Understanding the influence of exploration on the dynamics of policy-gradient algorithms

Adrien Bolland (adrien.bolland@uliege.be)

January 11, 2024

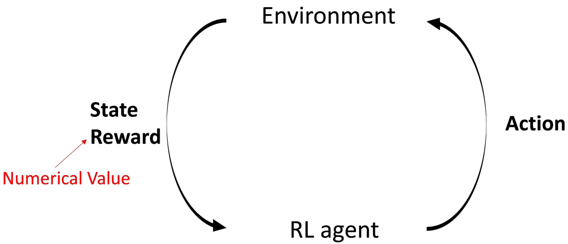
Abstract

- Policy gradients are effective reinforcement learning algorithms.
- During optimization, the policy should remain sufficiently stochastic.
- Why does optimizing stochastic policies perform better ?
- In practice, we achieve this result with intrinsic exploration bonuses.
- How good is this new learning objective ?

Introduction

Reinforcement Learning

Reinforcement learning agents make decisions in a system based on the observed states in order to maximize the expected sum of future rewards gathered.



- Requires an oracle model.
- Differentiates between optimization and execution time.
- Solves offline a nonconvex stochastic optimization problem.

Notations

Some reinforcement learning notations:

- $s \in \mathcal{S}$ for the states,
- $a \in \mathcal{A}$ for the actions,
- $h \in H$ for the histories of states and actions,
- p_0 for the initial state distribution,
- p for the transition distribution,
- ρ for the reward function,
- $\eta(a|h)$ for the history-dependent stochastic policies,
- $\pi(a|s)$ for the stationary Markov stochastic policies,
- $\mu(s)$ for the stationary Markov deterministic policies.

Direct Policy Search – Optimization problem

Definition (Problem Statement)

In direct policy search we look for a policy η^* maximizing the expected discounted sum of rewards (i.e., the expected return of the policy):

$$J(\eta) = \underset{\substack{s_0 \sim p_0(\cdot) \\ a_t \sim \eta(\cdot|h_t) \\ s_{t+1} \sim p(\cdot|s_t, a_t)}}{\mathbb{E}} \left[\sum_{t=0}^{\infty} \gamma^t \rho(s_t, a_t) \right].$$

Policy-gradient algorithms maximize this objective by iterative local optimization of a parametric function, typically a neural network by stochastic gradient ascent.

Direct Policy Optimization

Different types of policies

- Stochastic history-dependent policy $\eta_\theta \in \mathcal{E} = H \to \mathcal{P}(\mathcal{A})$
- Stochastic Markov policy $\pi_\theta \in \Pi = \mathcal{S} \to \mathcal{P}(\mathcal{A}) \subsetneq \mathcal{E}$
- Gaussian policy $\pi^{GP}(a|s) = \mathcal{N}(a|\mu_\theta(s), \Sigma_\theta(s))$
- Deterministic policy $\mu_\theta \in \mathcal{S} \to \mathcal{A} = M \subsetneq \Pi \subsetneq \mathcal{E}$

A policy is said to be affine, if the function approximators used to construct the functional form of the policy are affine functions of the parameter θ.

Direct Policy Search – Requirement

The policy shall remain sufficiently stochastic during the optimization procedure to avoid converging towards a locally optimal solution.

The role of the stochasticity of the policy

Optimization by Continuation

Research question

What is the effect of the choice of the functional parameterization of the policy on the learning objective and how shall it be optimized to converge towards an optimal policy.

Optimization by Continuation

We optimize a surrogate objective function f^p called continuation, of the true objective f, of the form

$$f^p(x) = \mathop{\mathbb{E}}_{y \sim p(\cdot|x)} [g(y)] \ .$$

- g is any function over a latent space $\mathcal{Y}$.
- p is the continuation distribution.
- There exists a p^* for which $f^{p^*} = f$.

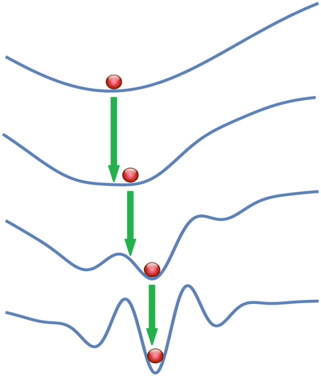
Algorithm 1 Optimization by Continuation

```
Provide a sequence $p_0 \succ p_1 \succ \cdots \succ p_{I-1}$
Provide an initial variable value $x_0^* \in \mathcal{X}$
for all $i = 0, 1, \ldots, I-1$ do
    $x_{i+1}^* \leftarrow$ Optimize the continuation $f^{p_i}$ by local search initialized at $x_i^*$
end for
return $x_I^*$
```

Optimization by Continuation – Illustration

Illustration for Gaussian continuations:

$$f^p(x) = \mathop{\mathbb{E}}_{y \sim \mathcal{N}(\cdot|x,\sigma)} [f(y)] \ .$$



Optimization Policies by Continuation

We define a continuation for the optimization variable $x = \theta$ and the latent variable $y = s_0, \theta_0, a_0, s_1, \ldots,$ where

$$p(y|x) = p(s_0) \prod_{t=0}^{\infty} \eta_{\theta_t}(a_t|h_t)\, q(\theta_t|\theta, \Lambda(h_t))\, p(s_{t+1}|s_t, a_t)$$

$$g(y) = \sum_{t=0}^{\infty} \gamma^t \rho(s_t, a_t)\ .$$

Continuation of the return

The continuation $f_\Lambda^q = f^p$ of the return of the policy $\eta_\theta \in \mathcal{E}$ corresponding to the distribution q and covariance function Λ, is defined $\forall \theta \in \mathbb{R}^{d_\Theta}$ as:

$$f_\Lambda^q(\theta) = \mathop{\mathbb{E}}_{\substack{s_0 \sim p_0(\cdot) \\ \theta_t \sim q(\cdot|\theta, \Lambda(h_t)) \\ a_t \sim \eta_{\theta_t}(\cdot|h_t) \\ s_{t+1} \sim p(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t \rho(s_t, a_t) \right]\ .$$

The continuation converges towards the return of η_θ in the limit as a covariance function Λ approaches zero.

Optimization Policies by Continuation

- Smoothing effect of the continuation through marginalizing the variables.
- Policy parameters may vary differently based on the time step.
- Factorization represents the causal effect of actions.

Result

Show that optimizing by policy-gradient (1) a policy with discounted variance, and (2) a policy with discounted entropy regularization, is equivalent to optimizing the continuation of the return of another policy.

Mirror Policies and Policy Optimization

Definition. We call a mirror policy of the original policy η_θ, under the continuation distribution q and covariance function Λ, any history-dependent policy $\eta'_\theta \in \mathcal{E}$ such that $\forall \theta \in \mathbb{R}^{d_\Theta}$:

$$f^q_\Lambda(\theta) = J(\eta'_\theta) \ .$$

Properties of mirror policies

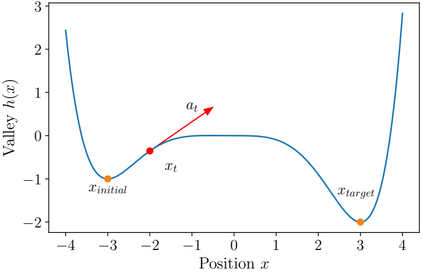
Existence of deterministic original policies

Let $\pi_\theta^{GP'}$ be an affine Gaussian policy with mean function μ_θ, and with covariance function $\Sigma'_\theta = \Sigma'$ constant with respect to the parameters of the policy (i.e., a function depending solely on the state). If $d_{\mathcal{A}} \leq d_\Theta$ and if $\nabla_\theta \mu_\theta(s)$ is full rank, then, there exists a continuation, with covariance Λ proportional to Σ', for which $\pi_\theta^{GP'}$ is a mirror policy of the original policy μ_θ.

If we schedule the variance of this Gaussian policy $\pi_\theta^{GP'}$ and optimize it by stochastic gradient ascent, it is equivalent to optimize the policy μ_θ by continuation.

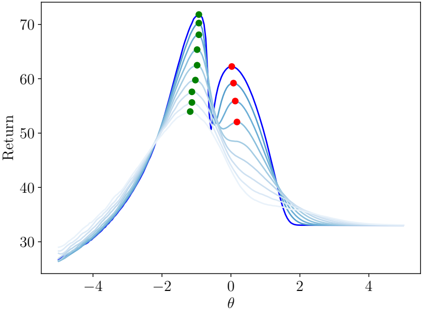
Illustration

We consider a car moving on a double-cliffed valley, and denote by x its position and by v its speed. The car starts in the highest cliff and perceives rewards proportional to the depth in the valley, an optimal sequence of actions would bring the car in the deepest cliff x_{target}.



Illustration

- Directly optimizing a deterministic policy $\mu_\theta(s) = \theta \times (x - x_{target})$ would result in a locally optimal policy.
- We optimize the Gaussian $\pi_\theta^{GP'}(a|s) = \mathcal{N}(a|\mu_\theta(s), \sigma')$ instead.
- $\pi_\theta^{GP'}$ is a mirror policy of μ_θ with continuation variance $\lambda = \sigma'/(x - x_{target})^2$.



A sufficiently large variance removes the local extrema of the the return of the mirror policy.

Discussion

- Exploration in the sense of enforcing the entropy of the policy has a smoothing effect on the return.
- The variance or entropy of the policy is part of the optimization process and shall be adjusted to avoid local extrema and not to locally maximize the return.
- As the variance has a smoothing role, there may be advantages to optimize history-dependent policies.

Learning objective with intrinsic exploration

Intrinsic Exploration

Learning objective

Policy gradient algorithms optimize by SGA the learning objective:

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

- Uncertainty-based motivations where the reward depends on a model prediction error.
- Entropy-based motivations where the reward depends on the state-action probability, typically :

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

Research question

We optimize a surrogate learning objective but we want the final solution computed by (stochastic) gradient ascent to be a near-optimal policy.

Research question

What are the required conditions to compute an optimal policy by (stochastic) gradient ascent on a learning objective ?

Convergence of SGA

Let us assume that we have unbiased gradient estimates of the learning objective function, and that we perform stochastic gradient ascent steps.

- Stochastic gradient ascent is guaranteed to converge towards a local maximum under mild conditions.
- If the function is (pseudo or quasi) concave, stochastic gradient ascent converges towards the global maximum.

Study of the Learning objective – Coherence

1. Coherence criterion

A learning objective L is ε-coherent if and only if

$$J(\pi_{\theta^*}) - J(\pi_{\theta^\dagger}) \leq \varepsilon \, , \tag{1}$$

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

The optimal parameter $\theta^\dagger$ corresponds to a policy at most suboptimal by ε.

Study of the Learning objective – Concavity

2. Pseudoconcavity criterion

A learning objective L is pseudoconcave if and only if

$$\exists!\, \theta^{\dagger} : \nabla L(\theta^{\dagger}) = 0 \wedge L(\theta^{\dagger}) = \max_{\theta} L(\theta)\ . \tag{2}$$

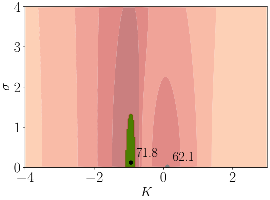
If the pseudoconcavity criterion is respected, there is a single optimum, and it is thus possible to globally optimize the learning objective function by (stochastic) gradient ascent.

Learning objective in the hill environment

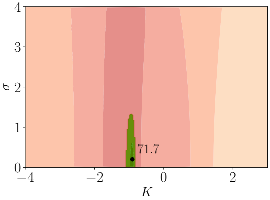
We optimize the policy $\pi^{GP}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

$$L(\theta) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi_\theta, \gamma}(\cdot) \\ a \sim \pi_\theta(\cdot|s)}} \left[\rho(s,a) + \lambda_1 \rho^s(s,a) + \lambda_2 \rho^a(s,a)\right] \ .$$

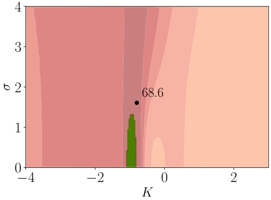
Learning objective in the hill environment



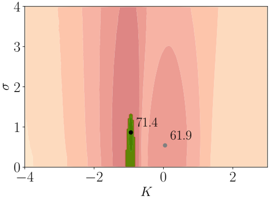
(a) $\lambda_1 = 0.05$ and $\lambda_2 = 0$



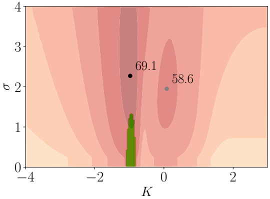
(b) $\lambda_1 = 0.1$ and $\lambda_2 = 0$



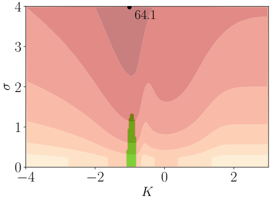
(c) $\lambda_1 = 1$ and $\lambda_2 = 0$



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



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



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

Discussion

- There is a tradeoff between both criteria.
- Balancing the criteria can be achieved by scheduling the weights.
- Entropy bonuses do not hold the same role as in value-based RL.

Related works

- The smoothing effect of entropy regularization has been long known.
- Optimizing entropy regularized objective is equivalent to robust optimization.

Research question

In practice, even pseudoconcave and coherent learning objective functions can be challenging to optimize with stochastic approximations.

Research question

What are the required conditions for exploration to accelerate the convergence speed of SGA ?

Probability of improvement of SGA

The improvement of learning objective f following the update direction $\hat{d}$ is

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

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product.

- The asymptotic convergence is deduced from the expectation of this random variable.
- In practice gradients are biased and the ascent algorithms modify the update directions.

Let us assume that all ascent steps lead to a constant variation of the objective, such that the policy improvement is proportional to $\mathbb{P}(X > 0)$.

Study of the stochastic ascent direction – Efficiency

3. Efficiency criterion

An exploration strategy is efficient if and only if

$$\forall^{\infty}\theta : \mathbb{P}(D > 0) > \mathbb{P}(G > 0)\,, \tag{3}$$

where $D = \langle \hat{d}, \nabla_\theta J(\pi_\theta) \rangle$ and $G = \langle \hat{g}, \nabla_\theta J(\pi_\theta) \rangle$.

Following the ascent direction $\hat{d} \approx \nabla_\theta L(\theta)$ has a higher probability of increasing the return of the policy than following the direction $\hat{g} \approx \nabla_\theta J(\pi_\theta)$.

Study of the stochastic ascent direction – Attraction

4. Attraction criterion

An exploration strategy is δ-attractive if and only if

$$\exists B(\theta^\dagger) : \theta^{int} \in B(\theta^\dagger) \land \forall^\infty \theta \in B(\theta^\dagger) : \mathbb{P}(D > 0) \geq \delta \,, \tag{4}$$

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

If the criterion is respected for large δ, 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 $\theta^\dagger$.

Sparse reward environment – Simple maze

Let us consider a maze environment consisting of a horizontal corridor composed of $S \in \mathbb{N}$ tiles.

- States $s \in \{1, \dots, S\}$ and actions $a \in \{-1\,(Left), +1\,(Right)\}$.
- Start at the first left-most state $s_0 = 1$.
- Stays idle with probability $p = 7/10$.
- Perceives a non-zero reward in the absorbing state $s = S$.

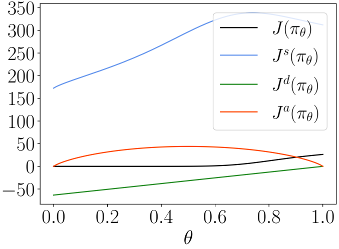
We optimize a one-parameter policy:

$$\pi_\theta(a|s) = \begin{cases} \theta & \text{if } a = 1 \\ 1 - \theta & \text{if } a = -1\,. \end{cases} \tag{5}$$

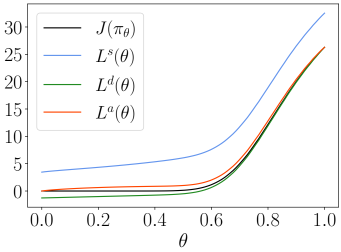
Learning objective functions in the maze

We consider two intrinsic reward bonuses:

$$\rho^s(s,a) = -\log d^{\pi_\theta,\gamma}(s)$$
$$\rho^a(s,a) = -\log \pi_\theta(a|s)$$
$$\rho^d(s,a) = (a-1)/2\ .$$



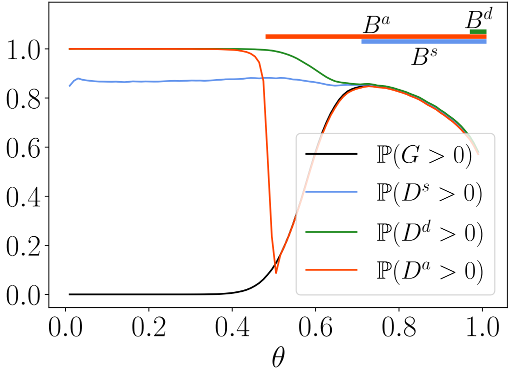
(a) Return



(b) Learning objectives

Probability of improving the return in the maze

Let us compute the probability that the gradient is in the correct direction.



Discussion

- Exploration terms are proxies to have more suited objective functions.
- The analysis is valid for any surrogate learning objective.
- In practice, entropy bonuses have good smoothing properties.
- Exploration is of paramount importance and further research could alleviate some folklore.

References

Adrien Bolland, Gilles Louppe, and Damien Ernst. Policy gradient algorithms implicitly optimize by continuation. *Transactions on Machine Learning Research*, 2023.

Hossein Mobahi and John Fisher III. A theoretical analysis of optimization by gaussian continuation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Adrien Bolland, Garspard Lambrechts, and Damien Ernst. Behind the myth of exploration in policy gradient. 2024.

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