Reinforcement Learning TP3 : Reinforcement Learning with Function Approximation

Plassier Vincent

1 Experiments

Thanks to our different experiences, we find that the REINFORCE algorithm works better for α_t defined by Adam's method. Indeed, for a constant step or an annealing method, the iterations oscillate without converging. The gradient compute has a large variance, so it is necessary to choose carefully α_t.
When parameter N is increased, the variance of the gradient decreases. This provides better accuracy but increases the complexity of the algorithm. Therefore, we must find a compromise between speed and precision.

The figure below was obtained by averaging 5 experiments with $n\_itr = 100$ policy parameters updates and $N = 60$. The curves represent the average reward in function of the number of policy parameters updates and also the evolution of θ_t :

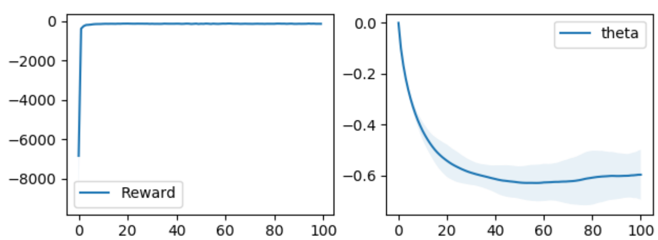

FIGURE 1: Adam's method

Now, we show the results obtained with $\alpha_t = 0.00001$:

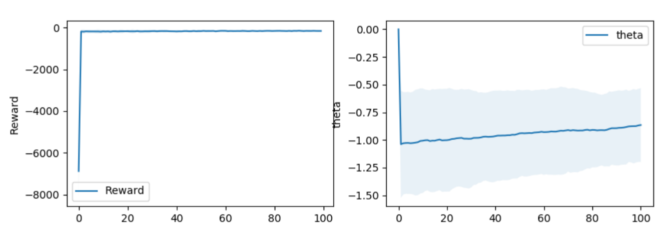

FIGURE 2: Constant method with a learning rate equals to 0.00001

2 Exploration in Policy Gradient

We find that adding an exploration bonus can improve the convergence speed. However, the β parameter must be correctly chosen. Indeed, for a choice of β too low, the algorithm does not perform enough exploration. Conversely, when β is too high, the algorithm does not perform enough exploitation and can not converge. We must find the correct compromise between exploration and exploitation.

We represent below the results obtained with $\beta = 10$ and $\beta = 50$:

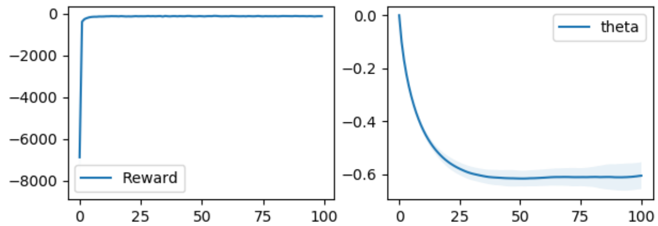

FIGURE 3: $\beta = 10$

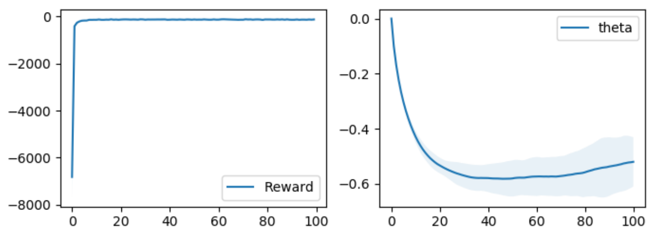

FIGURE 4: $\beta = 50$

We deduce that $\beta \simeq 10$ is a good choice.