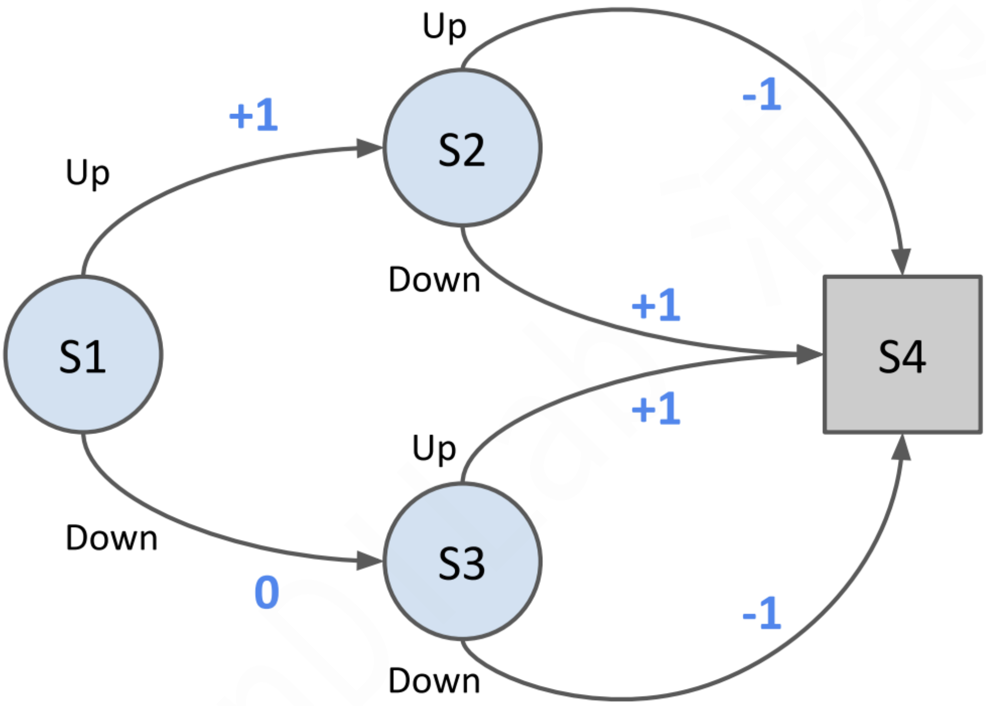
第一节课习题作业

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
$S2$	0	0	0	$1, r=1-2p$
$S3$	0	0	0	$1,\ r=2p-1$
$S4$	0	0	0	0

(注：题解中将 $S4$ 视为了特殊的终止状态，也可将 $S4$ 视为吸收态，则 $S4 \ \rightarrow \ S4$ 为 1)

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：

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

题解：我们可以使用题干中提示的公式获得各个时刻的状态价值函数：

$$V(s_4)=0$$

$$\begin{aligned}V(s_3)&=\pi(up|C)[p(r_t=+1|s_3,up)r_t+\gamma p(s_4|s_3,up)V(s_4)]+\ \pi(down|C)[p(r_t=-1|s_3,down)r_t+\gamma p(s_4|s_3,down)V(s_4)]\\&=p\times[1\times1+1\times\ 1\times\ 0]\ +\ (1-p)\times[1\times-1+1\times\ 1\times\ 0]\\&=2p-1\end{aligned}$$

$$\begin{aligned}V(s_2)&=\pi(up|C)[p(r_t=-1|s_2,up)r_t+\gamma p(s_2|s_4,up)V(s_4)]+\ \pi(down|C)[p(r_t=+1|s_2,down)r_t+\gamma p(s_4|s_2,down)V(s_4)]\\&=p\times[1\times-1+1\times\ 1\times\ 0]\ +\ (1-p)\times[1\times1+1\times\ 1\times\ 0]\\&=1-2p\end{aligned}$$

$$\begin{aligned}V(s_1)&=\pi(up|C)[p(r_t=+1|s_1,up)r_t+\gamma p(s_2|s_1,up)V(s_2)]+\ \pi(down|C)[p(r_t=0|s_1,down)r_t+\gamma p(s_3|s_1,down)V(s_3)]\\&=p\times[1\times1+1\times\ 1\times\ (1-2p)]\ +\ (1-p)\times[1\times0+1\times\ 1\times\ (2p-1)]\\&=-4p^2+5p-1\end{aligned}$$

为了求极值，我们计算 $V(s_1)$ 对概率 p 的导数，即：

$$\frac{\mathrm{d}V(s_1)}{\mathrm{d}p}=-8p+5$$

并取导数的零点，从而可得：当 $p=\dfrac{5}{8}$ 时，$V(s_1)$ 取到极大值，因此根据最优策略的定义，有：

$$\pi^\star(a|\phi(s))=\underset{\pi(a|\phi(s))}{\arg\max}\ V(s_1)=\begin{cases}\frac{5}{8},\ a=up\\\frac{3}{8},\ a=down\end{cases}|\ \phi(s)$$

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）

题解：动作价值函数的公式为：

$$
\begin{aligned}
Q_\pi(\phi(s_t), a_t) &= \mathbb{E}_{\phi(s_t), s_t, a_t \sim \pi}[\sum_{r_t} p(r_t|s_t, a_t) r_t + \gamma \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t) V(s_{t+1})] \\
&= \sum_{s_t} p(s_t|\phi(s_t))[\sum_{r_t} p(r_t|s_t, a_t) r_t + \gamma \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t) V(s_{t+1})]
\end{aligned}
$$

由于我们只能观测到状态的表征信息 $\phi(s_t)$ ，所以上式中引入了 $p(s_t|\phi(s_t))$ ，即在观测到 $\phi(s_t)$ 时，实际为为具体状态（ s_1， s_2， s_3 ）的概率

根据题干中的 MDP 图示，可以推导出各个状态 s_1， s_2， s_3 在策略 $\pi^\star(a|\phi(s))$ 下的分布规律（ s_4 为终态，故不存在观测 $\phi(s_4)$ 的事件），具体过程为：

- $p(\phi(s_t)) = 1$
- $p(s_i|\phi(s_t)) = \dfrac{p(s_i, \phi(s_t))}{p(\phi(s_t))} = \dfrac{p(s_i)}{p(\phi(s_t))} = p(s_i)$

考虑到三个状态之间的转移关系，有：

- $p(s_1|\phi(s_t)) + p(s_2|\phi(s_t)) + p(s_3|\phi(s_t)) = 1$
- $p(s_1|\phi(s_t)) = p(s_2|\phi(s_t)) + p(s_3|\phi(s_t))$
- $\dfrac{p(s_2|\phi(s_t))}{p(s_3|\phi(s_t))} = \dfrac{p(s_2|s_1, up)}{p(s_3|s_1, down)} = \dfrac{p}{1-p}$

因此，联立上述式子可以算出具体的概率值，即：

$$
\begin{aligned}
p(s_1|\phi(s_t)) &= \frac{1}{2} \\
p(s_2|\phi(s_t)) &= \frac{p}{2} \\
p(s_3|\phi(s_t)) &= \frac{1-p}{2}
\end{aligned}
$$

接下来，我们计算观测信息 $\phi(s_t)$ 所对应的动作价值函数，它是关于各个时刻的状态 s_t 的动作价值函数的期望：

$$
\begin{aligned}
Q_{\pi^\star}(\phi(s_t), up) &= \frac{1}{2} \times [1 \times 1 + 1 \times 1 \times (1-2p)] \ + \frac{p}{2} \times \ [1 \times -1 + 1 \times 1 \times \ 0] \ + \frac{1-p}{2} \times \ [1 \times 1 + 1 \times \ 1 \times 0] \\
&= \frac{3}{2} - 2p \ \text{(代入(2)中的最优值)} \\
&= \frac{1}{4}
\end{aligned}
$$

$$
\begin{aligned}
Q_{\pi^\star}(\phi(s_t), down) &= \frac{1}{2} \times [1 \times 0 + 1 \times 1 \times (2p-1)] \ + \frac{p}{2} \times \ [1 \times 1 + 1 \times 1 \times \ 0] \ + \frac{1-p}{2} \times \ [1 \times -1 + 1 \times \ 1 \times 0] \\
&= 2p - 1 \\
&= \frac{1}{4}
\end{aligned}
$$

根据上述计算结果我们可以看到，由于各个状态的表征信息是一致的，所以最终得到的最优动作价值函数，对于**不同的动作竟然是相等的，无法决策判别出究竟该选择哪一个动作。**这个例子也可以进一步泛化到更通用的场景中：环境观察信息处理的不妥当，神经网络优化的不够好，都可能造成类似这样的情形，使得 Value-Based RL 强化学习方法可能无法找到最优解。

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

题解：

将定义三类事件集合 $\mathcal{X}^{P<Q}$，$\mathcal{X}^{P=Q}$，$\mathcal{X}^{P>Q}$，将各个随机事件 $x \in \mathcal{X}$，分别按照概率质量函数在概率分布 P, Q 之间的相对大小，归纳到这三个互斥的事件集合中：

$$\mathcal{X}^{P<Q} = \{x | P(x) < Q(x)\}$$
$$\mathcal{X}^{P=Q} = \{x | P(x) = Q(x)\}$$
$$\mathcal{X}^{P>Q} = \{x | P(x) > Q(x)\}$$
$$\mathcal{X}^{\mathcal{P}>\mathcal{Q}} \cup \mathcal{X}^{\mathcal{P}=\mathcal{Q}} \cup \mathcal{X}^{\mathcal{P}<\mathcal{Q}} = \mathcal{X}$$
$$\mathcal{X}^{P>Q} \cap \mathcal{X}^{P=Q} = \mathcal{X}^{P>Q} \cap \mathcal{X}^{P<Q} = \mathcal{X}^{P=Q} \cap \mathcal{X}^{P<Q} = \emptyset$$

首先有：

$$\sum_{x \in \mathcal{X}} P(x) = \{\sum_{x \in \mathcal{X}^{P<Q}} + \sum_{x \in \mathcal{X}^{P=Q}} + \sum_{x \in \mathcal{X}^{P>Q}}\} P(x) = 1$$

$$\sum_{x\in\mathcal{X}} Q(x) = \{\sum_{x\in\mathcal{X}^{P<Q}} + \sum_{x\in\mathcal{X}^{P=Q}} + \sum_{x\in\mathcal{X}^{P>Q}}\}Q(x) = 1$$

将上两式做差，有：

$$\{\sum_{x\in\mathcal{X}^{P<Q}} + \sum_{x\in\mathcal{X}^{P>Q}}\}P(x) - \{\sum_{x\in\mathcal{X}^{P<Q}} + \sum_{x\in\mathcal{X}^{P>Q}}\}Q(x) = 0$$

即：

$$\sum_{x\in\mathcal{X}^{P>Q}}\{P(x) - Q(x)\} = \sum_{x\in\mathcal{X}^{P<Q}}\{Q(x) - P(x)\} = C$$

根据定义有：

$$\delta_{TV}(P,Q) = \sup_{A\subseteq\mathcal{X}} |P(A)\ -\ Q(A)| = P(\mathcal{X}^{P>Q}) = C$$

而：

$$\begin{aligned}\sum_{x\in\mathcal{X}} |P(x) - Q(x)| &= \sum_{x\in\mathcal{X}^{P>Q}}\{P(x) - Q(x)\} + \sum_{x\in\mathcal{X}^{P=Q}}\{P(x) - Q(x)\} + \sum_{x\in\mathcal{X}^{P<Q}}\{Q(x) - P(x)\} \\ &= C + 0 + C \\ &= 2C\end{aligned}$$

所以最终得到：

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