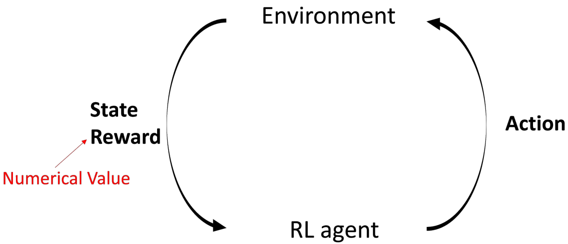
Behind the myth of exploration in policy gradients

Adrien Bolland (adrien.bolland@uliege.be)

February 19, 2024

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,
- p_0 for the initial state distribution,
- p for the transition distribution,
- ρ for the reward function,
- $\pi(a|s)$ for the stationary Markov stochastic policies.

Direct Policy Search – Optimization problem

Definition (Problem Statement)

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

$$J(\pi) = \mathop{\mathbb{E}}_{\substack{s_0 \sim p_0(\cdot) \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim p(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t \rho(s_t, a_t) \right] = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi,\gamma}(\cdot) \\ a \sim \pi(\cdot|s)}} [\rho(s, a)] .$$

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

Direct Policy Search – Requirement

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

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

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

Study of the Learning objective – Concavity

2. Pseudoconcavity criterion

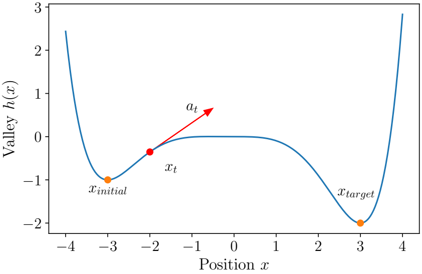
A learning objective L is pseudoconcave if, and only if,

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

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.

Environment with nonlinear dynamics – Hill environment

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}.



Learning objective in the hill environment

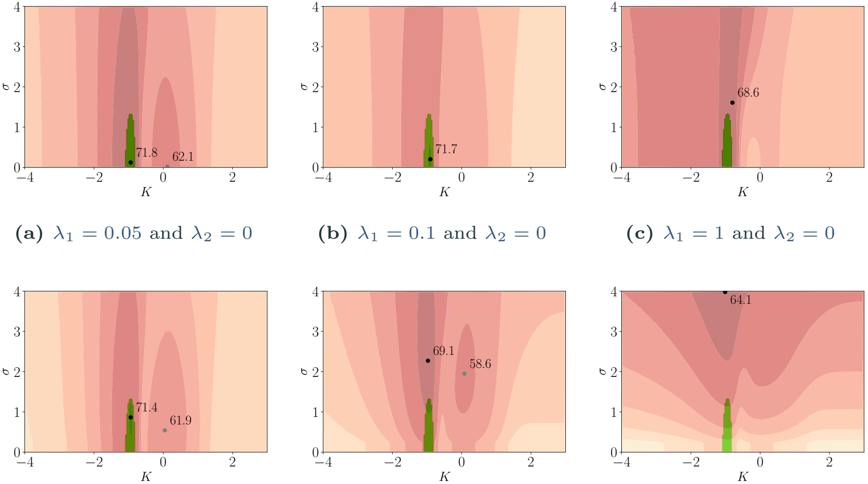
We consider the two intrinsic reward functions

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

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

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

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 δ-efficient if, and only if,

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

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

Following the ascent direction $\hat{d} \approx \nabla_\theta L(\theta)$ has a probability at least δ to improve the learning objective.

Study of the stochastic ascent direction – Attraction

4. Attraction criterion

An exploration strategy is δ-attractive if, and only if,

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

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

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 $s \in \{1, \ldots, S\}$ 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 $s = S$.

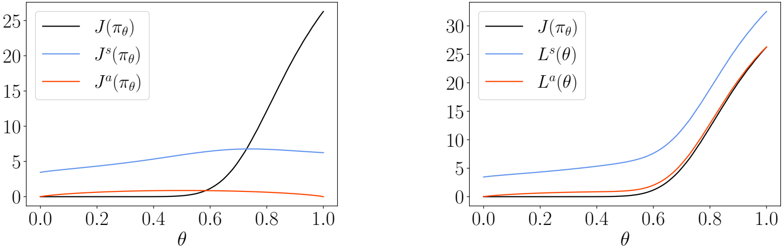
We optimize a one-parameter policy:

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

Learning objective functions in the maze

We consider two intrinsic reward bonuses:

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

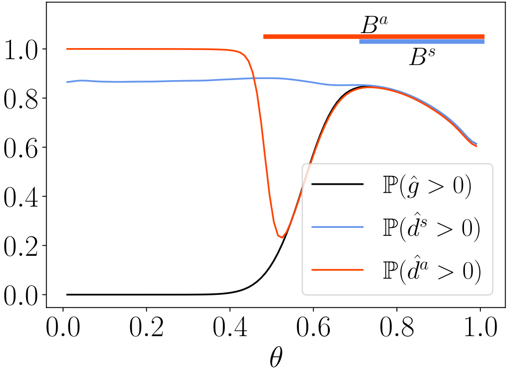


(a) Return

(b) Learning objectives

Probability of improving the return in the maze

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



Conclusion

- 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.