Summary 19: Trust Region Policy Optimization

Susan Cherry

This paper describes a procedure for optimizing policies, which guarantees monotonic improvement. Policy optimization algorithms can be grouped into three broad categories: policy iteration methods, policy gradient methods, and derivative-free optimization methods. This paper proves that minimizing an alternative objective function guarantees policy improvement.

The paper begins by describing an infinite-horizon Markov decision process. It then goes into the theoretical results of the paper. Their main result is that Equation 6 can be extended to general stochastic properties and need not apply only to just mixture policies. To do this, α can be replaced with a distance measure between π and $\tilde{\pi}$. Next, they present Algorithm 1, which is an approximate policy iteration scheme based on the policy improvement bound which is given in Equation 9. Algorithm 1 is a minorization-maximization algorithm. The trust region policy optimization which is described next is an approximation of Algorithm 1. It uses a constraint on the KL divergence instead of a penalty to allow for large updates.

The next section describes how to derive a practical algorithm from the theoretical results under finite sample counts and arbitrary parameterization. Equation 12 shows their proposed optimization problem for a policy update. Their experiments how that this update works empirically similar to the performance of the maximum KL divergence constraint, which is presented in Equation 11. Next, they derive how to approximate the objective and constraint functions using Monte Carlo simulation. They present two sampling schemes: simple path and vine. Vine gives a local estimate of the objective function with a much lower variance than the single path method. However, vine requires many more calls to the simulator.

Next two practical policy optimization algorithms are presented. They perform the following three steps: use single path or vine to collect state action pairs and Monte Carlo estimates of their Q-values, construct the estimated objective and constraint in Equation 14, approximately solve the constrained optimization problem to update the policy's parameter vector θ. They then summarize the relationship between the theory presented in earlier sections and this approximation.

The paper ends with experiments to determine the performance characteristics of the path and vine sampling, determine how the changes from using a fixed KL divergence affect performance, and determine if TRPO can be used to solve large scale problems. First, they describe a simulated robotic locomotion experiment. They use neural networks to represent the policy and use a cart pole balance problem to establish a standard baseline. They compare single path TRPO, vine TRPO, and several other standard algorithms. They show that single path and vine TRPO yield the best solutions. Their results provide empirical evidence that fixed KL divergence is a more robust way to choose step size. Next they evaluate TRPO by training policies for playing Atari games. They test the same algorithms and find that the vine and single path algorithms achieve reasonable results but don't consistently outperform other algorithms. The authors claim that this is because their approach wasn't designed specifically for this task. They conclude the paper by discussing avenues for future work.