PPO × Family 第一讲习题

💡
- 提交格式：请将答案汇总至单个文件内，".pdf"，".docx" 均可。
- 提交方式：
  发送邮件至 opendilab@pjlab.org.cn

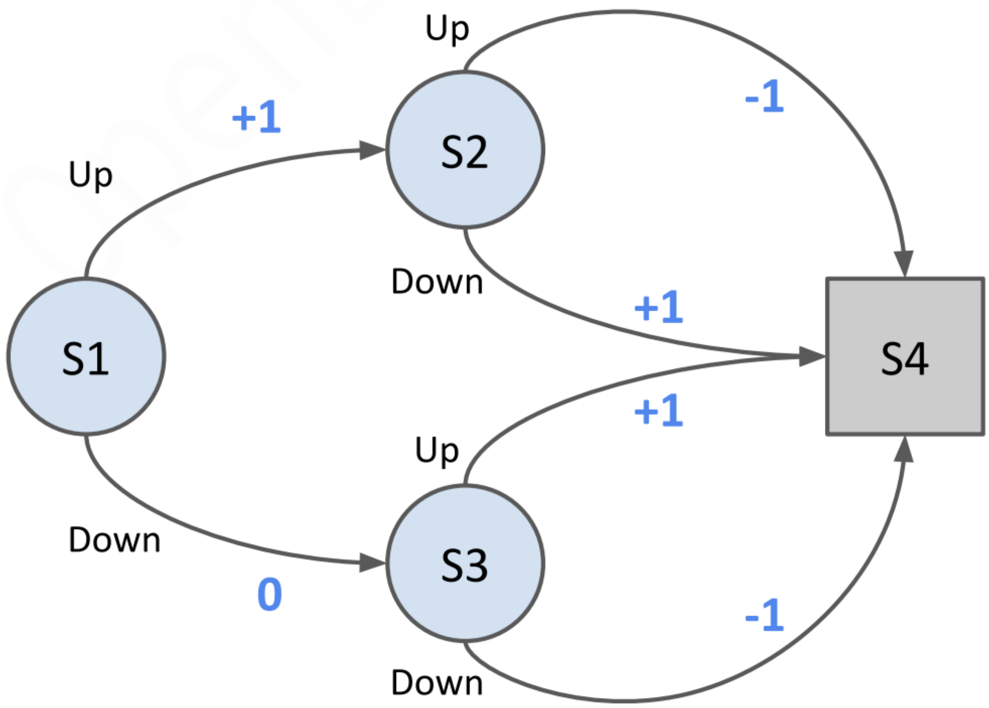
  请同学们严格按照下方格式命名邮箱主题/标题：

  【PPO × Family】+ 学生名 + vol.1 (第几节课）+ 作业提交日期

  示例：【PPO × Family】+ 喵小DI + vol.1 +20221207
- 提交截止时间为 2022.12.19 23:59 (GMT +8)，逾期作业将不会计入证书考量
- 如果其他问题请添加官方课程小助手微信（vx: OpenDILab)，备注「课程」，小助手将邀请您加入官方课程微信交流群；或发送邮件至 opendilab@pjlab.org.cn

题目1（MDP 求解）

如下图所示，是一个有限状态和长度的马尔科夫决策过程（MDP），$S1$ 是初始状态，$S4$ 是终止状态，对于每个状态，智能体可在动作集合 $A = \{\mathrm{Up},\ \mathrm{Down}\}$ 两种动作中选择一个执行，并获得相应的奖励。题目中使用折扣因子 $\gamma = 1$。另外，四个状态的表征信息完全相同，即 $\phi(s) = C$，其中 C 为某一常数。并且，由于表征信息相同，我们可以设 $\pi(up|\phi(s)) \;=\; p$



（四个状态的简单 MDP 示例）

1. 在**单步**状态转移的前提下，完成上述 MDP 的策略和奖励表

   （策略单步无法到达的状态用0表示即可，已给出 $S1$ 作为示例）

出发状态\到达状态	$S1$	$S2$	$S3$	$S4$
$S1$	0	$p, r=+1$	$1-p,\ r=0$	0
$S2$				
$S3$				
$S4$				

2. 尝试找到这个设定下**最优的随机性策略**，即确定 $\pi^\star(a|\phi(s))$ 。
   提示：可以表示出这个 MDP 下的状态价值函数，其中 r_t 是即时奖励：

$$V(s_t) = \sum_{a_t} \pi(a_t|\phi(s_t)) \left[\sum_{r_t} p(r_t|s_t, a_t) r_t \ + \ \gamma \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t) V(s_{t+1}) \right]$$

3. 在第二问得到的最优策略的基础上，计算动作价值函数 $Q_{\pi^\star}(\phi(s_t), up)$ 和 $Q_{\pi^\star}(\phi(s_t), down)$
   提示：执行 up 动作之后，能转移到的状态只有 $S2, S4$
   （有兴趣的同学可以以此来简单分析 Value-Based RL 方法和 Policy Gradient RL 方法的差异）

题目2 （Total Variation Distance 相关证明）

TRPO 的推导（补充材料）中有一个关键的不等式，给出了原函数和替代函数之间的定量关系：

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha$$
$$\text{where } \alpha = \max_s D_{\mathrm{KL}}(\pi(\cdot \mid s) \| \tilde{\pi}(\cdot \mid s)), \epsilon = \max_{s,a} |A_\pi(s,a)|$$

这个不等式的证明过程中，用到了一个重要的数学工具 total variation distance 来刻画两个概率分布之间的距离（WIKI链接），即对于两个定义在相同事件集合 $\mathcal{X}$ 上的概率分布 P, Q，他们的 total variance distance 为：

$$\delta_{TV}(P, Q) = \sup_{A \subseteq \mathcal{X}} |P(A) \ - \ Q(A)|, \ P(A) = \sum_{x \in A} P(x)$$

其中 A 是事件集合 $\mathcal{X}$ 的子集，不是 $\mathcal{X}$ 里的一个事件， $\sup$ 代表上确界。然而在一般实践中，又常常使用另一个形式（仅考虑离散事件集合）：

$$\delta_{TV}(P, Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$$

试证明两者的等价性