MLDS 2019 Spring HW4-3 - Actor-Critic

2019/06/08

adlxmlds@gmail.com

Time Schedule

- June 18th 19:30 Deadline
- June 22th 10:30 Deadline

Outline

	HW4-1	HW4-2	HW4-3
Pong	PG		AC
Breakout		DQN	AC
Improved Version			

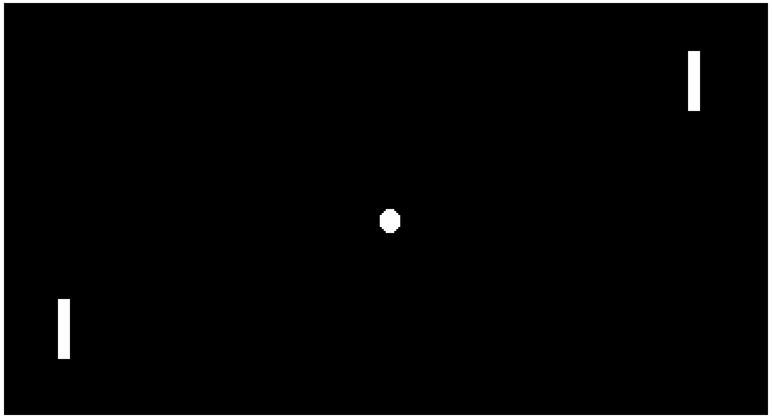
Outline

- Environment
- Actor-Critic
- Grading & Format
  - Grading Policy
  - Code Format
  - Report
  - Submission

Environment

Breakout

Pong



Actor Critic

online actor-critic algorithm:

1. take action $\mathbf{a} \sim \pi_\theta(\mathbf{a}|\mathbf{s})$, get $(\mathbf{s}, \mathbf{a}, \mathbf{s}', r)$
2. update $\hat{V}^\pi_\phi$ using target $r + \gamma \hat{V}^\pi_\phi(\mathbf{s}')$
3. evaluate $\hat{A}^\pi(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \gamma \hat{V}^\pi_\phi(\mathbf{s}') - \hat{V}^\pi_\phi(\mathbf{s})$
4. $\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(\mathbf{a}|\mathbf{s}) \hat{A}^\pi(\mathbf{s}, \mathbf{a})$
5. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

Improvements to Actor-Critic

- DDPG : https://arxiv.org/abs/1509.02971
- ACER : https://arxiv.org/pdf/1611.01224.pdf
- A3C : https://arxiv.org/abs/1602.01783
- A2C
- ACKTR : https://arxiv.org/abs/1708.05144
- PPO：https://arxiv.org/abs/1707.06347

https://zhuanlan.zhihu.com/p/50343077

Sample Efficeint Actor-critic with Experience Replay (ACER)

- Off policy → sample efficient, experience replay
- How? → By importance sampling

$$g^{\text{marg}} = \mathbb{E}_{x_t \sim \beta, a_t \sim \mu}\left[\rho_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t)\right], \quad \rho_t = \frac{\pi(a_t|x_t)}{\mu(a_t|x_t)}$$

- Problem : High variance cause by ρ_t
  - Use clipping → trade bias for variance

$$\bar{\rho}_t = \min\{c, \rho_t\}$$

- To correct bias, add a correction term

$$\begin{aligned} g^{\text{marg}} &= \mathbb{E}_{x_t a_t}\left[\rho_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t)\right] \\ &= \mathbb{E}_{x_t}\left[\mathbb{E}_{a_t}[\bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^\pi(x_t, a_t)] + \mathbb{E}_{a \sim \pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t) Q^\pi(x_t, a)\right)\right] \end{aligned}$$

$[x]_+ = x$ if $x > 0$ and it is zero otherwise.

Sample Efficeint Actor-critic with Experience Replay (ACER)

- How to evaluate Q^ π under samples from μ ?
  - Use Retrace Estimator

$$Q^{\text{ret}}(x_t, a_t) = r_t + \gamma \bar{\rho}_{t+1}[Q^{\text{ret}}(x_{t+1}, a_{t+1}) - Q(x_{t+1}, a_{t+1})] + \gamma V(x_{t+1})$$

  - And loss gradient

$$(Q^{\text{ret}}(x_t, a_t) - Q_{\theta_v}(x_t, a_t))\nabla_{\theta_v} Q_{\theta_v}(x_t, a_t))$$

$$\widehat{g}^{\text{marg}} = \mathbb{E}_{x_t}\left[\mathbb{E}_{a_t}[\bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t) Q^{ret}(x_t, a_t)] + \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t) Q_{\theta_v}(x_t, a)\right)\right]$$

Reduce variance by value function baseline yields

$$\begin{aligned}\widehat{g}_t^{\text{acer}} &= \bar{\rho}_t \nabla_\theta \log \pi_\theta(a_t|x_t)[Q^{\text{ret}}(x_t, a_t) - V_{\theta_v}(x_t)] \\ &+ \mathop{\mathbb{E}}_{a\sim\pi}\left(\left[\frac{\rho_t(a) - c}{\rho_t(a)}\right]_+ \nabla_\theta \log \pi_\theta(a|x_t)[Q_{\theta_v}(x_t, a) - V_{\theta_v}(x_t)]\right)\end{aligned}$$

Algorithm (ACER)

Algorithm 1 ACER for discrete actions (master algorithm)

// Assume global shared parameter vectors θ *and* θ_v.
// Assume ratio of replay r.
repeat
  Call ACER on-policy, Algorithm 2.
  $n \leftarrow \text{Possion}(r)$
  for $i \in \{1, \cdots, n\}$ **do**
    Call ACER off-policy, Algorithm 2.
  end for
until Max iteration or time reached.

Algorithm (ACER)

Algorithm 2 ACER for discrete actions

Reset gradients $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$.
Initialize parameters $\theta' \leftarrow \theta$ and $\theta'_v \leftarrow \theta_v$.
if not On-Policy **then**
  Sample the trajectory $\{x_0, a_0, r_0, \mu(\cdot|x_0), \cdots, x_k, a_k, r_k, \mu(\cdot|x_k)\}$ from the replay memory.
else
  Get state x_0
end if
for $i \in \{0, \cdots, k\}$ **do**
  Compute $f(\cdot|\phi_{\theta'}(x_i))$, $Q_{\theta'_v}(x_i, \cdot)$ and $f(\cdot|\phi_{\theta_a}(x_i))$.
  if On-Policy **then**
    Perform a_i according to $f(\cdot|\phi_{\theta'}(x_i))$
    Receive reward r_i and new state x_{i+1}
    $\mu(\cdot|x_i) \leftarrow f(\cdot|\phi_{\theta'}(x_i))$
  end if
  $\bar{\rho}_i \leftarrow \min\left\{1, \frac{f(a_i|\phi_{\theta'}(x_i))}{\mu(a_i|x_i)}\right\}$.
end for

$$Q^{ret} \leftarrow \begin{cases} 0 & \text{for terminal } x_k \\ \sum_a Q_{\theta'_v}(x_k, a) f(a|\phi_{\theta'}(x_k)) & \text{otherwise} \end{cases}$$

for $i \in \{k-1, \cdots, 0\}$ **do**
  $Q^{ret} \leftarrow r_i + \gamma Q^{ret}$
  $V_i \leftarrow \sum_a Q_{\theta'_v}(x_i, a) f(a|\phi_{\theta'}(x_i))$
  Computing quantities needed for trust region updating:

$$\begin{aligned} g &\leftarrow \min\{c, \rho_i(a_i)\} \nabla_{\phi_{\theta'}(x_i)} \log f(a_i|\phi_{\theta'}(x_i))(Q^{ret} - V_i) \\ &\quad + \sum_a \left[1 - \frac{c}{\rho_i(a)}\right]_+ f(a|\phi_{\theta'}(x_i)) \nabla_{\phi_{\theta'}(x_i)} \log f(a|\phi_{\theta'}(x_i))(Q_{\theta'_v}(x_i, a_i) - V_i) \\ k &\leftarrow \nabla_{\phi_{\theta'}(x_i)} D_{KL}\left[f(\cdot|\phi_{\theta_a}(x_i)) \| f(\cdot|\phi_{\theta'}(x_i))\right] \end{aligned}$$

  Accumulate gradients wrt θ': $d\theta' \leftarrow d\theta' + \frac{\partial \phi_{\theta'}(x_i)}{\partial \theta'}\left(g - \max\left\{0, \frac{k^T g - \delta}{\|k\|_2^2}\right\} k\right)$
  Accumulate gradients wrt θ'_v: $d\theta_v \leftarrow d\theta_v + \nabla_{\theta'_v}(Q^{ret} - Q_{\theta'_v}(x_i, a))^2$
  Update Retrace target: $Q^{ret} \leftarrow \bar{\rho}_i \left(Q^{ret} - Q_{\theta'_v}(x_i, a_i)\right) + V_i$
end for
Perform asynchronous update of θ using $d\theta$ and of θ_v using $d\theta_v$.
Updating the average policy network: $\theta_a \leftarrow \alpha\theta_a + (1-\alpha)\theta$

Submission

- Submit your presentation files to: google drive

Grading & Format

Slides

- Describe your actor-critic model on Pong and Breakout
- Plot the learning curve and compare with 4-1 and 4-2 to show the performance of your actor-critic model on Pong & Breakout
  - X-axis: number of time steps
  - Y-axis: average reward in last 100 episodes
- Reproduce 1 improvement method of actor-critic (Allow any resource)
  - Describe the method
  - Plot the learning curve and compare with 4-1 and 4-2, 4-3 to show the performance of your improvement

Related Materials

- Course & Tutorial:
    - Berkeley Deep Reinforcement Learning, Fall 2017
    - David Silver RL course
    - Nips 2016 RL tutorial
- Blog:
    - Andrej Karpathy’ s blog
    - Arthur Juliani’ s Blog
- Text Book:
    - Reinforcement Learning: An Introduction
- Repo:
    - https://github.com/williamFalcon/DeepRLHacks