Policy gradient methods

March 28, 2023

Chapter 13 of Sutton-Barto

1 Policy gradient methods

Methods that parameterize the policy (may or may not parameterize the value function)

Let $\pi(a|s,\theta) = Pr(A_t = a|S_t = s, \theta)$

Example: Parameterized Boltzmann policy

$$\pi(a|s,\theta) = \frac{e^{\theta^T \phi(s,a)}}{\sum_{b \in A(s)} \theta^T \phi(s,b)}$$

$\phi(s,a)$: features associated with (s,a) tuples

$$\pi(a|s,\theta) = \frac{e^{h(s,a\theta)}}{\sum_{b \in A(s)} e^{h(s,b,\theta)}}$$

h can be via LFA or Neural network based parameterization

Let $\theta \in \mathbb{R}^{d'}$ and $J : \mathbb{R}^{d'} \to \mathbb{R}$ be the performance function.

Then,

$$\theta_{t+1} = \theta_t + \alpha \nabla \hat{J}(\theta_t)$$

Here, $\nabla \hat{J}(\theta_t)$ is the estimate of $\nabla J(\theta_t)$

Example : Goal: Find λ such that average queue length as function of λ is minimized

$\theta = \lambda$, $J(\theta) = \mathbb{E}\left[Q(\theta)\right]$

Assumption (on policy) : $\pi(a|s,\theta) > 0 \forall a, s, \theta$
Example of $J(\theta)$: value function v_{π_θ} under policy π_θ

Policy gradient theorem : [Episode setting with $\gamma = 1$]

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla_\theta \pi(a|s,\theta)$$

Here, $J(\theta) = v_{\pi_\theta}(s_0)$ where $s_0 \in S$ is same given state.
Consider

$$\begin{aligned}
\nabla J(\theta) &= \nabla_\theta v_{\pi_\theta}(s_0) \\
&= \nabla_\theta \left(\sum_{a \in A(s_0)} \pi(a|s_0,\theta) q_\pi(s_0,a) \right) \\
&= \sum_a \left(\nabla \pi(a|s_0,\theta) q_\pi(s_0,a) + \pi(a|s_0,\theta) \nabla \left(\sum_{s',r} p(s',r|s_0,a)(r + v_\pi(s')) \right) \right)
\end{aligned}$$

$$\begin{aligned}
\nabla \left(\sum_{s',r} p(s',r|s_0,a)(r + v_\pi(s')) \right) &= \sum_{s',r} p(s',r|s_0,a) \nabla v_\pi(s') \\
&= \sum_{s'} \overline{p}(s'|s_0,a) \nabla v_\pi(s') \\
\text{where, } \overline{p}(s',|s_0,a) &= \sum_r p(s',r|s_0,a)
\end{aligned}$$

$$\begin{aligned}
\nabla J(\theta) &= \sum_a \left[\nabla \pi(a|s_0,\theta) q_\pi(s_0,a) + \pi(a|s_0,\theta) \left(\sum_{s'} \overline{p}(s'|s_0,a) \nabla v_\pi(s') \right) \right] \\
&= \sum_a \left[\nabla \pi(a|s_0,\theta) q_\pi(s,a) + \pi(a|s_0,\theta) \sum_{s'} \overline{p}(s'|s_0,a) \right. \\
&\times \left[\sum_{a'} \left(\nabla \pi(a'|s',\theta) q_\pi(s',a') + \pi(a'|s',\theta) \times \sum_{s"} \overline{p}(s"|s',a') \times \nabla v_\pi(s") \right) \right]
\end{aligned}$$

Probability of given from s_0 to x in k steps under π

$$\sum_{x \in S} \sum_{k=0}^{\infty} Pr(s_0 \to x, k, \pi) \sum_a \nabla \pi(a|x,\theta) q_\pi(x,a)$$

Then,

$$\begin{aligned}
\nabla J(\theta) &= \nabla v_{\pi_\theta}(s_0) \\
&= \sum_x \eta(x) \sum_a \nabla\pi(a|x,\theta) q_\pi(x,a) \\
&= \sum_{s'} \eta(s') \sum_x \left(\frac{\eta(x)}{\sum_{s'} \eta(s')} \right) \sum_a \nabla\pi(a|x,\theta) q_\pi(x,a) \\
&\propto \sum_x \mu(x) \sum_a \nabla\pi(a|x,\theta) \mu(x) \times q_\pi(x,a)
\end{aligned}$$

2 Reinforce: Monte-Carlo policy gradient

Note:

$$\begin{aligned}
\nabla J(\theta) &\propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla_\theta \pi(a|s,\theta) \\
&= \mathop{\mathbb{E}}_\pi \left[\sum_a q_\pi(s_t,a) \nabla\pi(a|s_t,\theta) \right] \\
&= \mathop{\mathbb{E}}_\pi \left[\sum_a \pi(a|s_t,\theta) q_\pi(s_t,a) \frac{\nabla\pi(a|s_t,\theta)}{\pi(a|s_t,\theta)} \right] \\
&= \mathop{\mathbb{E}}_\pi \left[q_\pi(s_t,A_t) \frac{\nabla\pi(A_t|s_t,\theta)}{\pi(A_t|s_t,\theta)} \right] \\
&= \mathop{\mathbb{E}}_\pi \left[G_t \frac{\nabla\pi(A_t|s_t,\theta)}{\pi(A_t|s_t,\theta)} \right]
\end{aligned}$$

Where G_t is the return from time t.

Reinforce

$$\begin{aligned}
\theta_{t+1} &= \theta_t + \alpha G_t \frac{\nabla\pi(A_t|s_t,\theta_t)}{\pi(A_t|s_t,\theta_t)} \\
&= \theta_t + \alpha \left(\mathop{\mathbb{E}}_\pi \left[G_t \frac{\nabla\pi(A_t|s_t,\theta_t)}{\pi(A_t|s_t,\theta_t)} \middle| f_t \right] + M_{t+1} \right)
\end{aligned}$$

Where, $M_{t+1} = \frac{G_t \nabla\pi(A_t|s_t,\theta_t)}{\pi(A_t|s_t,\theta_t)} - \mathbb{E}_\pi \left[G_t \frac{\nabla\pi(A_t|s_t,\theta_t)}{\pi(A_t|s_t,\theta_t)} \middle| f_t \right]$

Here, $f_t = \sigma(\theta_s, S_s, A_s, s \le t)$, $t \ge 0$, $\{\theta_s \le c, s_t \le b, A_s \le a, s \le t\} \in f_t$

(M_t, f_t), $t \ge 0$ is a martingale difference sequence, $\mathbb{E}[M_{t+1}|f_t] = 0$

If $\sum_t \alpha_t M_{t+1} < \infty$ (happens if $\sum_t \alpha_t^2 < \infty$, $E[M_{t+1}^2|f_t] \leq k(1+\|\theta_t\|^2)$)

$\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$

ODE: $\theta(t) = \mathbb{E}_\pi \left[G_t \frac{\nabla\pi(A_t|s_t,\theta_t)}{\pi(A_t|s_t,\theta_t)} \right] = \nabla J(\theta_t)$

Stationary points $\{\theta | \nabla J(\theta) = 0\}$
Under some conditions, can show that $\theta_t \to$ local maxima of J.
Pemantle (1990)

3 Reinforce with Baseline

Let $b : S \to \mathbb{R}$ be a certain function, we call this the baseline function.

The policy gradient can be generalized as follows

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a \left(q_\pi(s,a) - b(s)\right) \nabla\pi(a|s,\theta)$$

$$\sum_a b(s)\nabla\pi(a|s,\theta) = b(s)\sum_a \nabla\pi(a|s,\theta)$$

where, $\sum_a \nabla\pi(a|s,\theta) = 0$

$$\theta_{t+1} = \theta_t + \alpha(G_t - b(s_t))\frac{\nabla\pi(A_t|s_t,\theta_t)}{\pi(A_t|s_t,\theta_t)}$$

A good choice of $b(s_t)$ is $v_\pi(s_t)$.

$$\theta_{t+1} = \theta_t + \alpha(G_t - \hat{v}(s_t,w_t))\frac{\nabla\pi(A_t|s_t,\theta_t)}{\pi(A_t|s_t,\theta_t)}$$

Incremental update algorithm

$$\text{(PG update) } \theta_{t+1} = \theta_t + \alpha\left(R_{t+1} + \gamma\hat{v}(s_{t+1},w_t) - \hat{v}(s_t,w_t)\right)\nabla log\pi(A_t|s_t,\theta_t)$$
$$\text{(TD update) } w_{t+1} = w_t + \beta\left(R_{t+1} + \gamma G(s_{t+1},w_t) - \hat{v}(s_t,w_t)\right)\nabla\hat{v}(s_t,w_t)$$

$\sum_t \alpha_t = \sum_t \beta_t = \infty$, $\sum_t \alpha_t^2, \sum_t \beta_t^2 < \infty$, $\frac{\alpha_t}{\beta_t} \to 0$ as $t \to \infty$
This is called **Actor-Critic algorithm**.