Deterministic Policy Gradient Algorithms

Changyou Chen

Department of Computer Science and Engineering
Universitpy at Buffalo, SUNY
changyou@buffalo.edu

May 2, 2019

Stochastic Policy Gradient

$$
\begin{aligned}
J(\pi_\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(s, a) r(s, a) \mathrm{d}a\mathrm{d}s \\
&= \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}\left[r(s, a)\right]
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^\pi(s, a) \mathrm{d}a\mathrm{d}s \\
&= \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)\right]
\end{aligned}
$$

Deterministic Policy Gradient

- Consider a deterministic policy: $\mu_\theta : \mathcal{S} \to \mathcal{A}$. Let $\rho^\mu(s)$ be the discounted state distribution:

$$
\begin{aligned}
J(\mu_\theta) &= \int_{\mathcal{S}} \rho^\mu(s) r(s, \mu_\theta(s)) \mathrm{d}s \\
&= \mathbb{E}_{s\sim\rho^\mu}\left[r(s, \mu_\theta(s))\right]
\end{aligned}
$$

Theorem 1 (Deterministic Policy Gradient Theorem). *Suppose that the MDP satisfies conditions A.1 (see Appendix; these imply that $\nabla_\theta \mu_\theta(s)$ and $\nabla_a Q^\mu(s,a)$ exist and that the deterministic policy gradient exists. Then,*

$$
\begin{aligned}
\nabla_\theta J(\mu_\theta) &= \int_{\mathcal{S}} \rho^\mu(s) \nabla_\theta \mu_\theta(s) \left.\nabla_a Q^\mu(s,a)\right|_{a=\mu_\theta(s)} \mathrm{d}s \\
&= \mathbb{E}_{s\sim\rho^\mu}\left[\nabla_\theta \mu_\theta(s) \left.\nabla_a Q^\mu(s,a)\right|_{a=\mu_\theta(s)}\right] \quad (9)
\end{aligned}
$$

On Policy Deterministic Actor-Critic

- Use another DNN (called critic) to approximate the Q-function, as in DQN.

$$
\begin{aligned}
\delta_t &= r_t + \gamma Q^w(s_{t+1}, a_{t+1}) - Q^w(s_t, a_t) \\
w_{t+1} &= w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \\
\theta_{t+1} &= \theta_t + \alpha_\theta \nabla_\theta \mu_\theta(s_t) \left. \nabla_a Q^w(s_t, a_t) \right|_{a=\mu_\theta(s)}
\end{aligned}
$$

Off Policy Deterministic Actor-Critic

- learn a deterministic target policy $\mu(s)$ from trajectories generated by an arbitrary stochastic behavior policy $\pi(s, a)$.

$$
\begin{aligned}
J_\beta(\mu_\theta) &= \int_{\mathcal{S}} \rho^\beta(s) V^\mu(s) \mathrm{d}s \\
&= \int_{\mathcal{S}} \rho^\beta(s) Q^\mu(s, \mu_\theta(s)) \mathrm{d}s
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\theta J_\beta(\mu_\theta) &\approx \int_{\mathcal{S}} \rho^\beta(s) \nabla_\theta \mu_\theta(a|s) Q^\mu(s, a) \mathrm{d}s \\
&= \mathbb{E}_{s \sim \rho^\beta} \left[\nabla_\theta \mu_\theta(s) \; \nabla_a Q^\mu(s, a)|_{a=\mu_\theta(s)} \right]
\end{aligned}
$$

$$
\begin{aligned}
\delta_t &= r_t + \gamma Q^w(s_{t+1}, \mu_\theta(s_{t+1})) - Q^w(s_t, a_t) \\
w_{t+1} &= w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \\
\theta_{t+1} &= \theta_t + \alpha_\theta \nabla_\theta \mu_\theta(s_t) \; \nabla_a Q^w(s_t, a_t)|_{a=\mu_\theta(s)}
\end{aligned}
$$

Experiments

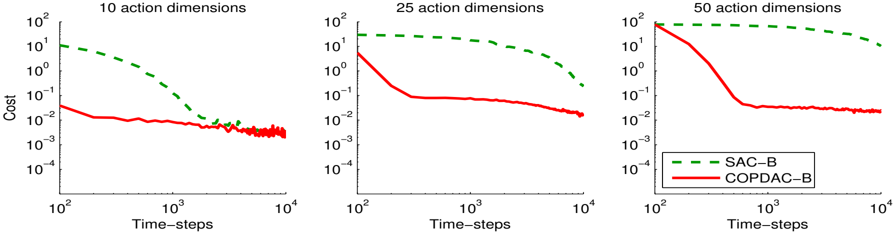
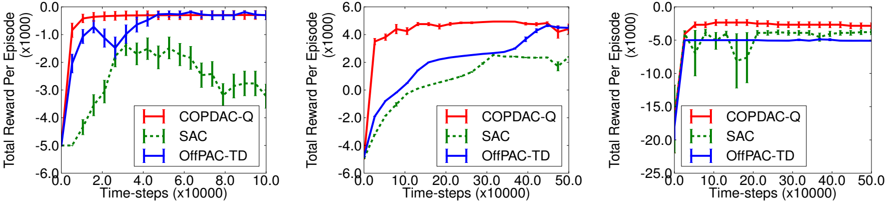


Figure 1. Comparison of stochastic actor-critic (SAC-B) and deterministic actor-critic (COPDAC-B) on the continuous bandit task.



(a) Mountain Car (b) Pendulum (c) 2D Puddle World

Figure 2. Comparison of stochastic on-policy actor-critic (SAC), stochastic off-policy actor-critic (OffPAC), and deterministic off-policy actor-critic (COPDAC) on continuous-action reinforcement learning. Each point is the average test performance of the mean policy.