Interleaved Quadrature and Optimization for Planning in Continuous State-Action Space

Nishad Gothoskar

LIS

July 27 2018

Value Function: Discrete vs. Continuous

Discrete

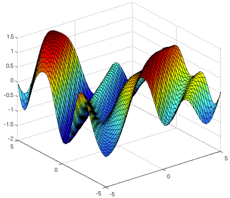
$$V^{\pi}(s) = \max_{a \in A} \Big[\sum_{s' \in S} p_{s'|s,a}(s'|s,a)\big(R(s'|s,a) + \gamma V^{\pi}(s')\big)\Big]$$

Continuous

$$V^{\pi}(s) = \max_{a \in A} \int_{s' \in S} p_{s'|s,a}(s'|s,a)\big(R(s'|s,a) + \gamma^{\Delta t} V^{\pi}(s')\big)\ \mathrm{d}s'$$

In the continuous setting, a value update consists of a maximation over an integral. More specifically, this is a maximization in a continuous action space and an integral of the product of a continuous state space value function and transition model.

Gaussian Process in RL



As opposed to discretizing the state space, we use a Gaussian Process to model the value function across the entire domain. Previous work in [1][2][3] has presented approaches to using GPs as value functions and learning a policy for a simple reinforcement learning problem (mountain car/swinging pendulum). There are many added complexities when working in larger state spaces and with complex dynamics models.

Background

Gaussian processes output a distribution of value at a state, which allows us to explicitly model our uncertainty about its value.

$$y^*|\mathbf{y} \sim \mathcal{N}(K_* K^{-1}\mathbf{y}, K_{**} - K_* K^{-1} K_*^T)$$

In addition, when using Gaussian kernels for the GP and assuming Gaussian or mixture of Gaussian transition models, the following integral will have a closed form solution (mean and variance). By taking this into account when designing an action selection policy, we can tune the "hyperparameters" of our final policy.

Problem Statement

In this work, we present a framework for planning in stochastic continuous state-action spaces. As opposed to discretizing the state space, we use a GP to model the value across the entirety of the state space. We use this model to actively select states to update in each iteration, therefore extending previous methods to be feasible in large high-dimensional state spaces. We also use the value GP, paired with a learned transition model, to estimate the value and uncertainty of taking certain actions, aiding in the action selection in a continuous domain. We show this method of active state selection and interleaved quadrature and optimization allows us to learn robust policies efficiently. We evaluate these policies with both a point robot in simulation and a block pushing scenario on the PR2.