The Policy Gradient and Actor-Critic Architectures for Deep Reinforcement Learning

Tom Wilson

April 2022

Abstract

In this project, I build an intuition about a popular class of reinforcement learning algorithms, called policy gradients. I implement a generalized framework for constructing policy-based learning agents, using function approximators to directly predict actions that will result in better playouts from a state observation. This is done in an interative process by sampling actions from a probability distribution.

1 The Policy Gradient

Unlike **value-based** methods in Reinforcement Learnging where we are using a function to approximate $V_\theta(s) \approx V^\pi(s) \in \mathbb{R}$ or $Q_\theta(s,a) \approx Q^\pi(s,a) \in \mathbb{R}$, we can also use a **policy-based** approach to directly parameterize the policy.

$$\pi_\theta(a|s) = \mathbb{P}[a|s;\theta]$$

Instead of extracting our policy from a value function, we use an approximation to model the policy mapping from states to actions, and update our approximation from the playout data. Using a stochastic policy, we remove the need to introduce an explicit exploration hyperparameter. It has been shown that policy-based RL has better convergence properties, as well as greater effectiveness in high-dimensional continuous action spaces. However, they typically have high variance, and are not guaranteed to converge to a globally optimal policy. The policy gradient is defined as follows:

$$\nabla_\theta J \approx \mathbb{E}[\sum_t^\infty A_t \nabla_\theta \pi_\theta(a_t|s_t)]$$