Lecture 13: Policy-gradient algorithm

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

Policy gradient

Policy Gradient Theorem

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

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

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

called (improper) discounted state distribution.

Policy gradient

REINFORCE (Williams (1988, 1992)):

- Given an episode $x_0, a_0, x_1, a_1, \cdots, x_{T-1}, a_{T-1}, x_T$, starting from $t = T$ backwards (i.e. $t = T, T-1, \cdots, 0$), update θ as follows:

$$\theta \leftarrow \theta + \beta \nabla_\theta \log \pi_\theta(a_t|x_t) \sum_{\tau=t}^{T} \alpha^{\tau-t} r_t.$$

- Monte-Carlo policy: $Q_\theta(x_t, a_t) \approx \sum_{\tau=t}^{T} \alpha^{\tau-t} r_t$.

Variance reduction (Control Variates Method)

We are interested in computing

$$E[f(x)] \approx \underbrace{\frac{1}{N}\sum_{i=1}^{N} f(x_i)}_{F} \quad x_i \sim P(x)$$

But F may have high variance.

- Solution: replace F with F' such that

$$E[F] = E[F'],\ Var(F') \leq Var(F)$$

Variance reduction

Consider function $\phi(x)$ such that $E[\phi(x)] = 0$,

$$E[f(x) - \phi(x)] = E[f(x)]$$

$$Var(f(x) - \phi(x)) = Var(f(x)) - 2Cov(f(x), \phi(x)) + Var(\phi(x))$$

- The variance can be reduced when $\phi(x)$ is strongly correlated with $f(x)$.

Variance reduction

Note that

$$
\begin{aligned}
& E[\nabla_\theta \log \pi_\theta(u_k|x_k) b(x_k) | x_0 = i] \\
= & \sum_{x_k} b(x_k) P_\theta(x_k|x_0 = i) \Big(\sum_{u_k} \nabla_\theta \log \pi_\theta(u_k|x_k) \pi_\theta(u_k|x_k)\Big) \\
= & \sum_{x_k} b(x_k) P_\theta(x_k|x_0 = i) \Big(\sum_{u_k} \frac{\nabla_\theta \pi_\theta(u_k|x_k)}{\cancel{\pi_\theta(u_k|x_k)}} \cancel{\pi_\theta(u_k|x_k)}\Big) \\
= & \sum_{x_k} b(x_k) P_\theta(x_k|x_0 = i) \underbrace{\nabla_\theta \Big(\cancelto{1}{\sum_{u_k} \pi_\theta(u_k|x_k)}\Big)}_{0} \\
= & 0
\end{aligned}
$$

Variance reduction

Estimate $\nabla_{\theta_t} J(i)$ as

$$\nabla_{\theta_t} J(i) = \sum_{k=0}^{T} \alpha^k \nabla \log \pi_{\theta_t}(u_k|x_k) \times (\underbrace{r(x_k, u_k) + \alpha V_{\theta_t}(x_{k+1}) - V_{\theta_t}(x_k)}_{\text{TD error}})$$

$$\theta_{t+1} = \theta_t + \beta_t \nabla J_{\theta_t}(i)$$

Policy gradient

- Function approximation:

$$\nabla_\theta J_\theta \approx E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a|x)\hat{Q}_{\underset{\text{weights}}{\underline{w}}}(x,a)]$$

- SGD:

$$\begin{aligned}\theta &\leftarrow \theta + \beta \nabla_\theta J_\theta \\ &= \theta_t + \beta_t \nabla \theta_t \log \pi_{\theta_t}(a_t|x_t)\hat{Q}_{w_t}(x_t, a_t)\end{aligned}$$

- REINFORCE (Williams's (1988, 1992)): Monte-Carlo policy

$$\hat{Q}(x_t, a_t) = \sum_{\tau=t}^{\infty} \alpha^{\tau-t} r_t = V_t$$

Actor-Critic

- Advantage Actor-Critic:

$$A(s,a) = \hat{Q}_w(s,a) - \hat{V}_v(s)$$

- TD Actor-Critic:

$$A(s,a) = r + \alpha \hat{V}_v(s') - \hat{V}_v(s)$$

Note: TD error estimates the advantage function.

Actor-Critic

- Natural Actor-Critic (parametrization independent):
  Note that $A(s, a)$ depends on policy parameter.

$$\nabla_\theta J(\theta) = E[\nabla_\theta \log \pi_\theta(a|x) A^{\pi_\theta}(x, a)]$$

- Natural Actor-Critic:

$$\nabla_\theta^{nat} \pi_\theta(a|x) = G_\theta^{-1} \nabla_\theta \pi_\theta(a|x)$$

$$\Downarrow$$

$$\nabla_\theta^{nat} J(\theta) = G_\theta^{-1} \nabla_\theta J(\theta)$$

where G_θ is the Fisher information matrix,

$$G_\theta = E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a|x) \nabla_\theta \log \pi_\theta(a|x)]$$

Actor-Critic

- Policy gradient

$$\begin{array}{rc} & \max_\theta E\left[\log \pi_\theta(a|x) A_\theta(x,a)\right] \\ \text{subject to:} & \|\tilde{\theta} - \theta\| \le \epsilon \end{array}$$

- Natural policy gradient

$$\begin{array}{rc} & \max_\theta E\left[\log \pi_\theta(a|x) A_\theta(x,a)\right] \\ \text{subject to:} & \mathsf{KL}(\pi_{\tilde{\theta}}, \pi_\theta) \le \epsilon \end{array}$$

Actor-Critic with Neural Networks

NN implementation:

- Critic: double-Q, target-Q, clipped-Q
- Actor: Weighted cross-entropy loss

$$L = -\sum_{(x,a)} A(x,a) \log \pi_\theta(a|x)$$

- Use

$$A_w(x,a) = (\nabla_\theta \log \pi_\theta(a|x))^T w$$

Score function as features:

$$\nabla A_w(x,a) = \nabla_\theta \log \pi_\theta(a|x)$$

$$\nabla_\theta J(\theta) = E[\nabla_\theta \log \pi_\theta(a|x) \nabla_\theta \log \pi_\theta(a|x)^T w] = G_\theta w$$

$$\implies \nabla_\theta^{nat} J(\theta) = w.$$

Actor-Critic with Neural Networks

- Deep deterministic policy gradient (DDPG) (Lillicrap et al. 2016):

$$J(\theta) \approx E[Q_w(x, \mu_\theta(x))]$$
$$\nabla_\theta J(\theta) = E[\nabla_a Q_w(x, a)|_{a=\mu_\theta(x)} \nabla_\theta \mu_\theta(x)]$$

  $\mu_\theta(x)$: deterministic policy

  - Implementation: Loss function

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

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

References

- Chapter 4.4 of Csaba Szepevari, *Algorithms for Reinforcement Learning,* Morgan Claypool, 2010.
- Ronald J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning*. Machine Learning, 8(3):229-256, 1992.
- Sutton, McAllester, Singh, Mansour, *Policy gradient methods for reinforcement learning with function approximation: actor-critic algorithms with value function approximation*, 1999.
- S. Kakade, *A natural policy gradient*. NeurIPS, 2002.
- David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, Martin Riedmiller, *Deterministic policy gradient algorithms*. ICML. 2014.
- 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).

References (cont'd)

- John Schulman, et al.*High-dimensional continuous control using generalized advantage estimation.* ICLR 2016

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