Lecture 13: Advanced Actor Critic Methods

Logistics:

- HW5 has been released. Due dates for this and other assignments will be on Mondays, per popular vote
- Final project description will be released soon. Find a partner! Post on Ed if you're looking for a partner.
- Midterm has been graded. Do review the questions you got wrong.

Where are we going? The second half of the semester will be a mix of advanced topics. Many of these questions are still under active development, so we might not know the correct answer to all of them. We will look a few categories of things:

- Building effective algorithms
  - Practical considerations for implementing actor critic methods (today's lecture)
  - How models can help
  - How to do exploration?
  - How to borrow ideas from generative AI (inference, LLMs)
- Different problem settings:
  - multi-agent RL (alpha star)
  - RLHF
  - Offline RL
- Student presentations!

1 Review: generalized policy improvement

- Estimate the Q function. Mention connection with FQI
- Optimize the policy using the Q function. This can be done in two ways. For discrete actions, just take max. For continuous actions, do gradients.

2 Implementing Actor Critic Methods

Recall FQI

$$y_i = \max_a (r(s,a) + \gamma \mathbb{E}[V(s')]) \tag{1}$$

$$\min_\phi \, (V_\phi(s_i) - y_i)^2 \tag{2}$$

Recall generalized policy improvement

- policy architecture. Shared layers for actor and critic?
- critic architecture
- double DQN, dueling networks
- replay buffers
- target networks
- exploration: OU noise, Gaussian noise, noisy nets, parameter space noise
- note that we're using off-policy methods
- n-step returns no gradients through target network

- common benchmarks
- TD3 multiple Q networks trick

conceptual point: we're doing dynamic programming, but we're making a function's output for one input be similar to it's output at another input

Advice on experimenting with methods:

- Start with something that works. With every change, ensure that the method still works
- Implementing from scratch is really difficult. Avoid at all costs (except in this course)
- start with easy tasks (but, note that everything works on cartpole)
- Run experiments on multiple random seeds. Results can be much, much noisier than supervised learning methods

2.1 Comparing Common Methods

DDPG.

SVG(0)

NAF

TD3 [?] is DDPG with three tricks: additive clipped noise on actions, double critics and actors, delayed actors update.

SAC [?] is DDPG with an extra entropy term (more on this in a future lecture). When TD3 came out, SAC was re-implemented on top of TD3 and got improved performance.

3 Analysis

Maximization bias:

$$\mathbb{E}[\max(x_1, x_2)] > \max(\mathbb{E}[x_1], \mathbb{E}[x_1]) = \mu \tag{3}$$

Reparametrization trick

Deterministic policy gradient theory

4 Advanced Topics

- HER
- handling random seeds

References

Learning Objectives

- Replay buffers
- Target Networks
- N-step returns

TD3 Fujimoto 18

~~3 key~~ (2)

4 ~~key~~ Key Ideas:

a) Target networks

warm up - rock paper scissors

~~Q(s,a)~~

$$y = r_i + \gamma Q_{old}(s, a \sim \pi(a|s))$$

$$(Q_{new} - y)^2$$

$$\theta^{old} \leftarrow \tau \theta^{new} + (1-\tau)\theta^{old}$$

$$\tau = 0.01$$

b) Replay Buffers

$\tau \sim \pi$ → $B \cup \{\tau\}$

update(τ) → update B

d) Learn 2 Q networks

Q_1, Q_2. $y = r_i + \gamma \min_{j=1,2} Q^j_{old}(s, a)$

c) N-step returns

$$r_i + \gamma r_{i+1} + \ldots + \gamma^N r_{i+N} + \gamma^{N+1} V_\theta(s_{i+N})$$

slides

SPG(0)

A2C

DDPG

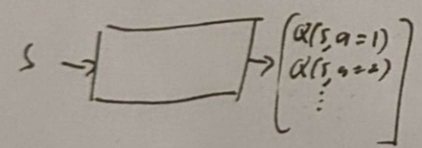
TD3

(1) Gen. Policy Iter/Actor Critic

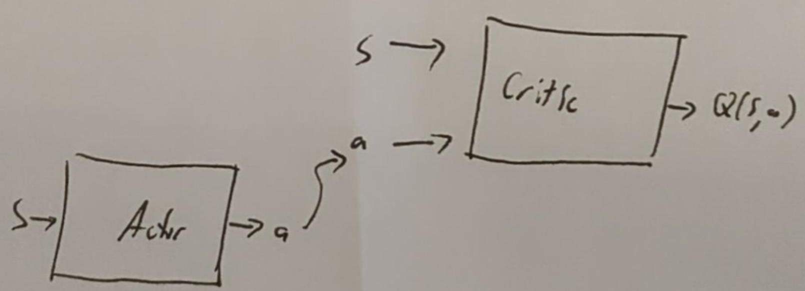
update Q ⟷ update π w/ Q

Arch

~~DQN~~ Discrete Actn



~~DDPG~~ Cts Actn



Exploration

~~Option 1 (DDPG): Deterministic Actor~~

Critic loss: $y_i = r_i + \gamma Q(s_{i+1}, a = \pi(s_{i+1}))$

$$\frac{1}{N}\sum_i (y_i - Q(s_i, a_i))^2$$

Actor loss $\max_\pi \left(Q(s, a = \pi(s)) \right)$

$$\nabla_\theta = \nabla_a Q(s,a) \nabla_\theta \pi_\theta(s)$$

Stoch. Actor:

$$\max_\pi \mathbb{E}_{a \sim \pi(a|s)}[Q(s,a)]$$

$$\nabla_\theta = \mathbb{E}_{a \sim \pi(a|s)}[Q(s,a) \nabla_\theta \log \pi(a|s)]$$

Q: Do stoch. policies get high rewards

~~Exploration?~~

(3) Videos

(4) Advice on experiments.