Lecture 12: Actor-Critic Methods

Course: Reinforcement Learning Theory
Instructor: Lei Ying
Department of EECS
University of Michigan, Ann Arbor

Policy iteration

Given policy μ_k,

- complete policy evaluation to compute J_{μ_k}
- complete policy improvement to obtain $J_{\mu_{k+1}}$

Impossible to have complete policy evaluation and policy improvement due to finite samples or function approximation.

Actor-Critic

Actor-critic: generalized policy-iteration

- Critic: estimate the value of the current policy (Q-function in general, because the actor needs the action values)
  TD(λ), double-Q, clipped double-Q, etc.
- Actor:
  - ϵ-greedy based on the current Q-function
  - Policy-gradient

Actor-Critic

Given Q-function (action value function),

$$J_{\theta^*} = \max_{\theta} E\left[\sum_a Q(x,a)\pi_\theta(a|x)\right]$$

$\pi_\theta(a|x)$: deterministic or stochastic policy parameterized by θ

Actor-Citic

Examples:

- finite action space. Gibbs policies:

$$\pi_\theta(a|x) = \frac{\exp\left(\theta^T \phi(x,a)\right)}{\sum_{u\in A} \exp\left(\theta^T \phi(x,u)\right)}$$

  where $\phi(x,u)$ is the feature vector.

- continuous action space. Gaussian policies:

$$\pi(a|x) = \frac{1}{\sqrt{(2\pi)^d \mathrm{det}(\Sigma_\theta(x))}} \exp\left\{-(a-\mu_\theta(x))^T \Sigma_\theta^{-1}(x)(a-\mu_\theta(x))\right\}$$

  where $\mu_\theta(x)$: mean (e.g. $\mu_\theta(x) = \phi^T(x)\theta$)
  $\Sigma_\theta(x)$: covariance (often $\Sigma_\theta(x) = \sigma^2 I$)

Actor-critic

- Goal: $\max_\theta J_\theta$

  e.g. $\theta \leftarrow \theta + \beta \nabla_\theta J_\theta$
- Question: how to compute $\nabla_\theta J_\theta$?

Actor-critic

Likelihood ratio trick:

$$\begin{aligned}\nabla_\theta \pi_\theta(a|s) &= \pi_\theta(a|s)\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \\ &= \pi_\theta(a|s)\nabla_\theta \log \pi_\theta(a|s)\end{aligned}$$

$\nabla_\theta \log \pi_\theta(a|x)$: score function

- Gibbs policy: $\nabla_\theta \log \pi_\theta(a|x) = \phi(x,a) - E_{\pi_\theta}[\phi(x,\cdot)]$
- Gaussian policy: $\nabla_\theta \log \pi_\theta(a|x) = \frac{(a-\mu(x))\phi(x)}{\sigma^2}$ where $\mu(x) = \phi^T(x)\theta$

Policy gradient

Policy Gradient Theorem

$$
\begin{aligned}
\nabla_\theta J_\theta =& E_{x_0 \sim \rho_0, u_k \sim \pi_\theta(u_k|x_k)} \left[\sum_{k=0}^{\infty} \alpha^k \nabla_\theta \log \pi_\theta(u_k|x_k) Q_\theta(x_k, u_k) \right] \\
=& E_{x \sim \rho_\theta, a \sim \pi_\theta(a|x)} [\nabla_\theta \log \pi_\theta(a|x) Q^{\pi_\theta}(x, a)],
\end{aligned}
$$

where $\rho_0(x)$ is the initial distribution of the states and

$$
\rho_\theta(x) = \sum_{k=0}^{\infty} \alpha^k \Pr\left(x_k = x\right),
$$

called (improper) discounted state distribution.

Policy-Gradient Theorem

- Recall $\pi_\theta(u|x)$ denotes a randomized (or deterministic) policy with parameter θ,

$$\pi_\theta(u|x) = P_\theta(\text{action} = u|\text{state} = x)$$

- Recall that

$$J_\theta(i) = E\left[\sum_{k=0}^{\infty} \alpha^k r(x_k, u_k)|x_0 = i\right]$$

- Let h be a sample path, i.e.

$$h = \{(x_0, u_0), (x_1, u_1), (x_2, u_2), \dots\} \text{ with } x_0 = i$$

and $h_t = \{(x_0, u_0), \dots, (x_t, u_t)\}$

Policy-Gradient Theorem

Then,

$$J_\theta(i) = \sum_{t=0}^{\infty} \alpha^t \left(\sum_{h_t} P_\theta(h_t) r(x_t, u_t) \right)$$

$$\begin{aligned}
\nabla J_\theta(i) &= \sum_{t=0}^{\infty} \alpha^t \left(\sum_{h_t} \nabla_\theta P_\theta(h_t) r(x_t, u_t) \right) \\
&= \sum_{t=0}^{\infty} \alpha^t \left(\sum_{h_t} \frac{\nabla_\theta P_\theta(h_t)}{P_\theta(h_t)} r(x_t, u_t) P_\theta(h_t) \right) \\
&= \sum_{t=0}^{\infty} \alpha^t E_{h_t}[\nabla \log P_\theta(h_t) r(x_t, u_t) | x_0 = i]
\end{aligned}$$

Policy-gradient algorithm

where

$$\nabla_\theta \log P_\theta(h_t) = \nabla_\theta(\log(\pi_\theta(u_0|x_0) \underbrace{P_{x_0x_1}(u_0)}_{\text{independent of } \theta} \dots \pi_\theta(u_t|x_t)))$$

$$= \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(u_k|x_k)$$

Thus,

$$\nabla_\theta J(i) = \sum_{t=0}^{\infty} \alpha^t E_{h_t} \left[r(x_t, u_t) \sum_{k=0}^{t} \nabla_\theta \log \pi_\theta(u_k|x_k) \middle| x_0 = i \right]$$

$$= E_h \left[\sum_{k=0}^{\infty} \nabla_\theta \log \pi_\theta(u_k|x_k) (\sum_{t=k}^{\infty} \alpha^t r(x_t, u_t)) | x_0 = i \right]$$

Policy-gradient algorithm

$$= \sum_{k=0}^{\infty} E\left[\nabla_\theta \log \pi_\theta(u_k|x_k)\left(\sum_{t=k}^{\infty} \alpha^t r(x_t, u_t)\right)|x_0 = i\right]$$

$$= \sum_{k=0}^{\infty} \sum_{x_k, u_k} E[\nabla_\theta \log \pi_\theta(u_k|x_k) \sum_{t \geq k} \alpha^t r(x_t, u_t)|x_k, u_k] \Pr(x_k, u_k|x_0 = i)$$

$$= \sum_{k=0}^{\infty} \sum_{x_k, u_k} \nabla_\theta \log \pi_\theta(u_k|x_k)\alpha^k Q_\theta(x_k, u_k) \Pr(x_k, u_k|x_0 = i)$$

$$= E\left[\sum_{k=0}^{\infty} \alpha^k \nabla_\theta \log \pi_\theta(u_k|x_k) Q_\theta(x_k, u_k)\middle| x_0 = i\right]$$

$$= \sum_{u,x} \nabla_\theta \log \pi_\theta(u|x) Q_\theta(x, u)\left(\sum_{k=0}^{\infty} \alpha^k \Pr\left(x_k = x|x_0 = i\right)\right)\pi_\theta(u|x)$$

References

- Chapter 4.4 of Csaba Szepevari, *Algorithms for Reinforcement Learning*, Morgan Claypool, 2010.
- R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation.* NeurIPS, 1999.
- The proof of the policy-gradient theorem is based on R. Srikant's lecture notes on *Policy Gradient Algorithm* available at `https://sites.google.com/illinois.edu/mdps-and-rl/lectures?authuser=1`

Acknowledgements: I would like to thank Alex Zhao for helping prepare the slides, and Honghao Wei and Zixian Yang for correcting typos/mistakes.