Reinforcement Learning with Rare Significant Events: Direct Policy Search vs. Gradient Policy Search

Paul Ecoffet, Nicolas Fontbonne, Jean-Baptiste André, Nicolas Bredeche

Reinforcement Learning with Rare Significant Events: Direct Policy Search vs. Gradient Policy Search

Paul Ecoffet
Institut des Systèmes Intelligents et de Robotique,
Sorbonne Université
Paris, France
paul.ecoffet@sorbonne-universite.fr

Nicolas Fontbonne
Institut des Systèmes Intelligents et de Robotique,
Sorbonne Université
Paris, France
nicolas.fontbonne@sorbonne-universite.fr

Jean-Baptiste André
Institut Jean Nicod, Département d'Études Cognitives,
École Normale Supérieure
Paris, France
jeanbaptisteandre@gmail.com

Nicolas Bredeche
Institut des Systèmes Intelligents et de Robotique,
Sorbonne Université
Paris, France
nicolas.bredeche@sorbonne-universite.fr

CCS CONCEPTS

• **Theory of computation → Evolutionary algorithms**; • **Computing methodologies → Reinforcement learning**.

KEYWORDS

reinforcement learning, rare significant events, on-policy, on-line, continuous state and action spaces, gradient policy search, direct policy search, evolutionary algorithms, PPO, CMAES

ACM Reference Format:
Paul Ecoffet, Nicolas Fontbonne, Jean-Baptiste André, and Nicolas Bredeche. 2021. Reinforcement Learning with Rare Significant Events: Direct Policy Search vs. Gradient Policy Search. In *Proceedings of the Genetic and Evolutionary Computation Conference 2021 (GECCO '21).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This paper presents a comparison between two methods for on-policy reinforcement learning with continuous state and action spaces, the gradient policy search method PPO [7] and the direct policy search method CMAES [6], for a particular class of reinforcement learning problems, that of **rare significant events** [1, 2, 5]. We consider a setup where significant events present unique opportunities to obtain a positive reward and stop the game, and each opportunity can either be seized for an immediate reward or ignored if the agent hopes to get a better reward in the future.

The problem is that of an agent who has to choose a partner to cooperate with, while a large number of partners are simply *not* interested in cooperating, regardless of what the agent has to offer. Formally, we consider an independent learner $x_\bullet$, called the *focal agent*, which is placed in an aspatial environment. At each time step, $x_\bullet$ is presented with either a *cooperative partner* $x_i^+ \in X^+$ or a *non-cooperative partner* $x_j^- \in X^-$. X^+ (resp. X^-) is the finite set of all cooperative (resp. non-cooperative) partners, with both i and $j \in \mathbb{N}$ and $i > 0, j \geq 0$. When presented with a non-cooperative partner x_j^-, the focal agent's reward will always be zero. When presented with a cooperative partner x_i^+, the focal agent's reward will depend on its own action and that of its partner.

Our objective is to endow the focal agent $x_\bullet$ with the ability to learn how to best cooperate, which implies to negotiate with its potential partners and decide whether cooperation is worth investing energy in, or not. The focal agent faces an individual learning problem as it must optimize its own gain over time in a competitive setup. For cooperation to occur between the focal agent and a partner, the partner must be willing to cooperate (ie. be one of x_i^+) and both the focal agent *and* the cooperative partner must estimate that one's own energy invested in cooperation is worth the benefits. In the setup used, each partner x_i^+ follow a specific *ad hoc* cooperative strategy, some with low expectations, other with high expectations.

The focal agent $x_\bullet$ interacts with the environment in a discrete time manner. At each time step $t = 0, 1, 2, ...$, $x_\bullet$ is in a state $s \in \mathbb{R}$ which describes its current partner's investment value, and plays a continuous value $a \in \mathbb{R}$ which represents its decision to cooperate ($a > 0$) or not ($a <= 0$).

Let π_θ be the parametrised policy of the focal agent, with $\theta \in \mathbb{R}^n$. The learning task is to search for θ^*, such as:

$$\theta^* = \underset{\theta}{argmax} J(\theta) \tag{1}$$

With J the global function to be optimized, defined as:

$$J(\theta) = \mathbb{E}\sum_t r_t \tag{2}$$

with reward r_t at time t. Rewards are defined such that $r \in \mathbb{R}$ and depends on the current state s and action a, and are produced according to the probability generator defined as follow:

$$r(s,a) = \begin{cases} payoff(s,a) & \text{with probability } p \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The probability $p \in [0,1]$ determines the probability to encounter a cooperative agent (i.e. one of x_i^+). The value of p depends on the setup, and determines how *rare* significant events occur

GECCO '21, July 10–14, 2021, Lille, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

when $p < 1.0$. A probability of $p = 1.0$ means the focal agent $x_\bullet$ encounters a cooperative partner at each time step t, with a possible positive reward (if cooperation is accepted by both agents) that depends on the *payoff* function. Non-zero rewards become rarer (but still possible) as $p \to 0$. $payoff(s, a)$ is non-zero *only* if both the focal agent *and* its cooperative partner accept to cooperate, with the exact payoff value depending on the amount of energy the focal agent and its partner invest in cooperation (more details in [4]).

The problem presented here is similar to that of Rare Significant Events as formulated in [5]. However, it differs on two aspects. Firstly, we consider on-line on-policy search of a parametrised policy, where the frequency of significant events cannot be controlled. Secondly, and even more importantly, a learning episode stops right after the focal agent and one cooperative agent have reached a consensus to cooperate. If no cooperation is triggered, an episode stops after a maximum number of iterations T, defined as:

$$T = \frac{100}{p} \text{ time steps} \tag{4}$$

It results that the expected number of significant events M is held constant independently from the value of p (i.e. $\mathbb{E}(M) = 100$). It is therefore possible to obtain episodes of different lengths but with the same number of significant events.

2 RESULTS

Both CMAES and PPO algorithms are used to learn the parameters of the focal agent's decision module, which is used to decide whether to cooperate or not based on the current partner's offer. CMAES and PPO are used to optimize a multi-layered Perceptron of 34 weights. We also use PPO with a deep neural network of 133894 dimensions, which could possibly benefit from over-parametrization [3].

Performance of the current policy is plotted every 4000 iterations, which corresponds to the batch size used by both PPO instances for learning. As episodes last significantly shorter than 4000 iterations this means the policy's performance is averaged. For CMAES, we extract the best policy of the current generation and re-evaluate it 10 times (i.e. for 10 episodes) to get a similarly averaged performance. Results are shown on figures with a data point every 1000 episodes.

Figure 1 show the performance of the agent throughout its learning with both PPO algorithms (termed PPO-MLP and PPO-DEEP) and the CMAES algorithm for different conditions of rare significant events ($p \in \{0.1, 0.2, 0.5\}$), as well as with the control condition when all events are significant ($p = 1.0$, taken from the previous Section). Each figure shows the mean performance of 24 independent runs per conditions, compiling each setup by tracing the median performance and 95% confidence interval from the 24 runs.

All three algorithms provide excellent and comparable results when only significant events are experienced ($p = 1.0$). However, results differ when significant events become rarer (i.e. $p < 1.0$). On the one hand, CMAES is only marginally impacted, with all setups showing convergence towards a similar performance value close to the optimal (above 40). On the other hand, PPO-DEEP and PPO-MLP are both are largely affected, with an even greater toll for PPO-MLP when $p << 1.0$. In the extreme case where $p = 0.1$, the average performance of 35.7 ± 5.2 for PPO-DEEP and 24.9 ± 4.2 of PPO-MLP, to be compared to 46.2 ± 3.2 for CMAES.

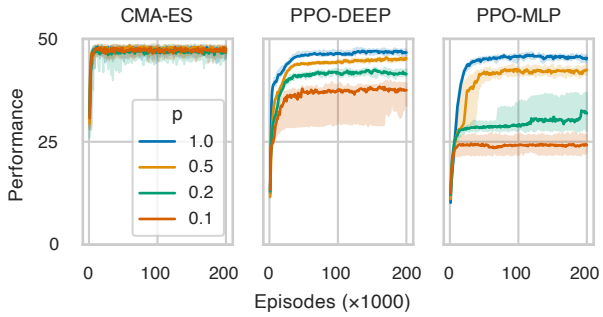

Figure 1: Performance of the best policies (median and 95% confidence interval) during learning with CMAES, PPO-DEEP and PPO-MLP for 3 conditions with rare significant events ($p \in \{0.1, 0.2, 0.5\}$) and 1 control condition ($p = 1.0$)

3 CONCLUDING REMARKS

While both methods provide similar results when the agent is always presented with significant events, policy search methods are not equals when such events become rarer. While the direct policy method is oblivious to rarity of significant events, the gradient policy search method (at least in its PPO implementation) suffers significantly from rarity.

The robustness of the direct policy search method can be expected as the sequential and temporal aspects of the task is lost within one episode. This is of course different for the gradient policy search method, where increased rarity means that many learning steps will be performed with zero-reward, resulting in poor and possibly misleading gradient information.

A comprehensive version of this work is available on Arxiv [4]. Source code is available on Github: https://github.com/PaulEcoffet/RLCoopExp/releases/tag/v1.1.

ACKNOWLEDGMENTS

This work is funded by ANR grant ANR-18-CE33-0006. Thanks to Yann Chevaleyre, Olivier Sigaud and Mathieu Seurin for feedbacks.

REFERENCES

[1] Shalabh Bhatnagar, Vivek S Borkar, and Madhukar Akarapu. 2006. A simulation-based algorithm for ergodic control of Markov chains conditioned on rare events. *Journal of Machine Learning Research* 7, Oct (2006), 1937–1962.

[2] Kamil Andrzej Ciosek and Shimon Whiteson. 2017. OFFER: Off-Environment Reinforcement Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder P. Singh and Shaul Markovitch (Eds.). AAAI Press, 1819–1825.

[3] Simon Du and Jason Lee. 2018. On the Power of Over-parametrization in Neural Networks with Quadratic Activation. In *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 1329–1338.

[4] Paul Ecoffet, Nicolas Fontbonne, Jean-Baptiste André, and Nicolas Bredeche. 2021. Policy Search with Rare Significant Events: Choosing the Right Partner to Cooperate with. (2021). arXiv:cs.LG/2103.06846

[5] Jordan Frank, Shie Mannor, and Doina Precup. 2008. Reinforcement Learning in the Presence of Rare Events. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, New York, NY, USA, 336–343.

[6] Nikolaus Hansen and Andreas Ostermeier. 2001. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation* 9, 2 (2001), 159–195.

[7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347