Paper Review

Policy Gradient Methods for Reinforcement Learning with Function Approximation
by Sutton et al.

Assael Jeremi, Auriau Vincent and Doyen Baptiste

OMA - CentraleSupélec - March 2019

Abstract

In this paper review project, we aim at present a fundamental theoritical paper in RL (Reinforcement Learning) published in NIPS 2000 titled *Policy Gradient Methods for Reinforcement Learning with Function Approximation* and authored by Sutton et al.

1 Introduction

RL methods aims at finding the best policy π^* over a certain policy space for a agent interacting with an environement. In general, these methods can be found in one of this category :

- Policy-based (direct method): The policy parameters are directly optimized by performing gradient ascent on a given agent objective function. One of its main drawback is that the method "learns" very slowly. Indeed, parameters updating is only done at the end of an episode and requires then a much more samples to converge towards an optimal policy.
- Value-based (undirect method): also called Q-learning, the method consists in value function approximation and aims at learning an approximate solution to the Bellman equation. This method is indirect because it does not try to optimize a policy over a policy space. One of its main drawback is the lack of guarantee that the method converges towards a global or a local optimum in the policy space.

The reviewed paper suggests to adopt an hybrid approach by combining these two kinds of methods into a single method one may call "actor-critic". With this method, a policy optimization scheme is adopted in the same time of a value function approximation. Furthermore, the authors show for the first time that a policy iteration method converges to a locally optimal policy and this with arbitrary differentiable function approximation. This results then overcomes one of the main drawback of a value-based only method mentioned above.

2 Theoritical results

In the following sections, we consider in accordance with the paper, the standard RL framework in which an agent interacts with a given MDP (Markov Decision Process).
We denote the state, action and reward at each time t by $s_t \in S$, $a_t \in A$ and $r_t \in \mathbb{R}+$. The state transition prabilities are $P^a_{ss'} = P(s_{t+1} = s'|s_t = s, a_t = a)$ and the expected rewards are $R^a_s = E[r_{t+1}|s_t = s, a_t = a]$. Finally, we denote the policy at each time conditionned by a set of paremeters θ by : $\pi(a, s, \theta) = P(a_t = a|s_t = s, \theta)$.
As in the article, the θ parameters are omitted from the usual notation $\pi(s, a)$

2.1 Further developments on previous methods

In a value-based approach, the policy selected by the method is the "greedy" policy with respect to the estimated value function.
If we denote, $\hat{Q}_k$ the estimated action-value function at a certain step k, then π_{k+1} the next step selected policy is the following : $\pi_{k+1}(s) = argmax_{a\in A}\hat{Q}_k(s, a)$.
Hence π_{k+1} acts greedy with respect to π_k (the selected policy at step k) and is deterministic. Nevertheless, it is in general more relevant to look for an optimal policy in the space of probabilistic policies rather than in the space of deterministic polices.
Indeed, according to Singh, Jaakkola, and Jordan (NIPS 1994), stochastic policies can yield considerably higher expected rewards ($E[\sum_{t=0}^{\infty} \gamma^t r_t]$ with γ the usual discount factor) than deterministic policies in the case of POMDP's settings (Partially Observable Markov Decision Process, a case of MDP more relevant to model real-world decision processes).

Thus, because of the only-deterministic research conducted by a value-based approach, using a policy gradient method seems inevitable. More precisely, if we constrain the policy space to a parametric space of function and denote by θ these parameters and ρ a policy performance mesure, the policy gradient method is a gradient ascent of the following form : $\theta_{t+1} = \theta_t + \alpha\frac{\partial\rho}{\partial\theta}$ where α is a positive coefficient rather small (the learning-rate).
The problem with this method is the slow "learning" performed by the agent. This is due to a high variance of gradient estimate.

Indeed, as we are going to see it later with the first theorem shown by this paper, we can express $\frac{\partial\rho}{\partial\theta}$ as $E_{\tau\sim p(\theta,\tau)}[g(\theta,\tau)]$ where τ represents a trajectory (the sequence of states s_t and actions a_t), p is a certain probability distribution and g a certain function.
The typical way to compute this gradient is then to sample a certain amount of trajectories $\tau \sim p(\theta,\tau)$ and average their $g(\theta,\tau)$ values.
If we call N this amount, then $E_{\tau\sim p(\theta,\tau)}[g(\theta,\tau)] \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta,\tau_i)$. As a result, the gradient is going to have variance, since its values depend on the sampled trajectories τ_i. This variance can be high. Results from sampling can indeed take very different values.
A simple way to reduce this variance would be to increase N so as to have an estimate closer to the expected value. Nevertheless, this means to increase computational cost and we can't afford N to be too high.
Another way, and this is precisely the way suggested by the article, would be to substract a certain baseline function depending only on states $s \in S$. For a well chosen baseline, this can reduce greatly the gradient variance and thus improve comvergence with policy gradient method.

Taking into account all the above explanations, the studied paper opts for the gradient policy approach and suggets a way to compute this gradient with function approximation. These results are presented in the following section with 2 theorems proven into the paper.

2.2 Established Theorems of the Paper

The first theorem of the paper is the generic gradient policy theorem.
The setting is the following : for the reward function ρ, two formulations are possible. One of them is called the "average-reward" and the other the "start-state". The paper studies and proves the result for the two formulations. Concerning this paper review, we will focus only on the "start-state" formulation for simplicity concerns and accordance with class materials.
In this formulation, $\rho(\pi) = E[\sum_{t=1}^{\infty}\gamma^{t-1}r_t|s_0,\pi]$ where s_0 is the initial state.
We define also the state-value function by :
$Q^{\pi}(s,a) = E[\sum_{k=1}^{\infty}\gamma^{k-1}r_{t+k}|s_t = s, a_t = a, \pi]$ and the discounted weighting of states: $d^{\pi} = \sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0,\pi)$. The first theorem is the following :

Theorem 1 (Policy Gradient) : For any MDP and the two performance formulations mentionend above :

$$\frac{\partial\rho}{\partial\theta} = \sum_s d^{\pi}(s)\sum_a \frac{\partial\pi(s,a)}{\partial\theta}Q^{\pi}(s,a) \tag{1}$$

Proof (start-state formulation only) :

(1) We first express $\frac{\partial V^\pi(s)}{\partial \theta}$ for any $s \in S$
By definition, $V^\pi(s) = E[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, \pi]$.
Hence, according to the law of total expectation :

$$\sum_a \pi(a,s) Q^\pi(a,s) = \sum_a \pi(a,s) E[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, a_t = a, \pi]$$

$$= E[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, \pi] = V^\pi(s)$$

Furthermore,

$$Q^\pi(a,s) = E[\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, a_t = a, \pi]$$

$$= R_s^a + \gamma E[\sum_{k=2}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, a_t = a, \pi]$$

$$= R_s^a + \sum_{s'} \gamma P_{ss'}^a V^\pi(s')$$

Then,

$$\frac{\partial V^\pi(s)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_a \pi(a,s) Q^\pi(a,s)$$

$$= \sum_a \frac{\partial \pi(a,s)}{\partial \theta} Q^\pi(a,s) + \pi(a,s) \frac{\partial Q^\pi(s,a)}{\partial \theta}$$

$$= \sum_a \frac{\partial \pi(a,s)}{\partial \theta} Q^\pi(a,s) + \pi(a,s) \frac{\partial}{\partial \theta} (R_s^a + \sum_{s'} \gamma P_{ss'}^a V^\pi(s'))$$

$$= \sum_a \frac{\partial \pi(a,s)}{\partial \theta} Q^\pi(a,s) + \pi(a,s) (\sum_{s'} \gamma P_{ss'}^a \frac{\partial}{\partial \theta} V^\pi(s'))$$

We can notice that the last equation has the form of a recursion on $\frac{\partial}{\partial \theta} V^\pi(s)$. The idea of the last part of the proof is then to exploit this recursive form and "unroll" the equations as mentionned by Sutton et al. in their article.

For this, we inroduce $P(s \to x, k, \pi)$, the probability from going to state x from state s, in k setps and under the π policy. We can also define this probability recursively: going from s to x in $k+1$ steps is equivalent to go from s to s' in k steps where s' is a one-step close to x state. Due to markovian property, we have the following recursive formula :

$$P(s \to x, k+1, \pi) = \sum_{s'} P(s \to s', k, \pi) P(s' \to x, 1, \pi)$$

$$P(s \to x, 0, \pi) = 1 \text{ and } P(s \to x, 1, \pi) = \sum_{s'} \pi(a,s) P_{ss'}^a$$

(2) We keep this result for the moment and go back to the previous equations For more simplicity, we also introduce another notation : $\phi(s) = \sum_a \frac{\partial \pi(a,s)}{\partial \theta} Q^\pi(a,s)$

$$\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} &= \phi(s) + \sum_a \pi(a,s)(\sum_{s'} \gamma P^a_{ss'} \frac{\partial}{\partial \theta} V^\pi(s')) \\
&= \phi(s) + \sum_{s'} \sum_a \gamma \pi(a,s) P^a_{ss'} \frac{\partial}{\partial \theta} V^\pi(s') \\
&= \phi(s) + \sum_{s'} \gamma P(s \to s', 1, \pi) \frac{\partial}{\partial \theta} V^\pi(s')
\end{aligned}$$

In the same way, we can now express $\frac{\partial}{\partial \theta} V^\pi(s')$ using the above equation and the recursive formula of $P(s \to x, k, \pi)$, which gives :

$$\frac{\partial V^\pi(s)}{\partial \theta} = \phi(s) + \sum_{s'} \gamma P(s \to s', 1, \pi)[\phi(s') + \sum_{s''} \gamma P(s' \to s'', 1, \pi) \frac{\partial}{\partial \theta} V^\pi(s'')]$$

$$= \phi(s) + \sum_{s'} \gamma P(s \to s', 1, \pi) \phi(s') + \sum_{s''} \gamma P(s \to s'', 2, \pi) \frac{\partial}{\partial \theta} V^\pi(s'')$$

$$= \phi(s) + \sum_{s'} \gamma P(s \to s', 1, \pi) \phi(s') + \sum_{s''} \gamma^2 P(s \to s'', 2, \pi) \phi(s'') + \sum_{s'''} \gamma P(s \to s''', 3, \pi) \frac{\partial}{\partial \theta} V^\pi(s''')$$

$$= ...$$

$$= \sum_{x \in S} \sum_k \gamma^k P(s \to x, k, \pi) \phi(x)$$

(3) We have then expressed $\frac{\partial V^\pi(s)}{\partial \theta}$ explicitly. To finish the proof, we express $\frac{\partial \rho}{\partial \theta}$ in function of $\frac{\partial V^\pi(s)}{\partial \theta}$ for a particular $s \in S$:

$$\begin{aligned}
\frac{\partial \rho}{\partial \theta} &= \frac{\partial}{\partial \theta} E[\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_0, \theta] = \frac{\partial V^\pi(s_0)}{\partial \theta} \\
&= \sum_{x \in S} \sum_k \gamma^k P(s_0 \to x, k, \pi) \phi(x) \\
&= \sum_{x \in S} \sum_k \gamma^k P(s_0 \to x, k, \pi) \sum_a \frac{\partial \pi(a,x)}{\partial \theta} Q^\pi(a,x) \\
&= \sum_{x \in S} d^\pi(x) \sum_a \frac{\partial \pi(a,x)}{\partial \theta} Q^\pi(a,x)
\end{aligned}$$

The first theorem gives us the direction to look at when performing gradient ascent. In order to compute this direction, we need to access to $Q^{\pi}(a,s)$. For that and with respect to the Q-learning method, we introduce a certain parametrized approximator $f_w : S \times A \rightarrow \mathbb{R}$.
Now we would like to replace $Q^{\pi}(a,s)$ in (1) by its approximator f_w. Is it possible or do we need to make further assumtions on f_w ? The reviewed paper proves that under certain realistic assumptions, the approximator can indeed replace the Q-function in the previous gradient expression.
Let $\hat{Q}^{\pi}(a,s)$ be an unbiased approximation of $Q^{\pi}(a,s)$. We can determine f_{w^*} the best approximator as $f_{w^*} = argmin_w[\hat{Q}^{\pi}(a,s) - f_w]^2$.
Which is equivalent to perform gradient descent on $w \in W$ space with descent direction : $[\hat{Q}^{\pi}(a,s) - f_w]\frac{\partial f_w}{\partial w}$. After convergence, we obtain the following : $[Q^{\pi}(a,s) - f_w*]\frac{\partial f_w*}{\partial w} = 0$
Thus the following expression and the second theorem of the paper:

$$\sum_{x \in S} d^{\pi}(x) \sum_{a} \pi(a,x)[Q^{\pi}(a,s) - f_w*]\frac{\partial f_w*}{\partial w} = 0 \tag{2}$$

We also want to update w parameters in a way that is similar to update θ parameters. For that the article looks for a parametrization that satisfies :

$$\frac{\partial f_w}{\partial w} = \frac{\partial log \pi}{\partial \theta} = \frac{\partial \pi}{\partial \theta} \times \frac{1}{\pi} \tag{3}$$

Theorem 2 (Policy Gradient with Function Approximation) If f_w (the best approximator) is determined using the above convergence and then respects (2) and is also compatible with policy parametrization mentionned in (3) :

$$\frac{\partial \rho}{\partial \theta} = \sum_{s} d^{\pi}(s) \sum_{a} \frac{\partial \pi(s,a)}{\partial \theta} f_w(s,a) \tag{4}$$

Proof : the proof is here pretty straight-forward. Indeed, we replace in (2) the expression of $\frac{\partial f_w*}{\partial w}$ by its policy π and policy gradient formulation in (3).
We then substract the obtained equality (which is 0) from the first theorem in (1) and we obtain the desired result of theorem 2.

As a commentary, the paper states that adding any $v(s)$ function to f_w where $s \in S$, does not affect the (4) result. Indeed, $\sum_a \frac{\partial \pi(a,s)}{\partial \theta} = 0$ due to probability normalization. Nevertheless, the choice of the function v can affect the variance of the gradient estimator and thus the convergence of gradient policy method.

These two theorems now give us the theoritical framework to compute policy gradient with function approximation. The next section describes more in details how the article presents the concrete implementation in a algorithmic method.

3 Algorithmic method

The following and last theorem of the paper states for the first time that a gradient policy method helped with function approximation is convergent. The optimum obtained is not global so at least local.
The result also gives the updates rule on both parameters w and θ used to paremeterize respectively f_w and policy π, in accordance with the 2 previous theorems.

Theorem 3 (Policy Iteration with Function Approximation) We consider f_w and π any differentiable function and policy that satisfy (4). We also consider that the Hessian of π is bounded and that the considered MDP has bounded rewards too.
For a specific $\{\alpha_k\}_k$ series of step-size that verify $\alpha_k \to 0$ and $\sum_k \alpha_k = \infty$ (for instance $\alpha_k = \frac{1}{k}$), if we denote $\pi_k = \pi(.,.,\theta_k)$ and perform the following updates:

$$w_k = w \text{ where } \sum_{x \in S} d^{\pi_k}(x) \sum_a \pi_k(a,x)[Q^{\pi_k}(a,xl) - f_w] \frac{\partial f_w}{\partial w} = 0 \quad (5)$$

$$\theta_{k+1} = \theta_k + \alpha_k \sum_s d^{\pi_k}(s) \sum_a \frac{\partial \pi_k(s,a)}{\partial \theta} f_{w_k}(s,a) \quad (6)$$

Then, $\frac{\partial \rho(\pi_k)}{\partial \theta} \to 0$ and policy gradient method converges.

Proof : Using theorems 1 et 2, we obtain that update direction of θ_k in this way is the gradient direction.
Furthermore, as the hessian of π and rewards are bounded, the hessian of ρ is then also bounded. We can refer to the link between $\frac{\partial \rho}{\partial \theta}$ and $\frac{\partial \pi}{\partial \theta}$ proven in theorem 1 to establish that.
Then the paper invokes a theorem from a textbook written by Bertsekas and Tsitsiklis (1996) which we haven't found on the web for free.
Nevertheless, if we also assume that $\sum_k \alpha_k^2 < \infty$ (the harmonic series complies) then, stochastic approximation theory gives us the result that the method converges almost surely (the steps are large enough to counter fluctuations and outliers and also small enough to make the update finer). Blum and Julius R. proved this result in "Approximation Methods which Converge with Probability one" (1954).

4 Discussion and Developements

First we propose to go a little bit further with theorem 1. Indeed, using the "log-trick", we have that $\frac{\partial log\pi}{\partial\theta} = \frac{\partial\pi}{\partial\theta} \times \frac{1}{\pi}$ which we can plug into (1) :

$$\begin{aligned}\frac{\partial\rho}{\partial\theta} &= \sum_s d^\pi(s) \sum_a \pi(s,a) \frac{\partial log\pi(s,a)}{\partial\theta} Q^\pi(s,a) \\ &= E_{s\sim d^\pi;a\sim\pi_\theta}[Q^\pi(s,a)\frac{\partial log\pi(s,a)}{\partial\theta}]\end{aligned}$$

Then we justify here the assumptions made in the 2.1 section and we can also see how we could perform stochastic gradient descent by estimating the mean with a random realization at each step.
More generally, theorem 1 as stated above is the foundation of many RL policy gradient algorithms developped later such as REINFORCE (Monte-Carlo policy gradient) ; Actor-critic ; A3C (Asynchronous Advantage Actor-Critic) ; DPG (Deterministic Policy Gradient) ; DDPG (Deep Deterministic Policy Gradient) and so on...

To discuss the article, we may say that the authors do not give concrete examples where their presented method is implemented and benchmarked numerically with other methods.
Furthermore, the convexity or non-convexity of the problem isn't mentionned and the article seems only to focus on the finding of local optimum policy and not global optimum policy.

5 Conclusion

To conclude, we suggest to go further by reading this article : **A Natural Policy Gradient** by Sham Kakade. The article gives more concrete examples on simple MDPs and test also the method on a more challenging MDP : the Tetris game.