Revisiting natural actor-critics with value function approximation

Matthieu Geist [a] Olivier Pietquin [a]

[a] *IMS Research Group, Supélec, Metz, France*

Reinforcement learning (RL) is generally considered as the machine learning answer to the optimal control problem. In this paradigm, an agent learns to control optimally a dynamic system through interactions. At each time step i, the dynamic system is in a given state s_i and receives from the agent a command (or action) a_i. According to its own dynamics, the system transits to a new state s_{i+1}, and a reward r_i is given to the agent. The objective is to learn a control policy maximizing the expected cumulative discounted reward.

Actor-critics approaches were among the first to be proposed for handling the RL problem [1]. In this setting, two structures are maintained, one for the actor (the control organ) and one for the critic (the value function which models the expected cumulative reward to be maximized). One advantage of such an approach is that it does not require knowledge about the system dynamics to learn an optimal policy. However, the introduction of the state-action value (or Q-) function [6] led to a focus of research community in pure critic methods, for which the control policy is derived from the Q-function and has no longer a specific representation. Actually, in contrast with value function, state-action value function allows deriving a greedy policy without knowing system dynamics, and function approximation (which is a way to handle large problems) is easier to combine with pure critic approaches. Pure critic algorithms therefore aim at learning this Q-function. However, actor-critics have numerous advantages over pure critics: a separate representation is maintained for the policy (in which we are ultimately interested), they somehow implicitly solve a problem known as dilemma between exploration and exploitation, they handle well large action spaces (which is not the case of pure critics, as some maxima over actions have always to be computed), and above all errors in the Q-function estimation can lead to bad derived policies.

A major march for actor-critics is the policy gradient with function approximation theorem [5, 3]. This result allows combining actor-critics with value function approximation, which was a major lack of the field. Another important improvement is the natural policy gradient [4] which replaces the gradient ascent over policy parameters by a natural gradient ascent improving consequently the efficiency of resulting algorithms. These results share the drawback that they lead to work with the advantage function which does not satisfy a Bellman equation. Consequently, derivation of practical algorithms is not straightforward, as it requires estimating the advantage function which is unnatural in RL. We reformulate (and re-prove) the theorems so as to work directly with the state-action value function [2]. This allows a very straightforward derivation of new actor-critic algorithms, some of them being proposed here. All results are given for the discounted cumulative reward case, however they can be easily extended to the average reward case.

References

[1] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. pages 535–549, 1988.

[2] Matthieu Geist and Olivier Pietquin. Revisiting natural actor-critics with value function approximation. In V. Torra, Y. Narukawa, and M. Daumas, editors, *Proceedings of 7th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2010)*, volume 6408 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 207–218, Perpinya (France), October 2010. Springer Verlag - Heidelberg Berlin.

[3] Vijay R. Konda and John N. Tsitsiklis. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems (NIPS 12)*, 2000.

[4] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Reinforcement Learning for Humanoid Robotics. In *third IEEE-RAS International Conference on Humanoid Robots (Humanoids 2003)*, 2003.

[5] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems (NIPS 12)*, pages 1057–1063, 2000.

[6] C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, England, 1989.