Reinforcement Learning

Policy Gradient-class 2

袁路展

北京邮电大学

2021 年 11 月 1 日

- AC 的结构、优势、缺陷。
- Advantage PG 解决的问题、理论最优的 baseline。
- 重要性采样

Policy Gradient

- Value-based 的方法很难在高维、连读动作空间中生成一个好的策略。
- Policy-based 方法直接优化策略。
- Policy-based 方法中核心问题是怎样评估策略。

Policy Gradient Theorem

在策略梯度中，我们的目标是优化：

$$J(\theta) = \mathbb{E}_{\tau\sim\rho_\theta}\left[\sum_{t=0}^{T}\gamma^t r\left(s_t, a_t, s_{t+1}\right)\right] = \int_\tau \rho_\theta(\tau)R(\tau)d\tau \tag{1}$$

使用数学方法对 $\rho_\theta(\tau)$ 进行分解，我们最终得到的 Policy Gradient Theorem 如下：

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} \rho^{\pi_\theta}(s)\int_{\mathcal{A}} \nabla_\theta \pi_\theta(a \mid s)Q^\pi(s,a)dads \tag{2}$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)Q^{\pi\theta}(s,a)\right] \tag{3}$$

后面主要用（3）式来表示。

Policy Gradient Theorem

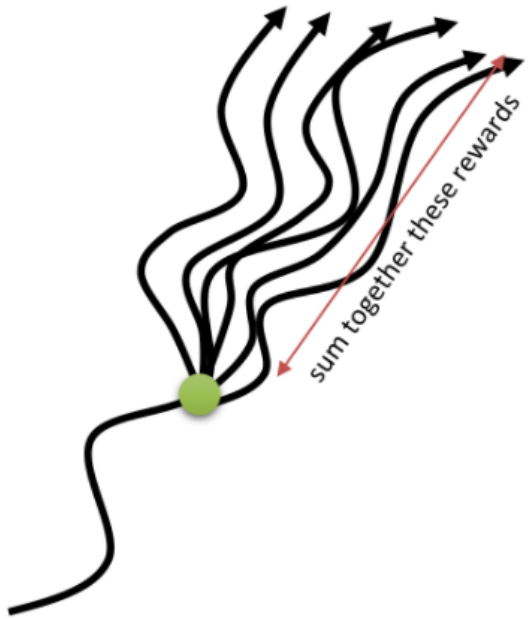
$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)\right]$$

它不在考虑整条 trajectory，更加关注在单个的 (s,a) 上。
注意在策略梯度里面，影响策略梯度的两个关键量：π 和 Q。

- π 是策略本身，就是我们要优化的方向.
- Q 是对后续路径的回报的期望，这个是我们本节围绕展开的点，AC、advantage PG、off-policy，都是围绕这个 Q 展开的。

High variance in PG

在策略梯度里面，存在高方差问题



考虑对于当前这一步的策略梯度，在计算的时候，我们需要计算后续的所有轨迹的回报的期望值。
但是在策略、环境的 transition 和 reward 都可能存在随机性，这会导致在这一步后面，可能会有不同的 return，Q 的估计不准确，会导致策略梯度存在方差。

High variance in PG

这也对应上节课讲的，high variance 的问题会在以下情况下更加严重

高维动作空间：后面的轨迹空间宽度广。

长程问题（long horizon）：后面的轨迹空间深度大。

有限样本：已知的东西太少了。

Intro to AC

回到策略梯度公式

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)\right]$$

问题是 $Q^{\pi_\theta}(s,a)$ 是未知的。
对后续的轨迹进行采样，求平均的方法显然在很多情况下不能实现。
结合之前的 value-based 的方法，直接学习 Q 的函数，就不用对后续的进行大规模采样。
策略梯度就变成：

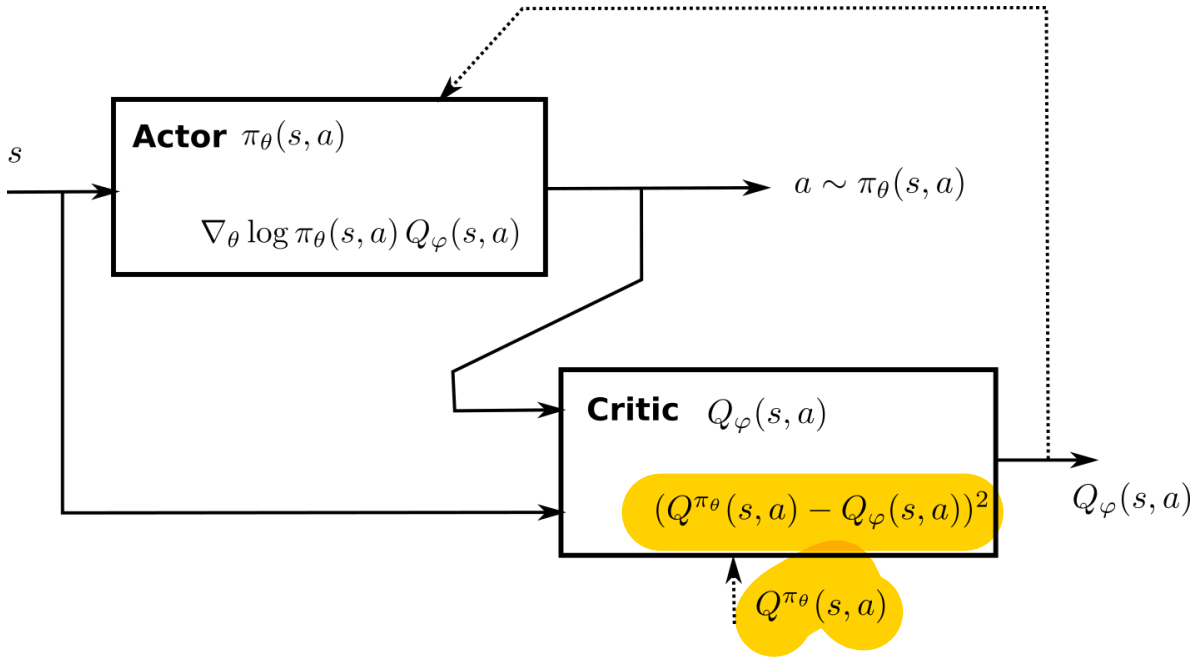
$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi\theta}\left[\nabla_\theta \log \pi_\theta(s,a) Q_\varphi(s,a))\right] \tag{4}$$

Actor-Critic

Actor-Critic 架构

Actor π 通过等式（4）进行参数更新

Critic 使用 value based 的方法来更新 Q



AC 架构的优缺点

优点

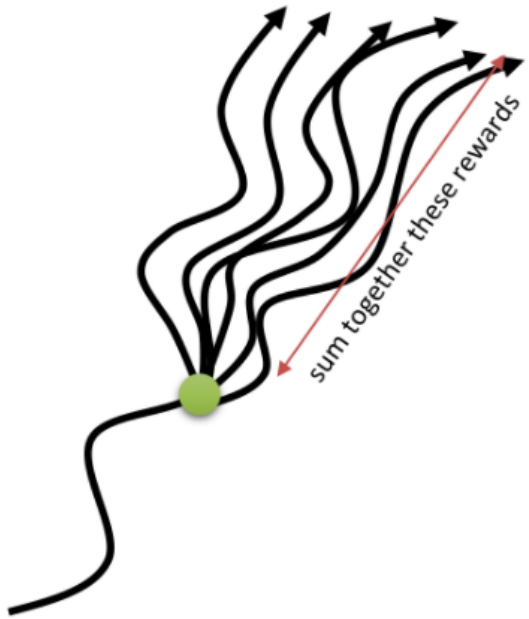
- Trajectories 在策略梯度里面没有了，之前的方法需要轨迹后面的评级，现在可以直接用其他网络计算得到的 Q 来评估。
- 我们可以引入 off-policy 的方法，因为评估策略可以用其他的网络，可以使用 replay buffer 来存储 transition。

缺点

- 高方差的问题仍旧存在。

High variance in PG

在策略梯度里面，存在高方差问题



对这个当前策略，后续执行得到的 return 的分布的估计问题。我们在计算策略梯度的时候，是从后面所有轨迹的分布中采样得到一个，然后计算梯度来更新的，缩小方差可以让所有采样得到的 s 下的策略梯度更加集中。

A simple demo

回到策略梯度公式

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)\right] \tag{5}$$

考虑策略梯度中的 $\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)$

- 其中 $\nabla_\theta \log \pi_\theta(s,a)$ 与后续的轨迹没有任何关系，就是在当前的 state，当前策略采取每个 action 的概率 $p(a|s)$.
- 后面的 Q(s,a) 代表执行这个 (s,a) 对应后续回报的期望，控制的是让策略向哪个 action 趋近的方向。

Demo of Baseline

减去一个 baseline，减小 $Var[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)]$
一个简单的 demo

action	a	b	c
概率 p	$0.5+\theta^2$	$0.3-0.5\theta^2$	$0.2-0.5\theta^2$
Q1	0	1	2
Q2	1000	1001	1002
$\nabla_\theta \log \pi_\theta(s,a) Q1^{\pi_\theta}(s,a)$	0	$-\theta$	$-\theta$
$\nabla_\theta \log \pi_\theta(s,a) Q2^{\pi_\theta}(s,a)$	2000θ	-1001θ	-1002θ

显然 Q2 下的策略梯度方差比 Q1 下的策略梯度方差大很多。
为了减小策略梯度的方差，使得不同 sa 采样下得到的策略梯度都更加集中与 Q 大的 action，一种很直接的方法是减去一个 baseline，来提高策略梯度向 Q 大的动作集中。

Analysis on baseline

减去 baseline 会改变策略梯度吗？
减去 baseline 之后，最原始的策略梯度就变成：

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)(Q^{\pi_\theta}(s,a) - b)\right]$$

证明：

$$\begin{aligned}
&\mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) b\right] \\
&= \int \rho_\theta(s) \int \pi_\theta(s,a) \nabla_\theta \log \pi_\theta(s,a) b \, da \, ds \\
&= \int \rho_\theta(s) \int \nabla_\theta \pi_\theta(s,a) b \, da \, ds \\
&= b\nabla_\theta \int \rho_\theta(s) \int \pi_\theta(s,a) \, da \, ds = b\nabla_\theta 1 = 0
\end{aligned}$$

Optimal Baseline

减去 baseline 不会改变策略梯度，那么设置 baseline 是多少合适？

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)(Q^{\pi_\theta}(s,a) - b)\right]$$

$$\begin{aligned}
&\mathrm{Var}\left[\nabla_\theta \log \pi_\theta(s,a)(Q^{\pi_\theta}(s,a) - b)\right] \\
&= \mathbb{E}\left[(g(\theta)(Q^{\pi_\theta}(s,a) - b))^2\right] - \mathbb{E}\left[(g(\theta)(Q^{\pi_\theta}(s,a) - b))\right]^2 \\
&= \mathbb{E}\left[(g(\theta)(Q^{\pi_\theta}(s,a) - b))^2\right]
\end{aligned}$$

$$\frac{d\,\mathrm{Var}}{db} = \frac{d}{db}\mathbb{E}\left[g(\theta)^2(Q^{\pi_\theta}(s,a) - b)^2\right]$$

展开二次项，根据期望计算法则，最后得到最优的 b

$$b = \frac{\mathbb{E}\left[g(\theta)^2 Q^{\pi_\theta}(s,a)\right]}{\mathbb{E}\left[g(\theta)^2\right]}$$

advantage function

在实际操作中，往往使用 V(s) 来作为 baseline
引入 advantage function：

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \tag{6}$$

把 advantage function 放到策略梯度里面就是

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{s\sim\rho_{\theta}, a\sim\pi_{\theta}} \left[\nabla_{\theta} \log \pi_{\theta}(s,a) A^{\pi}(s,a)\right] \tag{7}$$

放到 AC 架构里面就是：

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{s\sim\rho^{\pi}, a\sim\pi_{\theta}} \left[\nabla_{\theta} \log \pi_{\theta}(s,a) A_{\varphi}(s,a)\right] \tag{8}$$

代表算法有：A2C、A3C、GAE、MAGE

on-policy and off policy

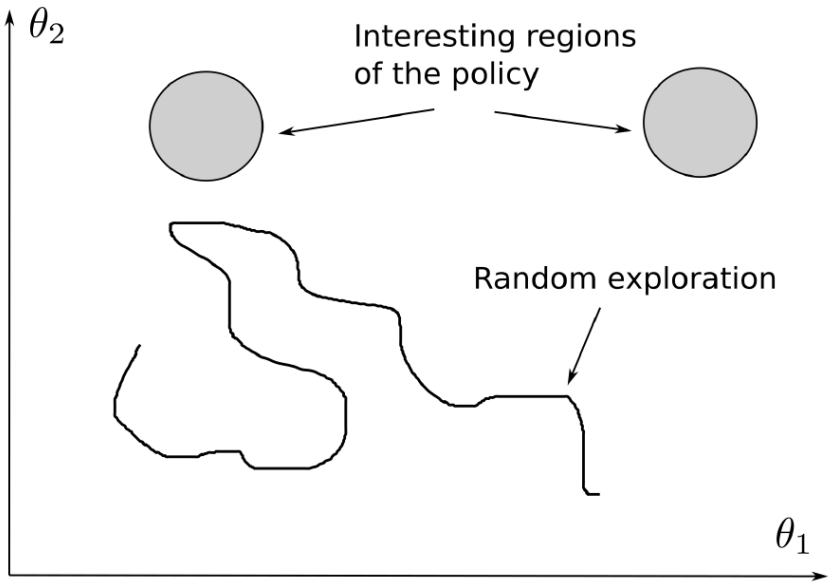
on-policy：优化出来是什么策略，就执行什么策略。
off-policy：优化出来的策略和执行的策略不一样。

- on policy，我心里知道我喝热水对治疗感冒有好处，我就喝热水.
- off-policy。我心里知道我喝热水对治疗感冒有好处，但是我选择不喝热水，喝冷水，探索喝冷水的结果，可能喝冷水对感冒更有用，不管结果如何，都让我心里知道的东西更多了。

最简单的区别 on-policy 和 off-policy 的两个算法：
SARSA 和 Q-learning

Disadvantages of on-policy

探索问题：
在 on-policy 的算法中，actor 的策略直接由梯度更新产生，很难保证对" 感兴趣" 区域的探索。



.png .bb

这问题在高维空间和稀疏奖励环境下会更加严重。

Disadvantages of on-policy

样本效率问题：

- 只能用当前策略采样的得到的样本，不能其他策略采集到的数据。
- 所有的数据用一次之后就丢弃了，但其实神经网络的方法，需要很多次训练才能最终达到收敛。

Advantage of off-policy

off-policy 的优点

- 更好的探索能力
- 可以用 experience replay memory

如果 actor 采取的动作不是根据同一个策略产生的，那么 critic 产生的反馈将会有带有误差。

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)\right]$$

怎么用 off-policy 的样本来评价策略？

重要性采样

- Off-policy 的方法学习一个 traget policy $\pi(s,a)$ 同时用一个 behavior policy $b(s,a)$ 来探索环境。
- Target policy $\pi(s,a)$ 用 behavior policy $b(s,a)$ 采集到的 tracjectory 来评估，这是否有问题？

回到最原始的策略梯度，我们的目标是最大化 trajectory 的 return:

$$J(\theta) = \mathbb{E}_{\tau\sim\rho_\theta}[R(\tau)] = \int_\tau \rho_\theta(\tau)R(\tau)d\tau \approx \frac{1}{N}\sum_{i=1}^{N} R(\tau_i) \tag{9}$$

当 trajectory 是 traget policy 采集出来的时候，等式是对的，但是当 trajectory 是 behavior policy 采集到的时候，我们真正得到的其实是：

$$\hat{J}(\theta) = \mathbb{E}_{\tau\sim\rho_b}[R(\tau)] = \int_\tau \rho_b(\tau)R(\tau)d\tau \tag{10}$$

importance sampling

怎么建立（9）和（10）之间的关系？重要性采样

$$
\begin{aligned}
J(\theta) &= \mathbb{E}_{\tau\sim\rho_\theta}[R(\tau)] \\
&= \int_\tau \rho_\theta(\tau)R(\tau)d\tau \\
&= \int_\tau \frac{\rho_b(\tau)}{\rho_b(\tau)}\rho_\theta(\tau)R(\tau)d\tau \\
&= \int_\tau \rho_b(\tau)\frac{\rho_\theta(\tau)}{\rho_b(\tau)}R(\tau)d\tau \\
&= \mathbb{E}_{\tau\sim\rho_b}\left[\frac{\rho_\theta(\tau)}{\rho_b(\tau)}R(\tau)\right]
\end{aligned}
\tag{11}
$$

其中 $\frac{\rho_\theta(\tau)}{\rho_b(\tau)}$ 被称作 importance sampling weight。

importance sampling weight

策略梯度就变为：

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\rho_b}\left[\nabla_\theta \log \rho_\theta(\tau)\frac{\rho_\theta(\tau)}{\rho_b(\tau)}R(\tau)\right] \tag{12}$$

现在问题就是怎样评估 importance sampling weight。

$$\frac{\rho_\theta(\tau)}{\rho_b(\tau)} = \frac{p_0\left(s_0\right)\prod_{t=0}^{T}\pi_\theta\left(s_t,a_t\right)p\left(s_{t+1}\mid s_t,a_t\right)}{p_0\left(s_0\right)\prod_{t=0}^{T}b\left(s_t,a_t\right)p\left(s_{t+1}\mid s_t,a_t\right)} = \prod_{t=0}^{T}\frac{\pi_\theta\left(s_t,a_t\right)}{b\left(s_t,a_t\right)} \tag{13}$$

有了 importance sampling weight, 策略的评估就可以变为：

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\rho_b}\left[\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta\left(s_t,a_t\right)\left(\prod_{t'=0}^{t}\frac{\pi_\theta\left(s_{t'},a_{t'}\right)}{b\left(s_{t'},a_{t'}\right)}\right)Q_b(s,a)\right]. \tag{14}$$

Review on this class

Advanced Policy Gradient

- Actor Critic
  - 不在需要整条 trajectory，更关注单个的 transition

  Advantage PG
  - 解决 high variance 的问题。

  off-policy PG?
  - 让 PG 的方法不再局限于 on-policy，有更加充分的探索能力和更好的样本效率。

Questions

Q&A