作业 8：策略梯度算法

陈达贵　深蓝学院

1　文字作业

1. (1 分) 我们在课堂上给出了 A2C 的算法片段，请回答它是在策略的还是离策略的？为什么？

2. (2 分) 本节课我们推导了策略梯度的算法。设定策略网络 π_θ 的建模方式是以状态为输入，以动作的概率分布为输出 (有限的离散动作)

   (a) 深度学习框架一般都采用**最小化**损失函数的方式去自动优化参数。根据策略梯度算法的更新公式，请给出在实际使用时，actor 的损失函数。(注意策略梯度算法是最大化回报值)

   (b) 在基于值函数的方法时，我们会使用 ε 的贪婪策略去保持更多的探索。思考在使用策略网络的时候，如何保持 agent 充分的探索。

2　编程作业

1. (7 分) 本作业主要是学习使用 A2C 算法去玩游戏

   (a) 安装 gym，并安装 atari 环境，调用 Breakout（打砖块）的游戏 https://gym.openai.com/envs/#atari

(b) 使用 A2C 算法去玩该游戏,画出总得分随着训练片段数 (episodes) 的曲线。注意以下几点

- 如何保持 Actor 的探索?
- 更新时为了防止梯度太大导致训练不稳定，我们会限制 reward 的幅度。
- 为了保证执行的效率，同一个动作可能会执行多帧