Natural Actor Critic

Jan Peters, Sethu Vijayakumar, Stefan Schaal
University of Southern California (USC)
Computational Motor Control and Learning Laboratory
3461 Watt Way, Los Angeles, CA 90089, USA

December 1, 2003

Reinforcement learning offers a promising framework to take planning for real-world systems towards true autonomy and versatility. However, applying reinforcement learning to high dimensional movement systems (such as real-world robots) in the presence of uncertainty and continuous state-action spaces remains an unsolved problem. In order to make progress towards solving this issue, we focus on a particular type of reinforcement learning methods, i.e., policy gradient methods. These methods are particularly interesting to the robotics community as they seem to scale better to continuous state-action problems and have been successfully applied on a variety of high-dimensional robots. However, the main disadvantages of these methods have been the high variance in the gradient estimate, the very slow convergence, and the dependence on baseline functions. In this poster, we show how these policy gradients can be improved in respect to each of these problems.

Our approach to policy gradients focuses on the natural policy gradient instead of the regular policy gradient. Natural policy gradients for reinforcement learning have first been suggested by Kakade [2] as 'average natural policy gradients', and subsequently been shown to be the true natural policy gradient by Bagnell & Schneider [1], and Peters et al. [3]. As shown by Kakade, natural policy gradients are particularly interesting due to the fact that they equal the parameters of the compatible function approximation. We present a general algorithm for estimating the natural gradient, the Natural Actor-Critic algorithm. This algorithm uses the fact that the compatible function approximation represents an advantage function which can be em-

bedded cleanly into the Bellman equation. It can be used in two different ways, i.e., in form of general temporal difference learning where reasonable basis functions for the value function are required in order to obtain an unbiased gradient, and in form of start-state reinforcement learning where the gradient is estimated using the property that the sum of all advantages along a roll-out has to equal the sum of rewards plus a single value function offset parameter. The later method is guaranteed to yield an unbiased estimate of the gradient and is well suited for learning and refining parameterized policies, even in the light of incomplete state information.

We show two examples of the application of the Natural Actor-Critic algorithm, one where it by far outperforms non-natural policy gradients in the classical cart-pole balancing system, and one for learning nonlinear dynamic motor primitives for humanoid robot control. From our experience, the start-state Natural Actor-Critic algorithm seems to be one of the most efficient model-free reinforcement learning techniques and offers a promising route for the development of reinforcement learning techniques for truly high dimensionally continuous state-action systems.

References

[1] James A. Bagnell and John Schneider. Covariant policy search. In *IJCAI*. International Joint Conference for Artificial Intelligence, 2003.

[2] Sham Kakade. A natural policy gradient. In *NIPS*. Advances in Neural Information Processing Systems, 2002.

[3] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Reinforcement learning for humanoid robotics. In *Humanoids*. IEEE Conference for Humanoid Robotics, 2003.