强化学习 LAB2

中国科学技术大学
University of Science and Technology of China

段逸凡 dyf0202@mail.ustc.edu.cn

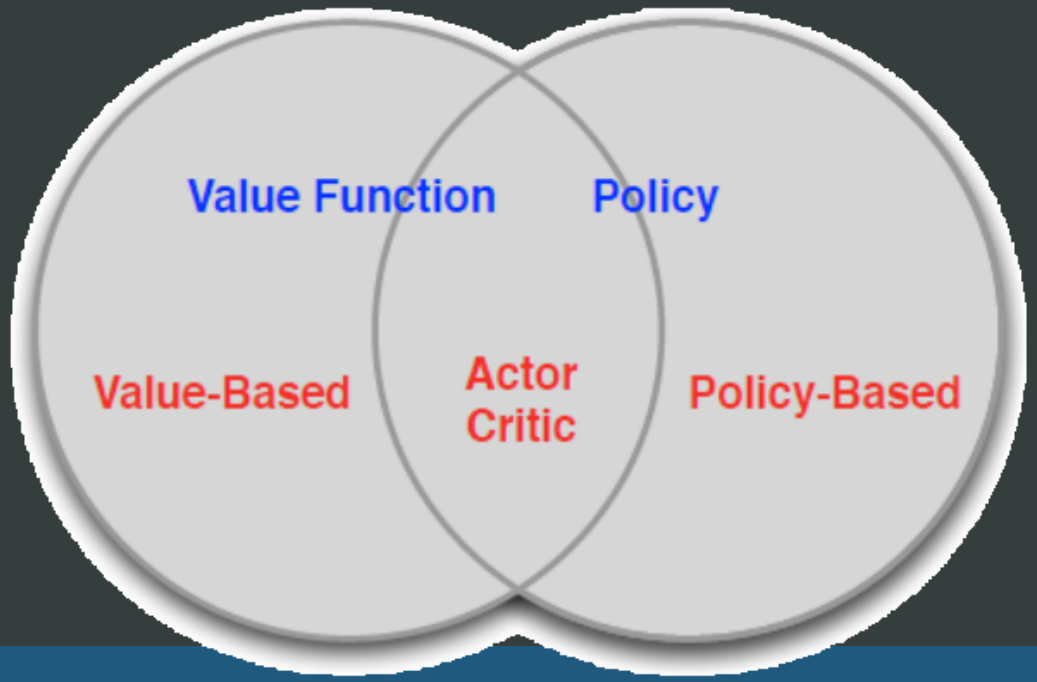
内容介绍

- 考察知识点
- 实验环境
- 实验任务
- 实验要求

考察知识点

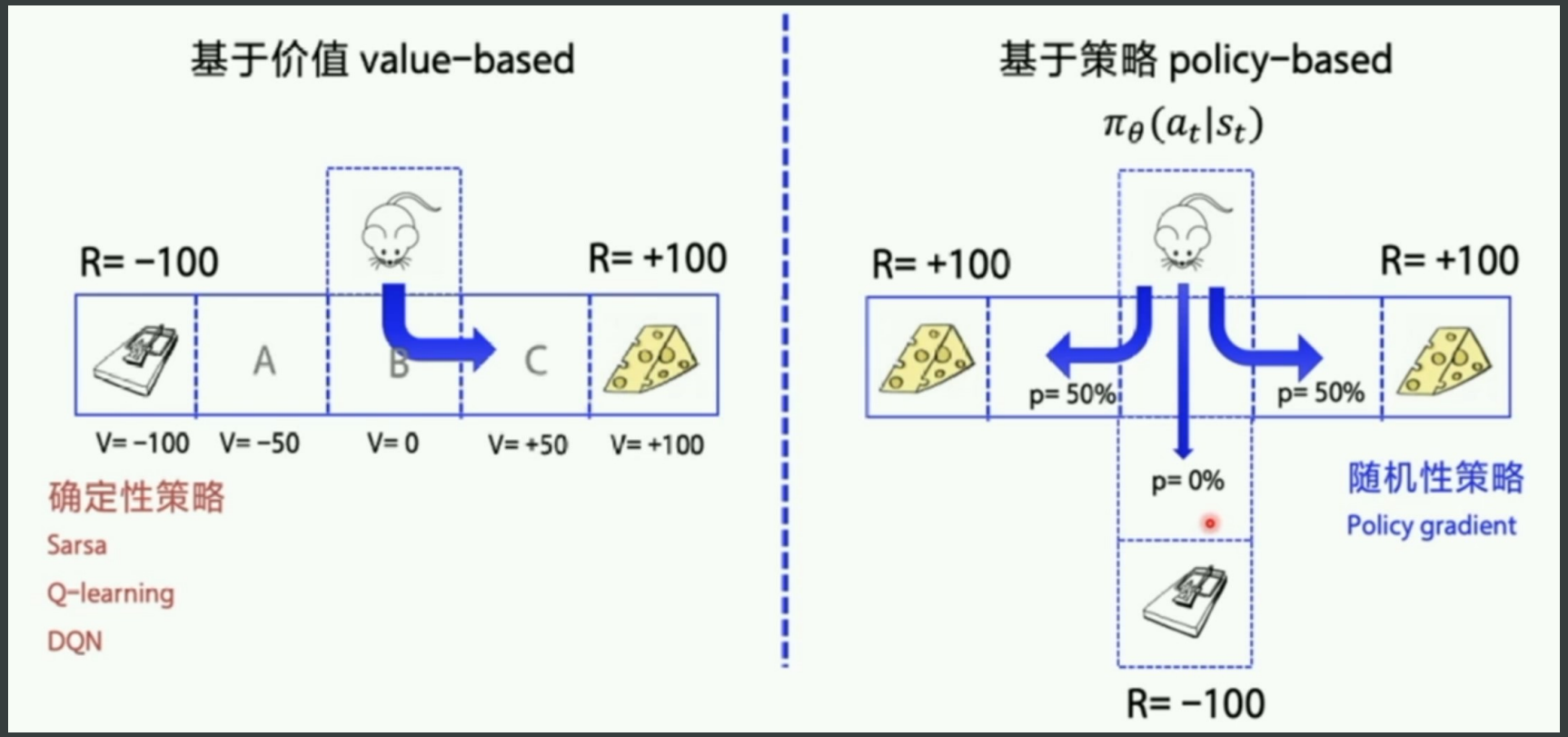
背景介绍

- 实验一考察了mc， sarsa， q-learning等 三种value-based方法
- 本次实验分为两个小实验，分别对Policy-Based以及Actor-Critic两种类型的方法进行考察

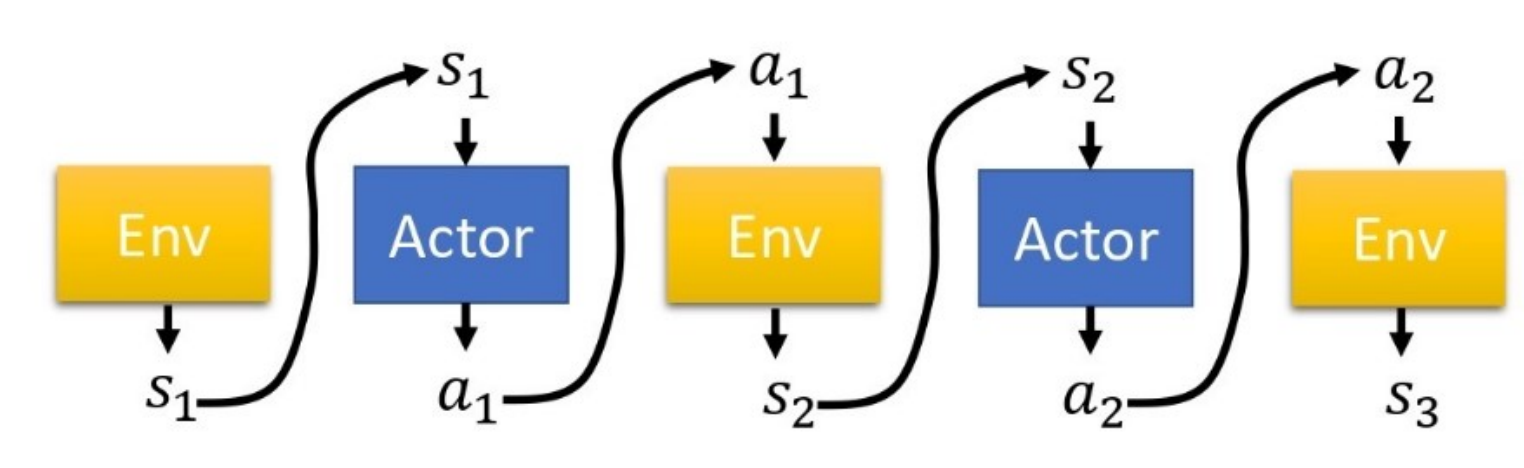


policy-based

- 详见ppt lec7，以及lec12



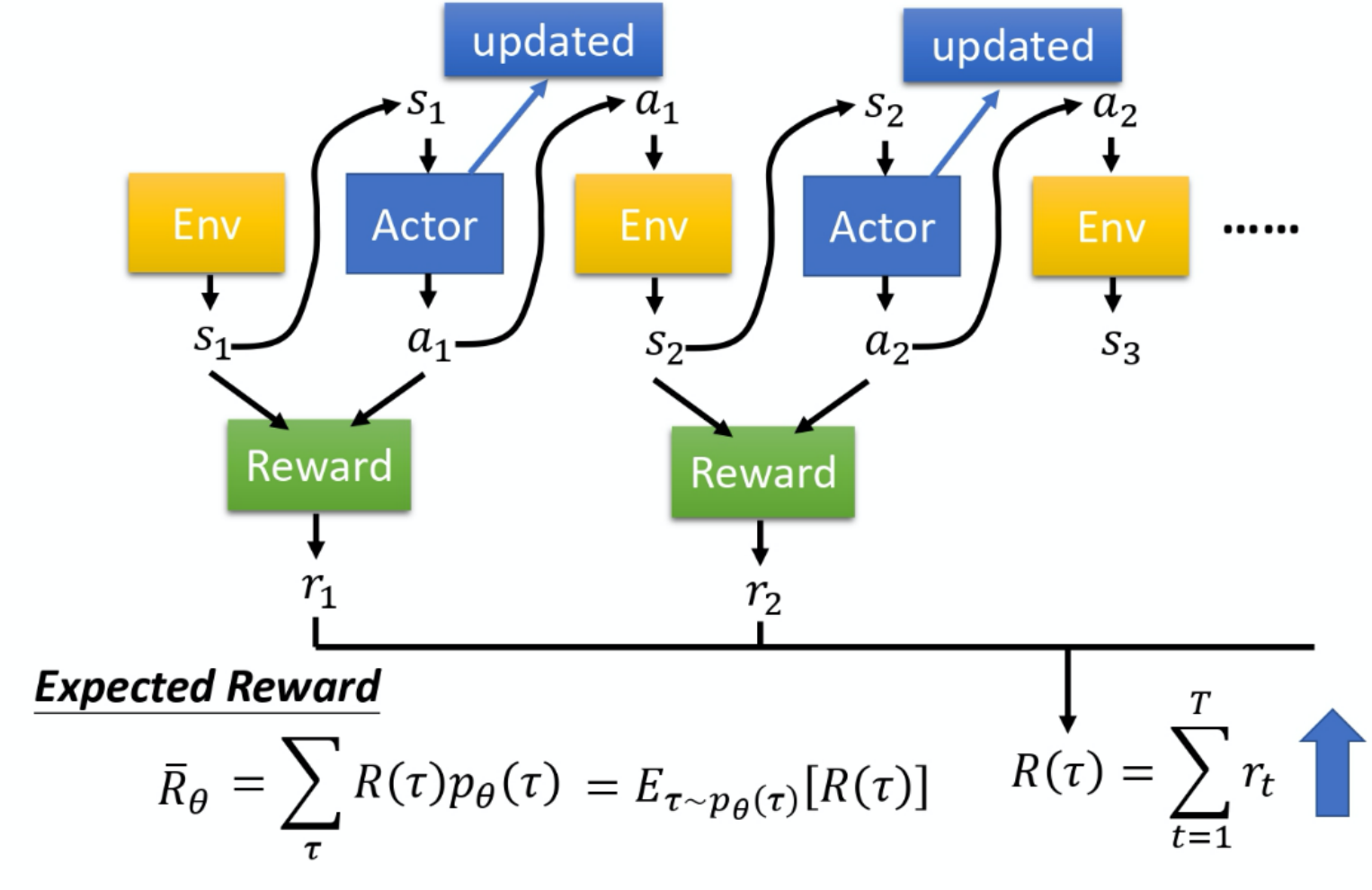
Policy Gradient



$$\text{Trajectory } \tau = \{s_1, a_1, s_2, a_2, \cdots, s_t, a_t\}$$

$$\begin{aligned} p_\theta(\tau) &= p(s_1)\, p_\theta(a_1|s_1)\, p(s_2|s_1, a_1)\, p_\theta(a_2|s_2)\, p(s_3|s_2, a_2) \cdots \\ &= p(s_1) \prod_{t=1}^{T} p_\theta(a_t|s_t)\, p(s_{t+1}|s_t, a_t) \end{aligned}$$

Policy Gradient



Policy Gradient

$$\nabla \bar{R}_\theta = E_{\tau \sim p_\theta(\tau)}[R(\tau)\nabla log p_\theta(\tau)]$$

Policy Gradient

Given policy π_θ

$$\begin{array}{lll} \tau^1: & (s_1^1, a_1^1) & R(\tau^1) \\ & (s_2^1, a_2^1) & R(\tau^1) \\ & \vdots & \vdots \\ \tau^2: & (s_1^2, a_1^2) & R(\tau^2) \\ & (s_2^2, a_2^2) & R(\tau^2) \\ & \vdots & \vdots \end{array}$$

only used once

Update Model

$$\theta \leftarrow \theta + \eta \nabla \bar{R}_\theta$$

$$\nabla \bar{R}_\theta = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} R(\tau^n) \nabla log p_\theta(a_t^n | s_t^n)$$

Data Collection

Policy Gradient

- 小tip

Tip 1: Add a Baseline

$$\theta \leftarrow \theta + \eta \nabla \bar{R}_\theta$$

It is possible that $R(\tau^n)$ is always positive.

$$\nabla \bar{R}_\theta \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} (R(\tau^n) \underline{- b}) \nabla log p_\theta(a_t^n | s_t^n) \qquad b \approx E[R(\tau)]$$

Policy Gradient

- 小tip

$$\nabla \bar{R}_\theta \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}(\cancel{R(\tau^n)} - b)\nabla log p_\theta(a_t^n|s_t^n)$$

$$\sum_{t'=t}^{T_n} r_{t'}^n \rightarrow \sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n$$

Add discount factor $\gamma < 1$

Actor-Critic

- 详见ppt lec7，以及lec12

baseline

$$\nabla \bar{R}_\theta \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\left(\underbrace{\sum_{t'=t}^{T_n}\gamma^{t'-t}r_{t'}^{n}}_{G_t^n} - \underline{b}\right)\nabla log p_\theta(a_t^n|s_t^n)$$

G_t^n : obtained via interaction

Very unstable

Actor-Critic

Actor-Critic

$$Q^{\pi_\theta}(s_t^n, a_t^n) - V^{\pi_\theta}(s_t^n)$$

$$V^{\pi_\theta}(s_t^n)$$

baseline

$$\nabla \bar{R}_\theta \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \boxed{\left(\sum_{t'=t}^{T_n} \gamma^{t'-t} r_{t'}^n - b \right)} \nabla log p_\theta(a_t^n | s_t^n)$$

G_t^n : obtained via interaction

$$E[G_t^n] = Q^{\pi_\theta}(s_t^n, a_t^n)$$

Advantage Actor-Critic(A2C)

$$Q^{\pi}(s_t^n, a_t^n) - V^{\pi}(s_t^n)$$

$$\Downarrow$$

$$r_t^n + V^{\pi}(s_{t+1}^n) - V^{\pi}(s_t^n)$$

$$Q^{\pi}(s_t^n, a_t^n) = E[r_t^n + V^{\pi}(s_{t+1}^n)]$$

$$Q^{\pi}(s_t^n, a_t^n) = r_t^n + V^{\pi}(s_{t+1}^n)$$

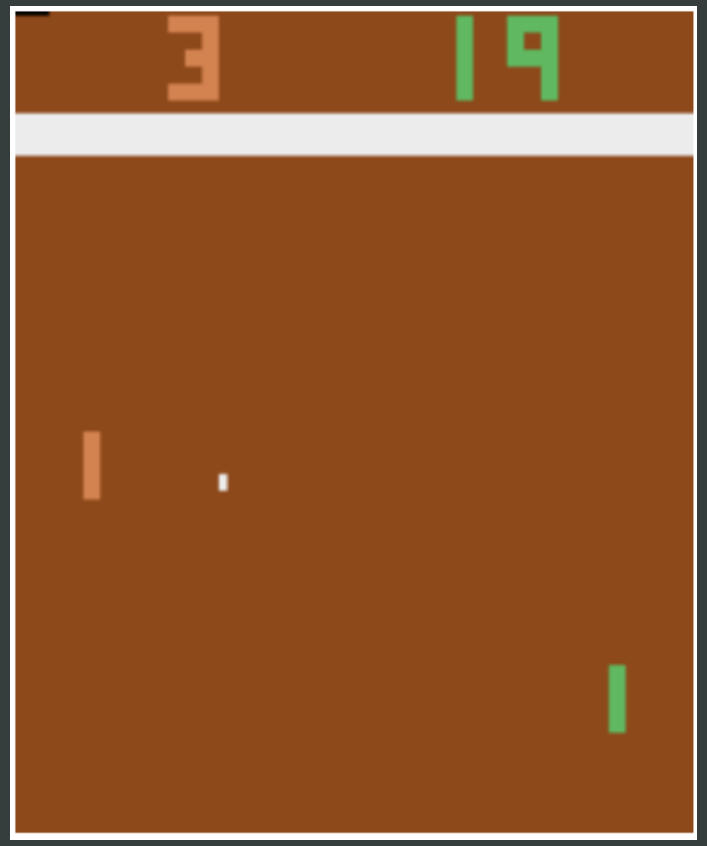
$$\nabla \bar{R}_\theta \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(r_t^n + V^{\pi}(s_{t+1}^n) - V^{\pi}(s_t^n)\right) \nabla log p_\theta(a_t^n | s_t^n)$$

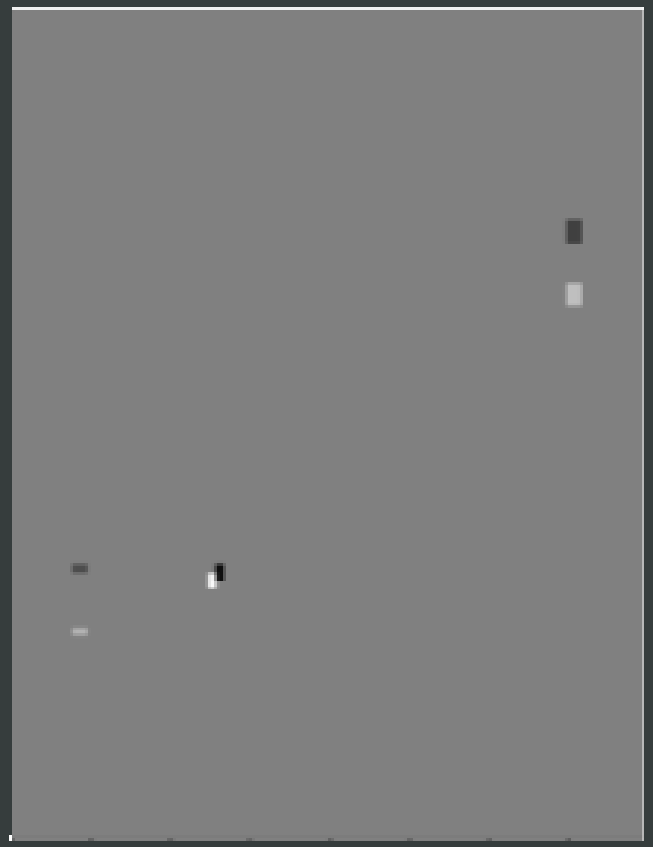
实验环境

Pong

- 实验环境：gym-Atari
  - PongDeterministic-v4
- state: 210x160x3 ->80*80
- action_space: 2
  - UP = 2
  - DOWN = 3
- 任意一方获得21分，游戏结束
- 得一分reward为1
- 输一分reward为-1

目标：-21->21



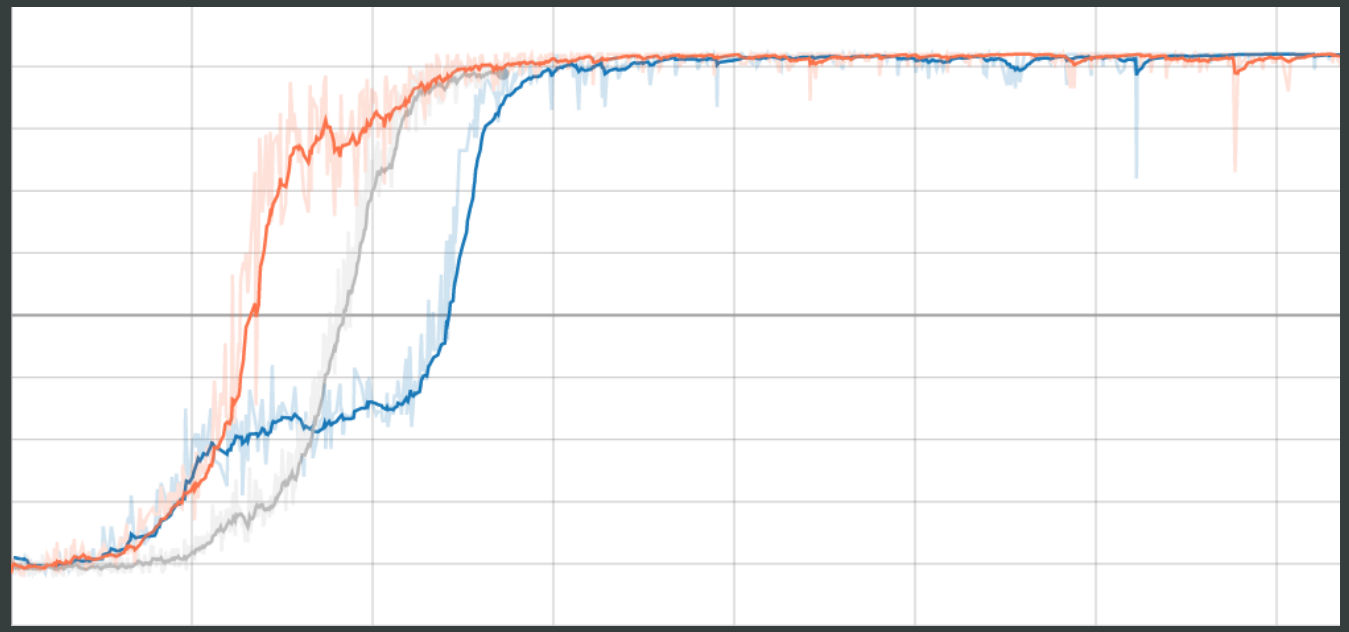


环境配置

- pip install gym[atari]
  - 如果出现 import gym 出现 AttributeError: module 'contextlib' has no attribute 'nullcontext'的错误，是因为gym版本和python版本不匹配，可以升级python版本或降gym版本
    - 例如：python 3.6 应该搭配 gym==0.15.7
- unzip ROMS.zip && cd Roms && python -m atari_py.import_roms .
- pytorch 安装教程：https://pytorch.org/get-started/locally/
- tensorflow 安装教程：https://www.tensorflow.org/install

tensorbord

- 运行 tensorboard --logdir tensorboard_logs， 后在浏览器中打开对应链接（如 http://localhost:6006/），可查看loss，以及reward曲线



注：PG在助教cpu上训练，约4h，600个batch后达到收敛

实验任务

实验任务

- 本实验提供一个Policy Gradient的代码框架
- 有五处TODO需要填空
  - 模型搭建
  - 前向传播
  - rewards计算
  - 动作选择
  - loss计算
- 实现PG后，在此基础上或者自己另外搭建，实现A2C算法

实验要求

注意事项

- 需要提交
  - 程序源文件（两个）
  - 实验报告一份（包括但不限于，核心代码注释，训练曲线，实现内容，两个算法优缺点对比）
- 实验DDL：2021.12.20 23：59：59
  - 延后一周*0.8
  - 期末*0.6
- 提交至ustcrl2021@163.com
- ！！！发送格式！！！：邮件以及附件命名均为 实验二-{姓名}-{学号}，例：实验二-张三-SA21000000

实验要求

- 评分细则：
  - Policy Gradient：40%
  - Actor-Critic：40%
  - report： 20%

创寰宇学府 育天下英才

感谢观看

中国科学技术大学
University of Science and Technology of China

段逸凡 dyf0202@mail.ustc.edu.cn