总结二：GAE

2022年2月3日 12:27

- **在PPO、TRPO等类似算法中都 常 使用 GAE 技术，这是一种对 advantage function 更稳定的估计方法**
- 论文：[1506.02438] High-Dimensional Continuous Control Using Generalized Advantage Estimation (arxiv.org)
- 参考：【强化学习技术 28】GAE - 知乎 (zhihu.com)
- GAE（Generalized Advantage Estimation）
  - 在 policy gradient 中，通过advantage function，提高稳定性，寻求 unbiased 和 various 之间的平衡
  - 常见的 policy loss 有（总结）：

$$g = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right], \tag{1}$$

where Ψ_t may be one of the following:

1. $\sum_{t=0}^{\infty} r_t$: total reward of the trajectory.
2. $\sum_{t'=t}^{\infty} r_{t'}$: reward following action a_t.
3. $\sum_{t'=t}^{\infty} r_{t'} - b(s_t)$: baselined version of previous formula.
4. $Q^\pi(s_t, a_t)$: state-action value function.
5. $A^\pi(s_t, a_t)$: advantage function.
6. $r_t + V^\pi(s_{t+1}) - V^\pi(s_t)$: TD residual.

$$V^\pi(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \qquad Q^\pi(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \tag{2}$$

$$A^\pi(s_t, a_t) := Q^\pi(s_t, a_t) - V^\pi(s_t), \quad \text{(Advantage function)}. \tag{3}$$

  - 之后人们引入了 discount rate：γ，即：

$$V^{\pi,\gamma}(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right] \qquad Q^{\pi,\gamma}(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty} \gamma^l r_{t+l}\right] \tag{4}$$

$$A^{\pi,\gamma}(s_t, a_t) := Q^{\pi,\gamma}(s_t, a_t) - V^{\pi,\gamma}(s_t). \tag{5}$$

$$g^\gamma := \mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[\sum_{t=0}^{\infty} A^{\pi,\gamma}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right]. \tag{6}$$

需要注意引入 γ 之后，policy loss（g） 的估计就 有偏了

  - 使用一个统计量 估计 这里的 A，要求 替换后的 loss 求出的梯度 无偏，即所谓 γ-just （不是 g 本身无偏）

Definition 1. *The estimator $\hat{A}_t$ is γ-just if*

$$\mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[\hat{A}_t(s_{0:\infty}, a_{0:\infty}) \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] = \mathbb{E}_{\substack{s_{0:\infty}\\ a_{0:\infty}}}\left[A^{\pi,\gamma}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right]. \tag{7}$$

那满足 γ-just 的 A 的估计有

- $\sum_{l=0}^{\infty} \gamma^l r_{t+l}$
- $Q^{\pi,\gamma}(s_t, a_t)$
- $A^{\pi,\gamma}(s_t, a_t)$
- $r_t + \gamma V^{\pi,\gamma}(s_{t+1}) - V^{\pi,\gamma}(s_t)$.

  - 主要基于这样的观点：赋予距离当前更远的时刻更小的权值，可以减少扰动（如 TD0 比 MC 更稳定）（数学证明？）
  - 论文提出了 GAE 也是一种 A 的 符合 γ-just 的估计，是从1-step 到 n-step 的 TD Advantage function（δ）的加权平均：

$$\hat{A}^{GAE(\gamma,\lambda)} = \frac{(\hat{A}^{(1)} + \lambda\hat{A}^{(2)} + \lambda^2\hat{A}^{(3)} \dots)}{\frac{(1-\lambda)(1+\lambda+\lambda^2 \dots)}{(1-\lambda)}} = \frac{(\hat{A}^{(1)} + \lambda\hat{A}^{(2)} + \lambda^2\hat{A}^{(3)} \dots)}{\frac{(1-\lambda^n)}{(1-\lambda)}} \approx (1-\lambda)(\hat{A}^{(1)} + \lambda\hat{A}^{(2)} + \lambda^2\hat{A}^{(3)} \dots)$$

$$
\begin{aligned}
\hat{A}_t^{GAE(\gamma,\lambda)} &:= (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda\hat{A}_t^{(2)} + \lambda^2\hat{A}_t^{(3)} + \cdots\right) \\
&= (1-\lambda)\left(\delta_t^V + \lambda\left(\delta_t^V + \gamma\delta_{t+1}^V\right) + \lambda^2\left(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V\right) + \cdots\right) \\
&= \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}^V
\end{aligned}
$$

实际上是为更近的 A 赋予更大的权重