantifies when the ascent step improves this objective. In practice, the expectation of the random variable X is positive and the estimate $\hat{d}$ is scaled and clipped by many algorithms, such that the sign of X is arguably of more importance than its norm. In the following, we study $\mathbb{P}(X > 0)$ and assume it to be sufficient to measure the efficiency of optimization algorithms. In other words, we assume that all ascent steps lead to a constant variation of the objective, such that the rate of policy improvement is proportional to $\mathbb{P}(X > 0)$.

In the case of a policy gradient, we first assume that the learning objectives respect the two previous criteria, and introduce two new criteria. The latter are independent (but not mutually exclusive) from those of Section 3. First, we say that an exploration strategy is δ-efficient if, and only if, following the ascent direction $\hat{d} \approx \nabla_\theta L(\theta)$ has a probability at least δ to increasing the learning objective $L(\theta)$ almost everywhere. Second, an exploration strategy is δ-attractive if, and only if, there exists a neighborhood of $\theta^\dagger$ containing the parameter θ^{int} maximizing the intrinsic return J^{int}, where the probability of increasing the return by following $\hat{d}$ is almost everywhere at least equal to δ. Note that each probability measure and random variable is a function of θ, which we do not explicitly write for the sake of keeping notations simple.

Efficiency criterion. An exploration strategy is δ-efficient if, and only if,

$$\forall^\infty \theta : \mathbb{P}(D > 0) \geq \delta , \tag{8}$$

where $D = \langle \hat{d}, \nabla_\theta L(\theta) \rangle$.

Attraction criterion. An exploration strategy is δ-attractive if, and only if,

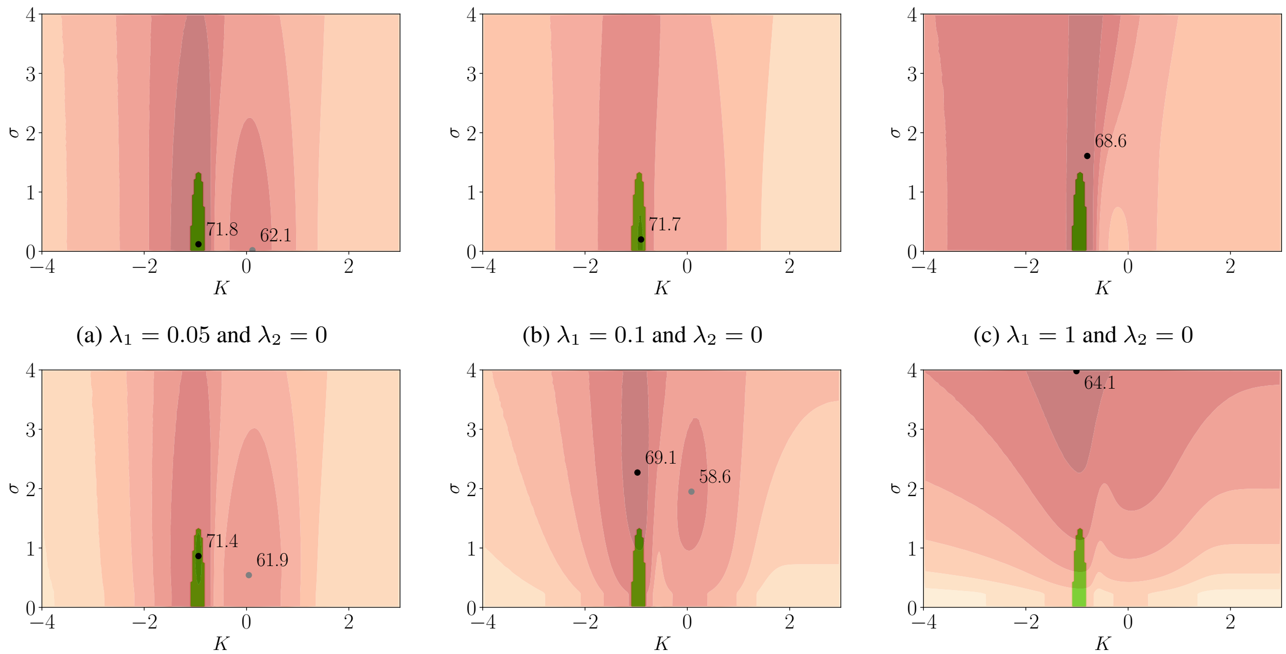
$$\exists B(\theta^\dagger) : \theta^{int} \in B(\theta^\dagger) , \tag{9}$$

such that

$$\forall^\infty \theta \in B(\theta^\dagger) : \mathbb{P}(G > 0) \geq \delta , \tag{10}$$

where $\theta^{int} = \text{argmax}_\theta J^{int}(\pi_\theta)$, $B(\theta^\dagger)$ is a ball centered in $\theta^\dagger$, and $G = \langle \hat{d}, \nabla_\theta J(\pi_\theta) \rangle$.

First, the efficiency criterion quantifies how often a stochastic gradient ascent step is going to improve the learning objective. The larger, the better the learning objective and its stochastic ascent direction approximations. Second, the rationale behind the attraction criterion is that in many exploration strategies, the intrinsic reward is dense, and it is then presumably easy to optimize the intrinsic return in the sense that $\mathbb{P}(\langle \hat{i}, \nabla_\theta J^{int}(\pi_\theta) \rangle > 0)$. It implies that it is easy to locally improve the learning objective by (solely) increasing the value of the intrinsic motivation terms. It furthermore implies that policy-gradient algorithms may be subject to converging towards θ^{int} rather than $\theta^\dagger$ when $\mathbb{P}(\langle \hat{d}, \nabla_\theta J(\pi_\theta) \rangle > 0)$ is small. If the criterion is respected for large δ, the latter is less likely to happen as policy gradients will eventually tend to improve the return of the policy if it approaches θ^{int} and enters the ball $B(\theta^\dagger)$; eventually converging towards $\theta^\dagger$.



(a) $\lambda_1 = 0.05$ and $\lambda_2 = 0$

(b) $\lambda_1 = 0.1$ and $\lambda_2 = 0$

(c) $\lambda_1 = 1$ and $\lambda_2 = 0$

(d) $\lambda_1 = 0$ and $\lambda_2 = 0.01$

(e) $\lambda_1 = 0$ and $\lambda_2 = 0.1$

(f) $\lambda_1 = 0$ and $\lambda_2 = 0.5$

Figure 2: Contour map of (scaled) learning objective functions for different values of λ_1 and λ_2. The darker the map, the larger the learning objective value. The green area represents the set $\Omega = \{\theta' | \max_\theta J(\pi_\theta) - J(\pi_{\theta'}) \leq \epsilon = 1\}$, such that when the parameter maximizing the learning objective is part of Ω, then the learning objective function is ϵ-coherent with $\epsilon = 1$. The black dot is the parameter $\theta^\dagger$ globally maximizing the learning objective and the grey dot is the local (non-global) maximum of the learning objective if it exists. Both are labeled with the return values of the corresponding policies.

4.2. Illustration of the Effect of Exploration on the Estimated Ascent Direction

Exploration is usually promoted and tested for problems where the reward function is sparse, typically in maze-environments (Islam et al., 2019; Liu & Abbeel, 2021; Guo et al., 2021). In this section, we first introduce a new maze-environment with sparse rewards where we illustrate the influence of exploration on the gradient estimates of the learning objective. To this end, we present two learning objective functions and elaborate on the influence of exploration on the performance of policy-gradient algorithms in the light of the efficiency and attraction criteria.

Let us consider a maze-environment consisting of a horizontal corridor composed of $S \in \mathbb{N}$ tiles. The state of the environment is the index of the tile $s \in \{1, \dots, S\}$, and the actions consists in going left $a = -1$ or right $a = +1$. When an action is taken, the agent stays idle with probability $p = 0.7$, and moves with probability $1 - p = 0.3$ in the direction indicated by the action, then $s' = \min(S, \max(1, s + a))$. The agent starts in state $s = 1$ and the target state $s = S = 15$ is absorbing. Zero rewards are observed except when the agent reaches the target state where a reward $r = 100$ is observed. A discount factor of $\gamma = 0.99$ is considered. Finally, we study the policy going with probability θ to the right and probability $1 - \theta$ to the left, and with density

$$\pi_\theta(a|s) = \begin{cases} \theta & \text{if } a = 1 \\ 1 - \theta & \text{if } a = -1 \; . \end{cases} \tag{11}$$

The return $J(\pi_\theta)$ is represented in black in Figure 3a as a function of θ along with two intrinsic returns, $J^a(\pi_\theta)$ in orange and $J^s(\pi_\theta)$ in blue. The intrinsic reward $\rho^a(s, a) = -\log \pi_\theta(a|s)$, from equation (4), and the intrinsic reward $\rho^s(s, a) = -\log d^{\pi_\theta, \gamma}(s)$, from equation (3), are used respectively. In Figure 3b, we illustrate the return of the policy without exploration $J(\pi_\theta)$, along with two learning objective functions, $L^a(\theta)$ and $L^s(\theta)$, using as exploration strategies the intrinsic returns $J^a(\pi_\theta)$ and $J^s(\pi_\theta)$. We observe that the return is a pseudoconcave function with respect to θ and the optimal parameter is $\theta^* = 1$. In addition, the two learning objectives respect the ϵ-coherence criterion for $\epsilon = 0$, implying that $\theta^* = \theta^\dagger$, and respect the pseudoconcavity criterion. It is important to note that with regard to the discussion from Section 3, there is no interest in optimizing the learning objectives rather than directly optimizing the return, as the latter is already pseudoconcave. In the following we illustrate how choosing a correct exploration strategy still deeply influences the policy-gradient algorithms when it comes to building gradient estimates.

Let us compute the estimate $\hat{g}$ and $\hat{d}$ relying on REINFORCE (Williams, 1992) by sampling 8 histories of length $T = 100$. In this particular environment, $\mathbb{P}(D > 0)$ equals $\mathbb{P}(G > 0)$, and equal the probability that the derivative is positive. We represent in Figure 3c this probability for the return and for both learning objectives. First, we see that the learning objectives are more efficient than the return, meaning they are δ-efficient for larger values of δ. Depending on the parameter, the objective $L^a(\theta)$ or $L^s(\theta)$ is best in that regard. Second, concerning the attraction criterion, we represent at the top of Figure 3c the intervals $B^a = [\theta^{int,a}, \theta^{\dagger,a}]$ and $B^s = [\theta^{int,s}, \theta^{\dagger,s}]$. They correspond to the smallest balls containing the maximizers of the intrinsic return and of the learning objective. Let the minima of the orange and blue curves over these intervals be denoted by δ^a and δ^s. By definition of the attraction criterion, it is thus respected for any values of δ at most equal to δ^a and δ^s, for $L^a(\theta)$ and $L^s(\theta)$, respectively. All these observations can eventually be explained as the computation of $\hat{g}$ is always zero when the target is not sampled in the histories, which is highly likely for policies with small values of θ. Adding the exploration terms here leads to policy-gradient algorithms that compute more easily an optimal policy while naive optimization without exploration would fail or be sample inefficient.

We have empirically shown that a well-chosen exploration strategy in policy gradients may not only remove local extrema from the objective function, but may also increase the probability that stochastic ascent steps improve the objective function. Under the previous assumptions, this probability measures the efficiency of algorithms. Furthermore, among different learning objectives respecting the coherence and pseudoconcavity criteria, it is best to choose one that has high values for δ in both the efficiency and attraction criteria. In Appendix A we use these criteria to study other reward-shaping strategies, and in Appendix B we extend the study to more complex environments where the policy is a deep neural network. In the experiments, we used REINFORCE estimates, yet the considerations generalize to any reinforcement learning technique where exploration can help to compute good estimates of the learning objective. Typically, estimating a critic by stochastic gradient ascent suffers from this problem as it is also built from an estimate computed from sampled rewards.

The problem discussed in this section strongly relates to overfitting or generalization in reinforcement learning. In situations where the same state and action pairs are repeatedly sampled with high probability, the policy may appear optimal by neglecting the rewards observed in state and action pairs sampled with low probability. The gradient estimates will then be zero with high probability, and the gradient updates will not lead to policy improvements. In the previous example, gradient estimates computed from policies with a small parameter value θ wrongly indicate that a stationary point has been reached as they equal zero with high probability. We quantify this effect with a novel definition of local optimality. We define as locally optimal policies over a space with probability Δ the policies that maximize the reward on expectation over a set of states and actions observed in a history with probability at least Δ. Formally, a policy π is locally optimal over a space with probability Δ if, and only if,

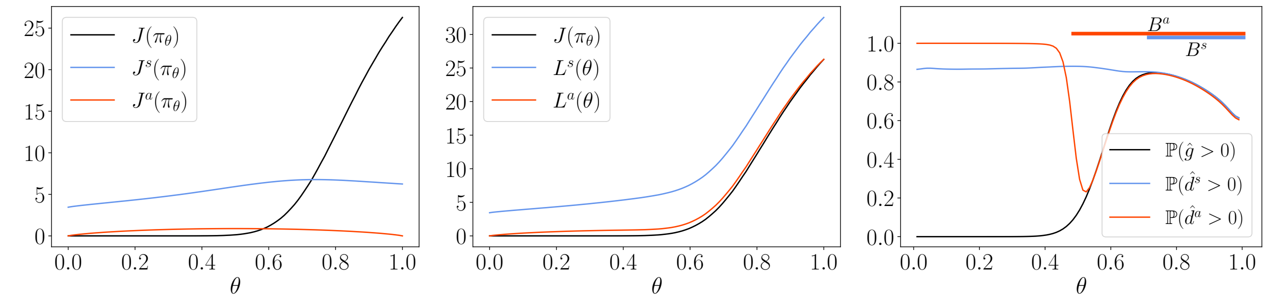
$$\exists\, \mathcal{E} \in \left\{ \mathcal{X} \middle| \int_{\mathcal{X}} d^{\pi,\gamma}(s)\pi(a|s)\, da ds \geq \Delta \right\} :$$
$$\pi \in \underset{\pi'}{\operatorname{argmax}} \int_{\mathcal{E}} d^{\pi',\gamma}(s)\pi'(a|s)\rho(a,s)\, da ds \,. \quad (12)$$

In the typical case of environments with sparse rewards, many policies observe with high probability state and action pairs with zero rewards and are locally optimal for large probabilities Δ. Typically, in the previous example, the joint set $\{1, \dots, S-2\} \times \{-1, 1\}$ is a set of state and action pairs $\mathcal{E}$ that respects the definition equation (12) for policies when θ is small for large values Δ. As we have shown, exploration mitigates the convergence of policy-gradient algorithms towards these locally optimal policies. Note that assuming a non-zero reward is uniformly distributed over the state and action space, exploration policies with uniform probabilities over visited states and actions are the best prior choice for sampling non-zero rewards with high probability. It can thus also be considered as the best choice of exploration to reduce the probability that the stochastic gradient ascent steps do not increase the objective value. Generally, such policy initialization priors may be learned from the framework developed by Lee et al. (2019).

5. Conclusion

In conclusion, this research takes a step towards dispelling misunderstandings about exploration through the study of its effects on the performance of policy-gradient algorithms. More particularly, we distinguished two effects exploration has on the optimization. First, it modifies the learning objective in order to remove local extrema. Second, it modifies the gradient estimates and increases the likelihood that the update steps lead to improved returns. These two phenomena were studied through four criteria that we introduced and illustrated.

These ideas apply to other direct policy optimization algorithms. Indeed, the four criteria do not assume any structure on the learning objective and can thus be straightforwardly applied to any objective function optimized by a direct policy search algorithm. In particular, for off-policy policy



(a) Return functions. (b) Learning objective functions. (c) Improvement probabilities.

Figure 3: Figure 3a represents the return of the policy along with two intrinsic return functions. In Figure 3b the return is also represented together with two learning objective functions, corresponding to the two intrinsic returns. Figure 3c illustrates the probability (estimated by Monte-Carlo) of positive stochastic gradient (derivative) estimates $J(\pi_\theta)$, $L^a(\theta)$, and $L^s(\theta)$. At the top of the figure, the intervals $B^a = [\theta^{int,a}, \theta^{\dagger,a}]$ and $B^s = [\theta^{int,s}, \theta^{\dagger,s}]$ are represented. These intervals represent the smallest balls containing the parameters maximizing the intrinsic return and the learning objective, for both exploration strategies.

gradient, we may simply consider that the off-policy objective is itself a surrogate or that the gradients of the return are biased estimates based on past histories. Ideas introduced in this work also apply to other reinforcement learning techniques. Typically, for value-based RL with sparse-reward environments, convergence towards a value function outputting zero is expected with high probability. This is mostly due to the low probability of sampling non-zero rewards by Monte-Carlo. The discussions from Section 4 then apply, and a similar analysis can be performed.

Our framework opens the door for further theoretical analysis, and the potential development of new criteria. We believe that deriving practical conditions on the exploration strategies, and the scheduling of the intrinsic return, for guaranteeing fast convergence should be the focus of attention. It could be achieved by bounding the policy improvement on expectation, which is nevertheless usually a hard task without strong assumptions. We furthermore believe that we provide a new lens on exploration necessary for interpreting and developing exploration strategies, in the sense of optimizing surrogate learning objective functions.

Acknowledgments

The authors would like thank Arnaud Delaunoy, Pascal Leroy, and Mathias Berger for valuable comments on this manuscript. Adrien Bolland gratefully acknowledges the financial support of a research fellowship of the F.R.S.-FNRS. Gaspard Lambrechts gratefully acknowledges the financial support of the F.R.S.-FNRS for his FRIA grant.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pp. 151–160. PMLR, 2019.

Aittahar, S., Bolland, A., Derval, G., and Ernst, D. Optimal control of renewable energy communities subject to network peak fees with model predictive control and reinforcement learning algorithms. *arXiv preprint arXiv:2401.16321*, 2024.

Ajalloeian, A. and Stich, S. U. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., et al. What matters in on-policy reinforcement learning? A large-scale empirical study. *arXiv preprint arXiv:2006.05990*, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Bedi, A. S., Chakraborty, S., Parayil, A., Sadler, B. M., Tokekar, P., and Koppel, A. On the hidden biases of policy mirror ascent in continuous action spaces. In *International Conference on Machine Learning*, pp. 1716–1731. PMLR, 2022.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. *Advances in Neural Information Processing Systems*, 29, 2016.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Bhatt, S., Koppel, A., and Krishnamurthy, V. Policy gradient using weak derivatives for reinforcement learning. In *Conference on Decision and Control (CDC)*, volume 58, pp. 5531–5537. IEEE, 2019.

Bolland, A., Louppe, G., and Ernst, D. Policy gradient algorithms implicitly optimize by continuation. *Transactions on Machine Learning Research*, 2023.

Bottou, L. Online learning and stochastic approximations. *Online learning in neural networks*, 17(9):142, 1998.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Boukas, I., Ernst, D., Théate, T., Bolland, A., Huynen, A., Buchwald, M., Wynants, C., and Cornélusse, B. A deep reinforcement learning framework for continuous intraday market bidding. *Machine Learning*, 110:2335–2387, 2021.

Brekelmans, R., Genewein, T., Grau-Moya, J., Delétang, G., Kunesch, M., Legg, S., and Ortega, P. Your policy regularizer is secretly an adversary. *arXiv preprint arXiv:2203.12592*, 2022.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Chen, J. and Luss, R. Stochastic gradient descent with biased but consistent gradient estimators. *arXiv preprint arXiv:1807.11880*, 2018.

Chung, W., Thomas, V., Machado, M. C., and Le Roux, N. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In *International Conference on Machine Learning*, pp. 1999–2009. PMLR, 2021.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pp. 1329–1338. PMLR, 2016.

Guo, Z. D., Azar, M. G., Saade, A., Thakoor, S., Piot, B., Pires, B. A., Valko, M., Mesnard, T., Lattimore, T., and Munos, R. Geometric entropic exploration. *arXiv preprint arXiv:2101.02055*, 2021.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2019.

Hinton, G., Srivastava, N., and Swersky, K. Neural networks for machine learning. Lecture 6a. Overview of mini-batch gradient descent, 2012.

Husain, H., Ciosek, K., and Tomioka, R. Regularized policies are reward robust. In *International Conference on Artificial Intelligence and Statistics*, pp. 64–72. PMLR, 2021.

Islam, R., Ahmed, Z., and Precup, D. Marginalized state distribution entropy regularization in policy optimization. *arXiv preprint arXiv:1912.05128*, 2019.

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E., Levine, S., and Salakhutdinov, R. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Liu, H. and Abbeel, P. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021.

Mangasarian, O. L. Pseudo-convex functions. In *Stochastic optimization models in finance*, pp. 23–32. Elsevier, 1975.

Manne, A. S. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

Mei, J., Xiao, C., Dai, B., Li, L., Szepesvári, C., and Schuurmans, D. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33: 21130–21140, 2020.

Mei, J., Dai, B., Xiao, C., Szepesvari, C., and Schuurmans, D. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Neu, G. and Pike-Burke, C. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

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

Tao, R. Y., François-Lavet, V., and Pineau, J. Novelty search in representational space for sample efficient exploration. *Advances in Neural Information Processing Systems*, 33: 8114–8126, 2020.

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

Williams, R. J. and Peng, J. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with reinforce. In *Conference on Artificial Intelligence*, volume 35, pp. 10887–10895. AAAI, 2021a.

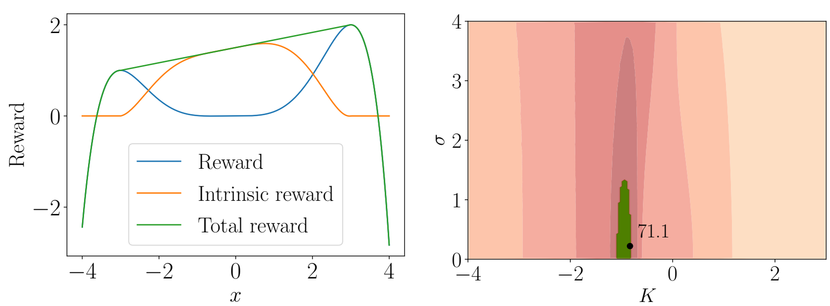
Zhang, T., Rashidinejad, P., Jiao, J., Tian, Y., Gonzalez, J. E., and Russell, S. Made: Exploration via maximizing deviation from explored regions. *Advances in Neural Information Processing Systems*, 34:9663–9680, 2021b.

Zhang, T., Xu, H., Wang, X., Wu, Y., Keutzer, K., Gonzalez, J. E., and Tian, Y. Noveld: A simple yet effective exploration criterion. *Advances in Neural Information Processing Systems*, 34:25217–25230, 2021c.

A. Reward Shaping and Exploration Strategies

As discussed in the manuscript, exploration strategies are reward-shaping strategies where the intrinsic reward bonuses are, among others, dependent on the policy parameters. This dependency makes the shaping strategies adaptive but makes the computation of gradients and the study of the learning objectives more complex. In this section, we study handcrafted reward-shaping strategies to have pseudoconcave and dense reward functions in the hill and maze environments. We then illustrate that the same criteria can be used to study these expert-knowledge based shaped rewards.

For the hill environment from Section 3, we illustrate in Figure 4a an intrinsic reward bonus making the sum of rewards in equation (2) concave. The corresponding learning objective has a unique maximum, which is part of the set $\Omega = \{\theta' | \max_\theta J(\pi_\theta) - J(\pi_{\theta'}) \leq \epsilon\}$ with $\epsilon = 1$ and $\theta = (K, \sigma)$. It can be seen in Figure 4b where the global maximum in black is within the set Ω in green. Both, the ϵ-coherence and the pseudoconcavity criteria are thus respected for $\epsilon = 1$. Here, the intrinsic reward function is a simple function independent of the policy π_θ. Finding such an intrinsic reward may be complex for other environments but the example underlines that exploration and reward shaping are mostly equivalent and that designing reward functions that are concave may help converging towards optimal policies.

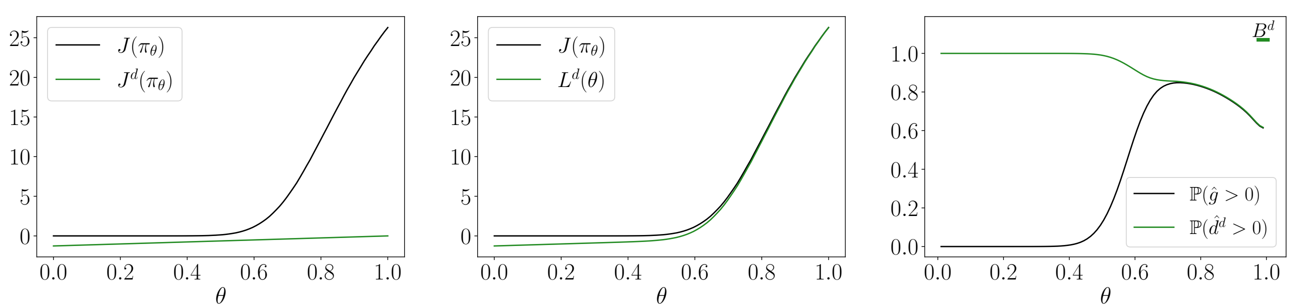


(a) Alternative reward function.

(b) Learning objective.

Figure 4: In Figure 4a, an alternative intrinsic reward function ensuring that the sum of rewards is a pseudoconcave function. In Figure 4b, the contour function of the learning objective.

For the maze environment, the return $J(\pi_\theta)$ is represented in black in Figure 5a together with the intrinsic return $J^d(\pi_\theta)$ in green. The latter is the return of the dense handcrafted reward function $\rho^d(s, a) = (a - 1)/2$ penalizing actions taken from a suboptimal policy. In Figure 5b, the corresponding learning objective function is shown. In the same experimental setting as in Section 4, we observe that the objective function is δ-efficient for higher values of δ compared to the already-discussed learning objectives. Furthermore, the attraction criterion is respected for any value of δ as the unique global maxima of the learning objective, intrinsic return, and return are all equals.



(a) Return.

(b) Learning objectives.

(c) Improvement probabilities.

Figure 5: In Figure 5a the return of the maze environment is represented together with the intrinsic return of a dense handcrafted reward function. Figure 5b represents the corresponding learning objective and Figure 5c the probability that the REINFORCE estimates are positive.

B. Extended Experiments

In this section, we introduce more complex environments and extend the experimental setting using neural networks for parameterizing the policies. In this setting, it is impracticable to compute and represent naively the objective functions and probability distributions from the different criteria. We therefore also introduce a new setting for evaluating the criteria along parameter trajectories.

We consider a passageway environment, inspired by that introduced by Tao et al. (2020), and represented in Figure 6. An agent has to move from rooms to rooms going through passageways and reach a final position. The passageways shall be opened by reaching positions where switches are located. To do so, the agent may choose actions that consist in moving to the four adjacent positions or to stay idle. We consider two reward settings: the dense passageway and the sparse passageway. In the first, rewards of -1 are perceived for every non-idle move, and rewards of 100 are perceived in the target position. In the second setting, zero-rewards are perceived everywhere, except in the target position where a bonus of 100 is provided. We consider a discount factor of $\gamma = 0.98$ and optimize a fully connected neural network of three hidden layers taking as input the position-pair and the switch-state and outputting a categorical distribution over actions.

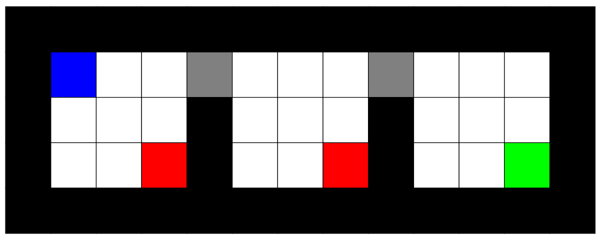
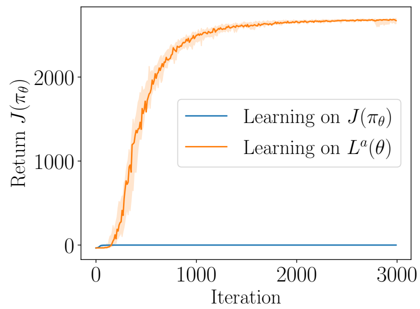


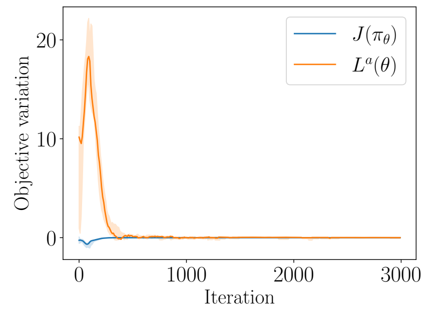
Figure 6: Illustration of the passageway environment. The agent moves within the black walls, starting from the blue tile, it must reach the red switches to open the gray doors, and reach the final green position.

In the dense passageway, we optimize the policy by maximizing three learning objective functions $J(\pi_\theta)$, $L^a(\theta)$, and $L^s(\theta)$. For the latter objective, we use a ten-components Gaussian mixture model over the sampled batch to approximate the state-visitation density. The optimization is performed using the Adam update rule (Kingma & Ba, 2014) with REINFORCE ascent directions computed over 64 histories of constant length $T = 100$. The length T of the histories is chosen such that the realization value T from a geometric distribution with success probability parameter $1 - \gamma$ has at least a cumulative probability of 0.85. We discuss the experimental setting and results for comparing the objective functions $J(\pi_\theta)$ and $L^a(\theta)$. The same setting and results hold for the second learning objective function $L^s(\theta)$, see Figure 7. First, on the one hand, we see in Figure 7a that optimizing the return ends up with high probability in a local optimum, where the policy keeps the agent at the original position. On the other hand, optimizing the learning objective $L^a(\theta)$ allows to converge towards an optimal policy. This learning objective is thus ϵ-coherent with $\epsilon \approx 0$. Second, to asses the pseudoconcavity criterion, we verify if any parameter θ computed during the stochastic gradient ascent steps on the return of the policy $J(\pi_\theta)$ is a local optimum of the learning objective function $L^a(\theta)$. To that end, we perform 5 gradient ascent steps (with the Adam update rule) on the learning objective starting at each parameter θ. In Figure 7b, we represent the expected improvement of the learning objective (in orange) and of the return (in blue). For the first hundreds of parameter values (even after local convergence of the return), it is possible to increase the learning objective, which results in a decrease of return. The policy parameters are thus no local optima of the learning objective. After more iterations, the improvement reaches zero, which is likely an artifact of the optimization algorithm used, namely REINFORCE with the Adam optimizer. These results indicate that, along this parameter trajectory, the return $J(\pi_\theta)$ has a local optimum (or saddle point), in opposition to the learning objective $L^a(\theta)$. The latter illustrates the validity of the pseudoconcavity criterion in that region of the parameter space.

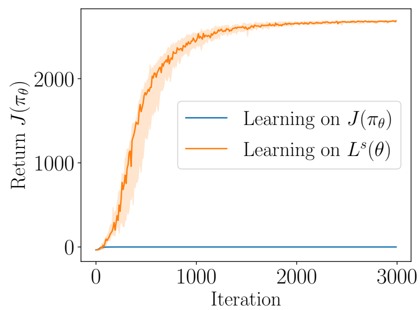
In the previous experiments with the dense-passageway environment, the local optimum exists due to the negative rewards associated to idle-actions. If we consider the sparse-passageway environment, the REINFORCE algorithm will eventually (after a likely long number of iterations) converge towards an optimal policy. There is only a single and global maximum. Yet, if we consider a smaller discount factor $\gamma = 0.95$, the relative importance of future rewards decreases, and by the previous rule of thumb, we may select a REINFORCE horizon of $T = 40$. The likelihood of randomly observing the target state is drastically decreased. On the one hand, the REINFORCE algorithm now fails again to find with high probability an optimal policy. On the other hand, the learning objective function $L^s(\theta)$ with intrinsic exploration rewards proportional to the likelihood of the state-visitation loglikelihood allows to rapidly find an optimal policy. The evolution of the return during optimization is represented in Figure 8a. We illustrate that these results can be justified by the efficiency and attraction



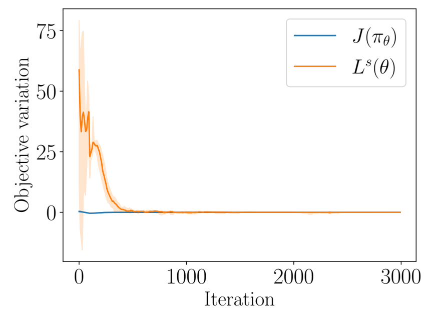
(a) Return during optimization.



(b) Objective functions improvement.



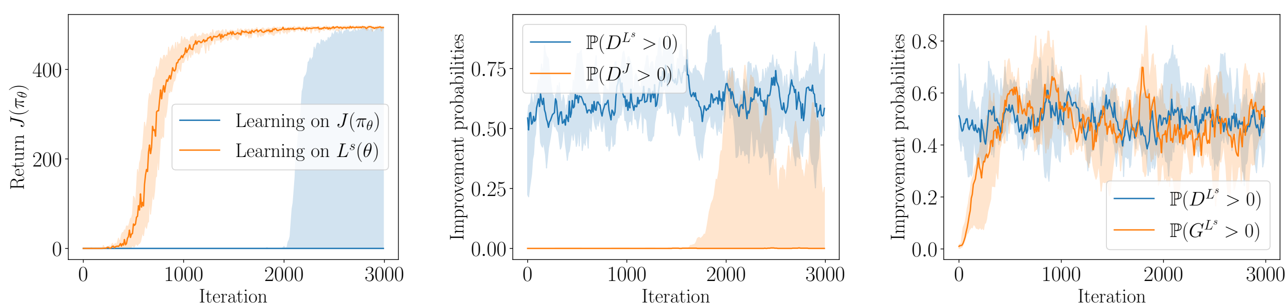
(c) Return during optimization.



(d) Objective functions improvement.

Figure 7: In Figure 7a, the evolution of the return of policies during the optimization is represented in the dense-passageway environment (with $\gamma = 0.98$). In orange, the learning objective $L^a(\theta)$ is optimized and in blue the return $J(\pi_\theta)$ is optimized performing Adam steps in REINFORCE directions. Note that the median, worst and best cases over five runs are represented for the orange curve. For the blue curve, the statistics are computed over the 4 runs that converged towards a return of zero (i.e., towards a non-global optimum). The fifth run, which is not represented, converged similarly to the policies optimized with the learning objective, and is a rare lucky event where no exploration is needed for reaching an optimal policy. Each iteration in Figure 7a corresponds to parameters that were computed with stochastic gradient ascent steps. For each parameter generated by the ascent on the return $J(\pi_\theta)$, we represent in Figure 7b the improvement (on average over 100 simulations) of the objective functions $L^a(\theta)$ and $J(\pi_\theta)$ after 5 Adam steps with REINFORCE directions of the objective $L^a(\theta)$. Similarly to Figure 7a, the median, worst and best cases over the 4 runs are represented. In Figure 7c and Figure 7d, the same experiment is performed using the learning objective $L^s(\theta)$ instead.

criteria. For each parameters obtained during the stochastic ascent steps on the return, we estimate the probability of improving both objective functions by stochastic gradient ascent, and illustrate the efficiency criterion. In Figure 8b, we observe that the improvement probability is negligible for the return. After some iterations, there is a sudden improvement of this probability, resulting from one lucky event out of the 5 where REINFORCE managed to converge towards an optimal policy. On the contrary, the probability of improving the learning objective remains much higher for each parameter. The efficiency of the learning objective with exploration is higher than that of the return. In order to illustrate the attraction criterion, we estimate the probability of improving the return and the learning objective, both by gradient ascent steps on the learning objective, for each parameters obtained during the stochastic ascent steps on the objective $L^s(\theta)$. As can be seen in Figure 8c, the probability of improving the learning objective is again high during the optimization. The probability of improving the return, on the contrary, is small at the beginning and increases after some iterations. This indicates that once the policy has a sufficiently large intrinsic return, the attraction criterion is respected for a high value δ.



(a) Return during optimization. (b) Parameters during ascent on $J(\pi_\theta)$. (c) Parameters during ascent on $L^s(\theta)$.

Figure 8: In Figure 8a, the median, worst and best cases over 5 runs of the evolution of the return during the optimization is represented for the learning objective functions $L^s(\theta)$ and $J(\pi_\theta)$. The optimization is performed by Adam steps in REINFORCE directions in the the sparse-passageway environment with $\gamma = 0.95$. Figure 8b provides the estimated probability of improving the return $J(\pi_\theta)$ and the learning objective $L^s(\theta)$ when following their REINFORCE gradient estimate. This value is estimated at each run of the optimization of the policy with learning objective $J(\pi_\theta)$. Figure 8c provides the estimated probability of improving the learning objective and the return when following the REINFORCE gradient estimate of the learning objective. These values are estimated at each run of the optimization of the policy with the learning objective $L^s(\theta)$. The probabilities were estimated with the frequencies of improving by more than 0.2 the objective functions when following 5 Adam ascent steps using REINFORCE update directions.