异步优势演员-评论家（A3C）：原理、公式、应用与实战

2025 年 9 月 21 日

1 引言

异步优势演员-评论家（Asynchronous Advantage Actor-Critic，A3C）通过多个并行工作线程采样、更新共享的策略与价值网络，避免经验回放带来的相关性问题。异步更新提升探索多样性，并充分利用多核 CPU 资源，是早期深度强化学习中的重要里程碑。

2 原理与公式

2.1 多工作线程目标

每个工作线程从全局参数 (θ, w) 拷贝本地副本 (θ', w')，采集长度为 n 的轨迹，并计算多步优势：

$$A_t = \sum_{k=0}^{n-1} \gamma^k r_{t+k+1} + \gamma^n V_w(s_{t+n}) - V_w(s_t). \tag{1}$$

本地累积梯度后再应用到全局参数。

2.2 异步更新

策略与价值的梯度分别为：

$$\nabla_\theta J \approx \sum_t \nabla_\theta \log \pi_\theta(a_t \mid s_t) A_t + \beta \nabla_\theta H[\pi_\theta(\cdot \mid s_t)], \tag{2}$$

$$\nabla_w L_V \approx \sum_t \partial_w \frac{1}{2} A_t^2, \tag{3}$$

其中 β 控制熵正则强度。计算完成后，工作线程将梯度异步地加到全局网络，再同步参数副本。

2.3 稳定性考量

异步执行会增加梯度噪声，需要通过较小的 rollout 长度、梯度裁剪、一致的学习率以及共享 RMSprop 统计量来维持稳定。适度的熵系数有助于避免线程过早收敛到同一策略。

3 应用与技巧

- **CPU 友好训练：**无需大型 replay buffer，即可充分利用多核服务器。
- **稀疏奖励：**多线程可并行探索不同轨迹，提高成功率。
- **离散/连续控制：**可扩展到高维 CNN 编码器或高斯策略。
- **实用建议：**定期同步参数，针对每个线程设置学习率退火，监控梯度范数以防发散。

4 Python 实战

脚本 `gen_a3c_figures.py` 模拟三个异步线程在“悬崖”网格世界中训练，使用多步优势估计更新共享策略与价值表，并输出汇总回报及策略概率热力图。

Listing 1: 脚本 $\mathrm{gen}_a 3c_f igures.py$

```
for worker in range(num_workers):
    states, actions, rewards = collect_rollout(theta_local[worker],
       V_local[worker])
    advantages = compute_n_step_advantages(states, rewards, V_global)
    grad_theta, grad_V = accumulate_gradients(states, actions,
       advantages)
    theta_global += actor_lr * grad_theta
    V_global += critic_lr * grad_V
```

5 实验结果

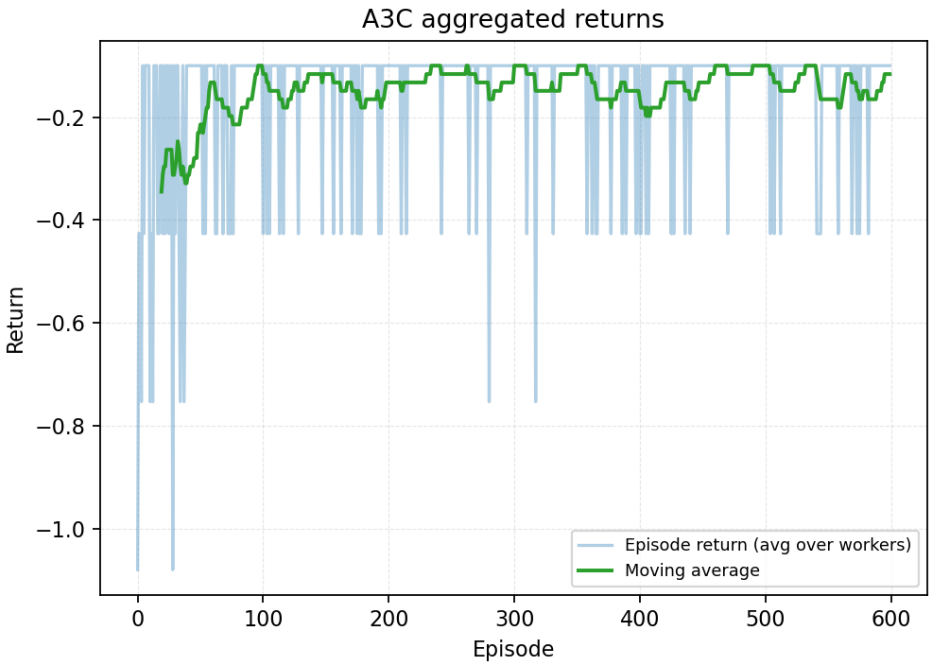

图 1: 多个线程异步更新后的回报曲线

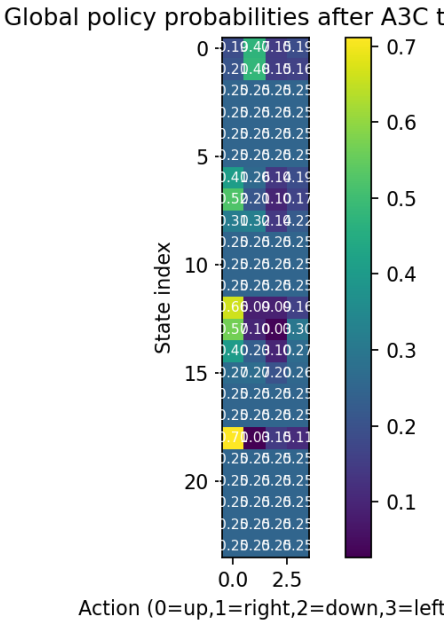

图 2: 最终策略的状态-动作概率热力图

6 总结

A3C 结合多步优势估计与异步线程，实现无需回放缓冲的稳定在线策略学习。合理选择 rollout 长度、熵系数与优化器可平衡梯度噪声与收敛速度。示例展示了多线程带来的收益以及策略概率如何集中在安全路径上。