Policy Gradient

Peng Lingwei

July 19, 2019

References

[1] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. 2014.

[2] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Contents

1 Stochastic Policy Gradient [2]

1.1 Sutton's proof (I have changed a little bit.)

Definition 1. ***(Average policy value).***

$$J(\pi) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\{r_1 + r_2 + \cdots + r_n|\pi\} = (d^\pi)^T \cdot \langle \pi, R_s^a\rangle \tag{1}$$

where $d^\pi(s) = \lim_{n\to\infty} P\{s_t = s|s_0, \pi\}$. Corresponding Q function is defined as

$$Q^\pi(s,a) = \sum_{t=1}^{\infty} \mathbb{E}\{r_t - J(\pi)|s_0 = s, a_0 = a, \pi\}, \forall s \in S, a \in A \tag{2}$$

Definition 2. ***(Discounted policy value).***

$$J(\pi) = \sum_{s_0} p_0(s_0)\mathbb{E}\left\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t|s_0, \pi\right\} \quad and \quad Q^\pi(s,a) = \mathbb{E}\left\{\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}|s_t = s, a_t = a, \pi\right\} \tag{3}$$

then, $d^\pi = \sum_{s_0} p_0(s_0) \sum_{t=0}^{\infty} \gamma^t \Pr\{s_t = s|s_0, \pi\}$, where $p_0(s)$ is an initial state distribution.

Theorem 1. ***(Stochastic Policy Gradient).***

$$\frac{\partial J}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) \tag{4}$$

Proof. Step1: For average-reward formulation:

$$\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} =& \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a), \quad \forall s \in S \\
=& \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial Q^\pi(s,a)}{\partial \theta} \\
=& \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} \left[R_s^a - J(\pi) + \sum_{s'} P_{ss'}^a V^\pi(s') \right] \right] \\
=& \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \left[-\frac{\partial J}{\partial \theta} + \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right] \right] \\
\frac{\partial J}{\partial \theta} =& \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \left[\sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right] \right] - \frac{\partial V^\pi(s)}{\partial \theta} \\
\sum_s d^\pi(s) \frac{\partial \rho}{\partial \theta} =& \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \sum_s d^\pi(s) \sum_a \pi(s,a) \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \\
&- \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \theta} \\
\frac{\partial J}{\partial \theta} =& \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \sum_{s'} d^\pi(s') \frac{\partial V^\pi(s')}{\partial \theta} - \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \theta} \\
=& \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)
\end{aligned}$$

Step2: For discounted policy value.

$$
\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} =& \frac{\partial}{\partial \theta} \sum_a \pi(s,a) Q^\pi(s,a), \quad \forall s \in S \\
=& \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial Q^\pi(s,a)}{\partial \theta} \right] \\
=& \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \frac{\partial}{\partial \theta} \left[R^a_{ss'} + \sum_{s'} \gamma P^a_{ss'} V^\pi(s') \right] \right] \\
=& \sum_a \left[\frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) + \pi(s,a) \left[\sum_{s'} \gamma P^a_{ss'} \frac{\partial}{\partial \theta} V^\pi(s') \right] \right] \\
& \vdots \\
=& \sum_{s'} \sum_{k=0}^{\infty} \gamma^k \Pr(s \to s', k, \pi) \sum_a \frac{\partial \pi(s',a)}{\partial \theta} Q^\pi(s',a) \\
\frac{\partial J}{\partial \theta} =& \sum_{s_0} p_0(s_0) \frac{\partial}{\partial \theta} \mathbb{E} \left\{ \sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_0, \pi \right\} = \sum_{s_0} p_0(s_0) \frac{\partial V^\pi(s_0)}{\partial \theta} \\
=& \sum_s \sum_{s_0} p_0(s_0) \sum_{k=0}^{\infty} \gamma^k \Pr\{s_0 \to s, k, \pi\} \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a) \\
=& \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \theta} Q^\pi(s,a)
\end{aligned}
$$

□

1.2 Revised

1. Initial state distribution with density $p_0(s)$.
2. Stationary transition dynamics distribution with conditional density: $p(s_{t+1}|s_t, a)$.
3. The discounted reward: $r_t^\gamma = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k)$.
4. The state expected total discounted reward: $V^\pi(s) = \mathbb{E}[r_0^\gamma | S_0 = s; \pi]$
5. The state-action expected total discounted reward: $Q^\pi(s,a) = \mathbb{E}[r_0^\gamma | S_0 = s, A_0 = a; \pi]$
6. Discounted state distribution $\rho^\pi(s) = \int_S \sum_{t=0}^{\infty} \gamma^t p_0(s_0) p(s_0 \to s, t, \pi) ds_0$
7. The performance of policy:

$$
\begin{aligned}
J(\pi) =& \int_S \rho^\pi(s) \int_A \pi_\theta(s,a) r(s,a) da ds \\
=& \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta}[r(s,a)]
\end{aligned}
$$

8. Stochastic Policy Gradient Theorem:

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) =& \int_S \rho^\pi(s) \int_A \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) da ds \\
=& \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)]
\end{aligned}
$$

1.3 Stochastic Actor-Critic Algorithms

We let $Q^w(s,a)$ to approximate $Q^\pi(s,a)$.

Theorem 2. *If* $Q^w(s,a) = (\nabla_\theta \log \pi_\theta(a|s))^T w$ *and* $w = \arg\min_w \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}\left[(Q^w(s,a) - Q^\pi(s,a))^2\right]$, *then*

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q^w(s,a)]$$

Proof.

$$\begin{aligned}
0 =& \nabla_w \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}\left[(Q^w(s,a) - Q^\pi(s,a))^2\right] \\
=& \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}\left[(Q^w(s,a) - Q^\pi(s,a))\nabla_w Q^w(w,a)\right] \\
=& \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s)(Q^w(s,a) - Q^\pi(s,a))\right] \\
\nabla_\theta J(\pi_\theta) =& \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q^w(s,a)]
\end{aligned}$$

□

Definition 3. ***(Stochastic Actor-Critic algorithm)***.

1. ***critic***: w_{critic}, *such that* $\mathbb{E}_{s\sim\rho^{\pi_{\theta_k}}, a\sim\pi_{\theta_k}}\left[\nabla_\theta \log \pi_{\theta_k}(a|s)(Q^{w_{critic}}(s,a) - Q^{\pi_{\theta_k}}(s,a))\right] = 0$.
2. ***actor***: $\theta_{k+1} = \theta_k + \alpha_k \mathbb{E}_{s\sim\rho^{\pi_{\theta_k}}, a\sim\pi_{\theta_k}}[\nabla_\theta \log \pi_{\theta_k}(a|s) Q^{w_{critic}}(s,a)]$

2 Deterministic Policy Gradient [1]

2.1 Basic Proof

Conditions

1. $p(s'|s,a), \nabla_a p(s'|s,a), \mu_\theta(s), \nabla_\theta \mu_\theta(s), r(s,a), \nabla_a r(s,a), p_0(s)$ are continuous in all parameters and variables s, a, s' and x.
2. $\exists b, L, \sup_s p_0(s) < b, \sup_{a,s,s'} p(s'|s,a) < b, \sup_{a,s} r(s,a) < b, \sup_{a,s,s'} \|\nabla_a p(s'|s,a)\| < L$, and $\sup_{a,s} \|\nabla_a r(s,a)\| < L$.

Definition 4. *(Greedy Policy, or Deterministic Policy)*

$$\mu_\theta(s) = \arg\max_a Q^{\mu_\theta}(s,a).$$

Definition 5. ***(Deterministic discounted policy value).***

$$J(\mu_\theta) = \int_S \rho^{\mu_\theta}(s) r(s, \mu_\theta(s)) ds = \mathbb{E}_{s\sim\rho^{\mu_\theta}}[r(s, \mu_\theta(s))]$$

$$Q^{\mu_\theta}(s,a) = \mathbb{E}\left\{\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} | s_t = s, a_t = a, \pi\right\}$$

$$\rho^{\mu_\theta}(s_0) = \int_S \sum_{t=0}^{\infty} \gamma^t p_0(s_0) p(s_0 \to s, t, \mu_\theta) ds_0$$

Theorem 3. ***(Deterministic Policy Gradient Theorem)***. *If preceeding conditions are satisfied, and $\nabla_\theta \mu_\theta(s)$ and $\nabla_a Q^{\mu_\theta}(s,a)$ exist, and that the deterministic policy gradient exists. Then,*

$$\begin{aligned}\nabla_\theta J(\mu_\theta) =& \int_S \rho_\theta^{\mu_\theta}(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta} ds \\ =& \mathbb{E}_{s\sim\rho^{\mu_\theta}} [\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta}]\end{aligned} \tag{5}$$

Proof.

$$\begin{aligned}
\nabla_\theta V^{\mu_\theta}(s) =& \nabla_\theta Q^{\mu_\theta}(s, \mu_\theta(s)) \\
=& \nabla_\theta \left(r(s,\mu_\theta(s)) + \int_S \gamma p(s'|s,\mu_\theta(s)) V^{\mu_\theta}(s') ds' \right) \\
=& \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} + \nabla_\theta \int_S \gamma p(s'|s,\mu_\theta(s)) V^{\mu_\theta}(s') ds' \\
=& \nabla_\theta \mu_\theta(s) \nabla_a r(s,a)|_{a=\mu_\theta(s)} \\
&+ \int_S \gamma \left(p(s'|s,\mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') + \nabla_\theta \mu_\theta(s) \nabla_a p(s'|s,a)|_{a=\mu_\theta(s)} V^{\mu_\theta}(s') \right) ds' \\
=& \nabla_\theta \mu_\theta(s) \nabla_a \left(r(s,a) + \int_S \gamma p(s'|s,a) V^{\mu_\theta}(s') ds' \right) |_{a=\mu_\theta(s)} \\
&+ \int_S \gamma p(s'|s,\mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s') ds' \\
=& \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \int_S \gamma p(s \to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s') ds' \\
&\vdots \\
=& \int_S \sum_{t=0}^{\infty} \gamma^t p(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' \\
\nabla_\theta J(\mu_\theta) =& \nabla_\theta \int_S p_0(s) V^{\mu_\theta}(s) ds \\
=& \int_S p_0(s) \nabla_\theta V^{\mu_\theta}(s) ds \\
=& \int_S \int_S \sum_{t=0}^{\infty} \gamma^t p_0(s) p(s \to s', t, \mu_\theta) \nabla_\theta \mu_\theta(s') \nabla_a Q^{\mu_\theta}(s',a)|_{a=\mu_\theta(s')} ds' ds \\
=& \int_S \int_S \sum_{t=0}^{\infty} \gamma^t p_0(s_0) p(s_0 \to s, t, \mu_\theta) ds_0 \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} ds \\
=& \int_S \rho^{\mu_\theta}(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} ds
\end{aligned}$$

□

2.2 The Relationship Between Stochastic and Deterministic Policy Gradient (Unfinished)

Theorem 4. *Deterministic policy gradient is a special case of the stochastic policy gradient.*