Improving Gradient Estimation by Incorporating Sensor Data

Gregory Lawrence
Computer Science Division
U.C. Berkeley
gregl@cs.berkeley.edu

Policy search algorithms have been effective in learning good policies in the reinforcement learning setting. Successful applications include learning controllers for helicopter flight [3] and robot locomotion [1]. A key step in many policy search algorithms is estimating the gradient of the objective function with respect to a given policy's parameters. The gradient is typically estimated from a set of policy trials. Each of these trials can be rather expensive, and so we prefer to minimize the total number of trials required to achieve a desired level of performance. During each policy trial, an agent may receive a considerable amount of sensory data from its environment. While the agent's controller may use this information in deciding which actions to take, the sensory data is usually ignored in the gradient estimation task. In this poster we show that by viewing the task of estimating the gradient as a structured probabilistic inference problem, we can improve the learning performance. We argue that in many instances, reasoning about sensory data obtained during policy execution is beneficial. We demonstrate the effectiveness of this approach by showing a reduction in the variance of the gradient estimates for a simulated dart throwing problem and quadruped locomotion task. Our prior work shows one method of exploiting the sensor data in the case of perfect sensing of the control noise [2]. This work removes this assumption by allowing the agent access to a sensor model that only measures the positions of observable joints.

The performance measure f for our dart throwing task is defined as the negative squared distance to the target, and in the quadruped problem it is the distance travelled during a single policy trial. Our policies specify the desired trajectories of each controllable joint, and a PD-controller applies torques in an attempt to follow these paths. Actuator noise is added in the simulation, and the agent observes the actual joint angles for each policy execution. By using an appropriate encoding of the sensor data obtained during each policy trial, unexpected sensor values can be used to explain away the deviations in the observed performance. For example, in the dart throwing problem suppose that the agent executes a policy and notices that the dart missed the target. Normally, an agent would want to change it's desired trajectory so that the next throw moves closer to the target. However, suppose that during the previous throw the agent noticed that it let go of the dart too soon. This helps to explain the miss and allows the agent to better infer which direction it needs to move in policy space to improve its performance.

We consider parameterized policies $\pi \in \mathbb{R}^d$ that encode how an agent chooses its actions given its past observations, and the reinforcement learning goal is to find an optimal policy π^* that maximizes the performance measure f. The gradient is estimated from a collection of policy trials by learning two components. The first component is a linear model between the policy parameters π and a transformation of the sensor data $\phi(s)$ that is almost uncorrelated with the policy parameters. At each time step we predict what the next observation should be as a function of the current observation and controls. This prediction is formed by using estimates, which are obtained in a pre-processing routine, of the current mass matrix, gravity compensation, and inertial terms used in the equations of motion. We project the difference between the observed states and the predicted states down to a low-dimensional subspace using a set of basis functions. The second component is a linear model between the policy parameters π augmented with the sensor data $\phi(s)$ and the observed performance f. If the sensor data correlates with the noise in the performance measure, then this relationship will be easier to learn when compared to a model that ignores the sensor data.

References

[1] N. Kohl and P. Stone. Machine learning for fast quadrupedal locomotion. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 2004.

[2] Gregory Lawrence, Noah Cowan, and Stuart Russell. Efficient gradient estimation for motor control learning. In *Proceedings of the Nineteenth International Conference on Uncertainty in Artificial Intelligence*, 2003.

[3] Andrew Y. Ng, H. Jin Kim, Michael Jordan, and Shankar Sastry. Autonomous helicopter flight via reinforcement learning. In *Advances in Neural Information Processing Systems*, 2003.