Guided Exploration in Policy Gradient Algorithms with Gaussian Process Function Approximation

Jakab Hunor

Gradient based direct policy optimization algorithms have strong convergence properties and can be used for model-free learning of control policies in complex high-dimensional Markov Decision Processes. The main drawback of the majority of Policy Gradient methods however is the high variance of the gradient estimate which is a result of approximating the Q-values in the gradient expression by Monte Carlo sampling of rewards accumulated through individual trajectories $Q^{\pi}(x,a) \sim \left(\sum_{j=o}^{H} \gamma^j r_j\right)$. In [1] we have presented a method for reducing the gradient variance by using a Gaussian Process based function approximator on state-action space to replace the high-variance(although unbiased) Monte Carlo estimation $Q(\cdot,\cdot) \sim GP(m_q, k_q)$. In this paper we present several ways to extend this method by exploiting the fully probabilistic nature of the Gaussian Process estimates to influence the directions of exploration. As we are mainly interested in robotics related learning problems, it is essential for us to develop methods that can be performed online. Also because of the nature of these learning problems, the number of trials that can be performed during the learning period, and the regions of the search-space that can be reached are limited both by physical and time constraints. Therefore random exploration in our methods performs poorly and needs to be replaced with guided exploration. In basic versions of Policy Gradient algorithms exploratory behavior is introduced by constructing the action-selection policy π from a deterministic controller output and an added exploratory noise: $\pi(a\|s) = f(s,\theta) + e$ where $e \sim N(0,\sigma^2)$. We replace the exploratory noise e by taking into account the GP predictive variance $k_q(s,a)$ at the state-action pair (s,a) which gives us useful information about how well the current region of the search-space has been explored. We design our search criteria based on the concept of exploitation and exploration. At each time-step for a given state s our controller returns a specific action $f(s,\theta) = a$. We sample a number of points from the neighborhood of a and get the predicted Q-values and variances for these points paired with state s. At this point we can choose to select an action from the sampled points which has the highest predicted Q-value (exploitation) or one that has the highest predictive variance (exploration). We define a parameterized measure which balances between these two. The exploratory noise's mean and variance will be set in such a way as to shift the policy π in the direction of the action selected in the above described way. We also investigate a simplistic approach where we only change the exploratory noise's variance from fixed σ^2 to the predictive variance of the selected state-action pair (s,a). We test the proposed methods on the inverted pendulum simulated control task in MATLAB where we compare its performance to our previous algorithm. We also perform tests on a pole-balancing control problem in a realistic simulation environment (ODE) to verify the behavior of the algorithm in uncertain environments.

Acknowledgements

This work was supported by programs co-financed by The SECTORAL OPERATIONAL PROGRAMME HUMAN RESOURCES DEVELOPMENT, Contract POSDRU 6/1.5/S/3 - "Doctoral studies: through science towards society".

References

[1] L. C. Hunor Jakab: Using Gaussian processes for variance reduction in policy gradient. *In ICAI2010: Proceedings of the 8th International Conference on Applied Informatics*, 2010.