DPG, DDPG, and TD3[1]

[1] Section 7.3: Yang&Ying

Deterministic policy gradient (DPG)

Deterministic Policy

- A parameterized critic w to estimate the Q-function: $Q_w(s, a)$.
- A parameterized actor θ that outputs an action $\mu_\theta(s)$ given state s.

Deterministic Policy Gradient Theorem

$$\nabla_\theta Q_w(x_0, \mu_\theta(x_0)) = \frac{1}{1-\alpha} E_{x\sim\rho_\theta}\left[\nabla_u Q_w(x, u)|_{u=\mu_\theta(x)} \nabla_\theta \mu_\theta(x)\right],$$

where

$$\rho_\theta(x) = (1-\alpha)\sum_{k=0}^{\infty} \alpha^k \Pr(x_k = x).$$

Proof

$$
\begin{aligned}
&\nabla_\theta Q_w(x_0, \mu_\theta(x_0)) \\
=&\nabla_\theta \left(r(x_0, \mu_\theta(x_0)) + \alpha \int_{\mathcal{S}} Q_w(x_1, \mu_\theta(x_1)) f(x_1|x_0; \mu_\theta(x_0)) dx_1 \right) \\
=&\nabla_u r(x_0, u)|_{u=\mu_\theta(x_0)} \nabla_\theta \mu_\theta(x_0) \\
&+ \alpha \left(\int_{\mathcal{S}} Q_w(x_1, \mu_\theta(x_1)) \nabla_u f(x_1|x_0; u)|_{u=\mu_\theta(x_0)} \nabla_\theta \mu_\theta(x_0) dx_1 \right) + \\
&+ \alpha \int_{\mathcal{S}} \nabla_\theta Q_w(x_1, \mu_\theta(x_1)) f(x_1|x_0; \mu_\theta(x_0)) dx_1 \\
=&\nabla_\theta \mu_\theta(x_0) \, \nabla_u \left(r(x_0, u) + \alpha \left(\int_{\mathcal{S}} Q_w(x_1, \mu_\theta(x_1)) f(x_1|x_0; u) dx_1 \right) \right) \Big|_{u=\mu_\theta(x_0)} \\
&+ \alpha \int_{\mathcal{S}} \nabla_\theta Q_w(x_1, \mu_\theta(x_1)) f(x_1|x_0; \mu_\theta(x_0)) dx_1
\end{aligned}
$$

Proof

$$\begin{aligned}&\nabla_\theta Q_w(x_0, \mu_\theta(x_0))\\ =&\nabla_\theta \mu_\theta(x_0)\, \nabla_u Q_w(x_0, u)|_{u=\mu_\theta(x_0)} + \alpha E_{x_1}\left[\nabla_\theta Q_w(x_1, \mu_\theta(x_1))\right].\end{aligned}$$

Repeatedly using the equation above, we obtain

$$\nabla_\theta Q_w(x_0, \mu_\theta(x_0)) = \frac{1}{1-\alpha} E_{x\sim\rho_\theta}\left[\nabla_u Q_w(x, u)|_{u=\mu_\theta(x)} \nabla_\theta \mu_\theta(x)\right].$$

Off-Policy Deterministic policy gradient (DPG)

Off Policy Deterministic Policy Gradient (Approximation)

$$\nabla_\theta Q_w(x_0, \mu_\theta(x_0)) \approx \frac{1}{1-\alpha} E_{x\sim\rho_{\tilde{\theta}}} \left[\nabla_u Q_w(x,u)|_{u=\mu_\theta(x)} \nabla_\theta \mu_\theta(x) \right],$$

where $\mu_{\tilde{\theta}}$ is the behavior policy.

Deep deterministic policy gradient (DDPG)

Critic

Q-network $Q_w(s, a)$: trained with mini-batch and temporal-difference.

Deep deterministic policy gradient (DDPG)

Critic

Q-network $Q_w(s, a)$: trained with mini-batch and temporal-difference.

Actor

Deterministic policy-network $\mu_\theta(x)$: trained with loss function

$$L(\theta) = - \sum_{s_b \in \text{minibatch}} Q_w(s_b, \mu_\theta(s_b)).$$

Deep deterministic policy gradient (DDPG)

Critic

Q-network $Q_w(s, a)$: trained with mini-batch and temporal-difference.

Actor

Deterministic policy-network $\mu_\theta(x)$: trained with loss function

$$L(\theta) = - \sum_{s_b \in \text{minibatch}} Q_w(s_b, \mu_\theta(s_b)).$$

Twin Delayed DDPG (TD3) (Fujimoto, van Hoof, Meger, 2018): Clipped double-Q + deterministic PG.

Reward Shaping

Motivation

Consider the following problem:

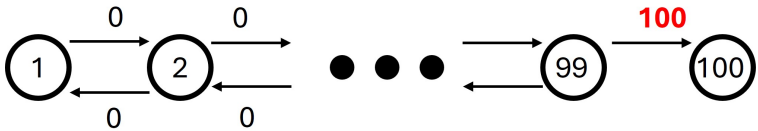

Figure: An MDP Example

Can we shape the reward function to encourage the agent to move towards the right?

Motivation

A naive approach: add an extra reward for going towards the right?

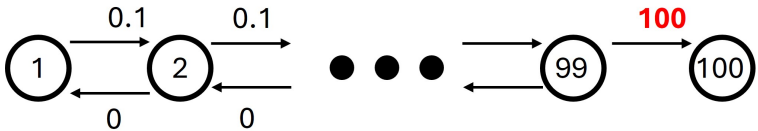

Figure: Extra rewards for going right

Motivation

A naive approach: add an extra reward for going towards the right?

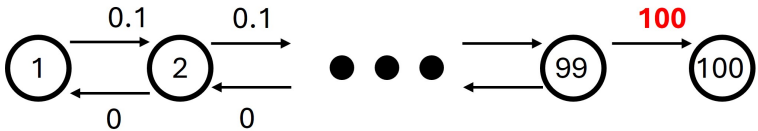

Figure: Extra rewards for going right

Issue: Now it is better for the agent to cycle repeatedly in the first circle than to go to the right.

Reward Shaping

Policy Invariance Reward Shaping

The optimal policy remains the same after reward shaping.

Reward Shaping

Policy Invariance Reward Shaping

The optimal policy remains the same after reward shaping.

Definition: Reward Shaping

Transfer a MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \alpha; R)$ to an MDP with a different reward function $\tilde{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \alpha; \tilde{R})$, where

$$\tilde{R} = R + F,$$

where $F : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbf{R}$ is a bounded real-valued function called the shaping reward function.

Potential-based Shaping Function

Potential-based Shaping Function

F is a potential-based shaping function if there exists a real-valued function $\phi : \mathcal{S} \to \mathbf{R}$ such that for all x, $a \in \mathcal{A}$, and $x' \in \mathcal{S}$,

$$F(x, a, x') = \alpha\phi(x') - \phi(x).$$

Potential-based Shaping Function

Potential-based Shaping Function

F is a potential-based shaping function if there exists a real-valued function $\phi : \mathcal{S} \rightarrow \mathbf{R}$ such that for all x, $a \in \mathcal{A}$, and $x' \in \mathcal{S}$,

$$F(x, a, x') = \alpha\phi(x') - \phi(x).$$

Theorem

Result 1: If F is a potential-based shaping function, then every optimal policy for M will also be an optimal policy for $\tilde{M}$ (and vice versa).
Result 2: If F is not a potential-based shaping function (e.g. no such ϕ exists satisfying the equation), then there exist a (proper) transition kernel and a reward function such that no optimal policy in $\tilde{M}$ is optimal in M.

Proof

Let Q^* ($\tilde{Q}^*$) be the optimal Q-function for the MDP M ($\tilde{M}$). From the definition, we have

$$Q^*(x,a) = E[R(x,a,x') + \alpha \max_u Q^*(x',u)],$$

so

$$Q^*(x,a) - \phi(x) = E[R(x,a,x') + \alpha\phi(x') - \phi(x) + \alpha \max_u (Q^*(x',u) - \phi(x'))].$$

Define $\hat{Q}(x,a) = Q^*(x,a) - \phi(x)$. Since $F(x,a,x') = \alpha\phi(x') - \phi(x)$, from the equation above, we obtain

$$\begin{aligned}\hat{Q}(x,a) &= E[R(x,a,x') + F(x,a,x') + \alpha \max_u \hat{Q}(x',u)] \\ &= E\left[\tilde{R}(x,a,x') + \alpha \max_u \hat{Q}(x',u)\right]\end{aligned}$$

Theory

Proof of sufficiency (continue)

Therefore $\hat{Q} = \tilde{Q}^*$ is the optimal Q-function for $\tilde{M}$. The optimal policy for $\tilde{M}$ is

$$\begin{aligned}\tilde{\pi}^* &\in \arg\max_u \hat{Q}(x,u) \\ &= \arg\max_u Q^*(x,u) - \phi(x) \\ &= \arg\max_u Q^*(x,u).\end{aligned}$$

Hence, $\tilde{\pi}^*$ is the same as the optimal policy π^*.

References

- Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. and Riedmiller, M., 2014, January. Deterministic policy gradient algorithms. In International Conference on machine learning (pp. 387-395). PMLR.
- Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra, *Continuous control with deep reinforcement learning*, arXiv preprint arXiv:1509.02971 (2015).
- Fujimoto, S., Hoof, H. and Meger, D., 2018, July. Addressing function approximation error in actor-critic methods. In International Conference on machine learning (pp. 1587-1596). PMLR.
- Ng, A.Y., Harada, D. and Russell, S., Policy invariance under reward transformations: Theory and application to reward shaping. In ICML (Vol. 99, pp. 278-287), June, 1999.
- The slides on reward shaping are based on the mini-project presentation in the Fall 2020 course (Group members: Wenjing Zhou, Jiachen Liu, Xinhao Sun, Yueyang Zhang, Ruixuan Zhang).