Reviewer of Policy Gradient

Guannan Hu

March 14, 2018

Reinforcement learning aims to learn a policy for an agent to maximize a sum of reward signals. The agent starts at an initial state $s_0 \sim P(s_0)$. Then, the agent repeatedly samples an action a_t from a policy $\pi_\theta(a_t|s_t)$ with parameters θ, receives a reward $r_t \sim P(r_t|s_t, a_t)$, and transitions to a subsequent state s_{t+1} according to the Markovian dynamics $P(s_{t+1}|a_t, s_t)$ of the environment. This generates a trajectory of states, actions and rewards $(s_0, a_0, r_0, s_1, a_1, ...)$. We abbreviate the trajectory after the initial state and action by τ.

The goal is maximize the discounted sum of rewards along sampled trajectories.

$$J(\theta) = \mathbb{E}_{s_0,a_0,\tau}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right] = \mathbb{E}_{s\sim\rho^\pi(s),a,\tau}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right],$$

where $\gamma \in [0, 1)$ is a discount parameter and $\rho^\pi(s) = \sum_{t=0}^{\infty}\gamma^t P^\pi(s_t = s)$ is the unnormalized discounted state visitation frequency.

Policy gradient methods differentiate the expected return objective with respect to the policy parameters and apply gradient-based optimization. The policy gradient can be written as an expectation amenable to Monte Carlo estimation

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_{s\sim\rho^\pi(s),a,\tau}\left[Q^\pi(s,a)\nabla\log\pi(a|s)\right] \\ &= \mathbb{E}_{s\sim\rho^\pi(s),a,\tau}\left[A^\pi(s,a)\nabla\log\pi(a|s)\right]\end{aligned} \tag{1}$$

where $Q_\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t|s_0 = s, a_0 = a\right]$ is the state-action value function. $V_\pi(s) = \mathbb{E}_a[Q_\pi(s, a)]$ is the value function, and $A_\pi = Q^\pi(s, a) - V_\pi(s)$ is the advantage function. the equality in the last line follows from the fact that $\mathbb{E}_a[\nabla\log\pi(a|s)] = 0$