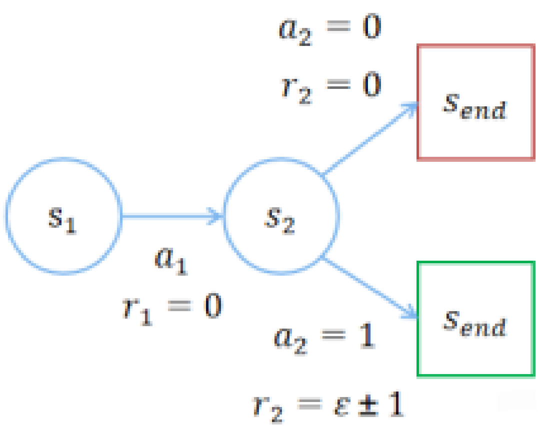
PPO × Family 第四讲习题

—— 02

理论题



1、请计算理论上的最优动作价值函数的数值 $Q(s_1, a_1)$， $Q(s_2, a_2 = 0)$， $Q(s_2, a_2 = 1)$ 与最优动作价值函数的分布 $Z(s_1, a_1)$， $Z(s_2, a_2 = 0)$， $Z(s_2, a_2 = 1)$，为了简化计算，令 $\gamma = 1$

（示例：比如 $Q(s_2, a_2 = 1) = \mathbb{E}(r(s_2, a_2)) + 0 = 2\epsilon$， $Z(s_2, a_2 = 1) = r(s_2, a_2) + 0 = \epsilon \pm 1$）

作答

1. 思路：

$$Z(s_t, a_t) = \sum_{i=0}^{end} \gamma^i r(s_{t+i}, a_{t+i})$$

$$\mathcal{T} Z(s, a) = r(s, a) + \gamma Z(s', a')$$

$$Q^*(s, a) = r(s, a) + \gamma \max_{a' \in \mathcal{A}} Q^*(s', a')$$

$$\mathcal{T} Q(s, a) = \mathbb{E}[r(s, a)] + \gamma \mathbb{E}_\pi [\max_{a' \in \mathcal{A}} Q(s', a')]$$

2. 答案：

$$
\begin{aligned}
Q(s_2, a_2 = 1) &= \mathbb{E}[r(s_2, a_2 = 1)] + 0 = 2\varepsilon \\
Q(s_2, a_2 = 0) &= \mathbb{E}[r(s_2, a_2 = 0)] + 0 = 0 \\
Q(s_1, a_1) &= \mathbb{E}[r(s_1, a_1 = 0)] + 2\varepsilon \\
&= 2\varepsilon
\end{aligned}
$$

$$
\begin{aligned}
Z(s_2, a_2 = 1) &= r(s_2, a_2 = 1) + 0 = 1 \pm \varepsilon \\
Z(s_2, a_2 = 0) &= r(s_2, a_2 = 0) + 0 = 0 \\
Z(s_1, a_1) &= r(s_1, a_1 = 0) \\
&\quad + \pi(a' = 0|s' = s_2) * Z(s_2, a_2 = 1) \\
&\quad + \pi(a' = 1|s' = s_2) * Z(s_2, a_2 = 0) \\
&= \frac{1 \pm \varepsilon}{2}(\text{假设两种动作等可能发生})
\end{aligned}
$$

2、假如当前策略的动作价值函数分布为 $Z(s_1, a_1) = \epsilon \pm 1$， $Z(s_2, a_2 = 0) = 0$， $Z(s_2, a_2 = 1) = -\epsilon \pm 1$。

i) 请计算使用贝尔曼最优算子后，目标动作价值函数的分布，$\mathcal{T}Z(s_1, a_1)$， $\mathcal{T}Z(s_2, a_2 = 0)$， $\mathcal{T}Z(s_2, a_2 = 1)$。

（示例：比如$\mathcal{T}Z(s_2, a_2 = 1) = r(s_2, a_2 = 1) + 0 = \epsilon \pm 1$）

作答

1. 思路：

$$
\mathcal{T}Z(s, a) = r(s, a) + \gamma Z(s', a')
$$

2. 答案：

$$
\begin{aligned}
\tau Z(s_2, a_2 = 1) &= r(s_2, a_2 = 1) + 0 = 1 \pm \varepsilon \\
\tau Z(s_2, a_2 = 0) &= r(s_2, a_2 = 0) + 0 = 0 \\
\tau Z(s_1, a_1) &= r(s_1, a_1 = 0) \\
&\quad + \pi(a' = 0|s' = s_2) * Z(s_2, a_2 = 1) \\
&\quad + \pi(a' = 1|s' = s_2) * Z(s_2, a_2 = 0) \\
&= \frac{\pm 1 - \varepsilon}{2}(\text{假设两种动作等可能发生})
\end{aligned}
$$

---

ii) 衡量两个概率分布之间的距离的测度方法有很多，其中之一为Wasserstein Metric [1]。假如标记 p 阶Wasserstein Metric 为 W_p ，其表达式如下：

$$W_p(\mu, \nu) = (\inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma} d(x,y)^p)^{\frac{1}{p}}$$

请计算使用贝尔曼最优算子前后，当前策略的动作价值函数与最优策略的动作价值函数的 1 阶 Wasserstein Metric。比较并讨论两者的差异，然后尝试分析这种差异的影响。

作答

1. 思路：Wasserstein Metric（WM）是用来刻画两个分布的最短距离，考虑到了两个分布的联合概率分布。

   先写出一阶表达式，然后可以看到是求积空间的期望最小

$$W_1(\mu, v) = (\inf \mathbb{E}_{(x,y)\ \gamma} d(x,y))$$

   where $\Gamma(\mu, \nu)$ is the set of all couplings of μ and ν. A coupling γ is a joint probability measure on $M \times M$ whose marginals are μ and ν on the first and second factors, respectively. That is,

$$\int_M \gamma(x,y)\,\mathrm{d}y = \mu(x)$$
$$\int_M \gamma(x,y)\,\mathrm{d}x = \nu(y)$$

2. 答案：

$$W_1(\tau Z(s_2, a_2 = 1), Z(s_2, a_2 = 1)) = 2\varepsilon$$
$$W_1(\tau Z(s_2, a_2 = 0), Z(s_2, a_2 = 0)) = 0$$
$$W_1(\tau Z(s_1, a_1), Z(s_1, a_1)) = \frac{3\varepsilon}{2}$$

   从 WM 测度可以看到策略优化的分布变化，这种变化主要是来源于在更新的时候考虑到了真实的 reward 信息，从而使得动作价值函数更符合实际情况。

请参考下文中的提示，证明机械狗笨笨需要采用的UCB算法的数学形式为：

$$a_t = \arg\max_a [Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}}]$$

其中 t 为第 t 次安装测试。 $N_t(a)$ 为累计至第 $t-1$ 次安装时，选择 a 厂商的次数。 $Q_t(a)$ 为累计至第 $t-1$ 次安装时的平均动作价值。 c 为平衡探索效益和利用效益之间的系数。

作答

因为每次如果假设回报都是 1 或 -1 ，那么就具有一个乐观上界 $\tilde{Q}$ 和悲观下界 $\hat{Q}$ ，两者和真实的估计 Q 存在一个 $\Delta = |\tilde{Q} - Q|$ ，由于我们的每次决策服从伯努利分布，总体呈二项分布，且其中

$$Q(a) = \frac{1}{N}\sum_{i=1}^{N} r_i(s_i, a)$$

是奖励函数的期望

为此我们可以引入 Chernoff-Hoeffding Bound， $p(|\tilde{Q} - Q| \le \sigma) \ge 1 - 2e^{-2n\sigma^2}$

当设定 $\sigma = \sqrt{2\frac{\log t}{N_t(a)}}$ 时， $p(|\tilde{Q} - Q| \le \sigma) \ge 1 - \frac{2}{t^4}$

再当 $T = 4$ 时， $p = 0.992$ ，即该上界随着策略执行次数的增加能够基本包括最大回报的情况，故为最优上界，当将 c 设为不同的值时，有不同的逼近效果，故可证明上式

---

ii)

霍夫丁不等式 (Hoeffding's inequality)：

如果 Z 为一系列独立同分布且有界的随机数， $Z_i \in [a, b]$， $-\infty < a \le b < +\infty$，那么对于所有的非负实数， $\delta \ge 0$，都有如下不等式成立：

$$P(\frac{1}{n}[\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])] \ge \delta]) \le \exp(-\frac{2n\delta^2}{(b-a)^2})$$

$$P(\frac{1}{n}[\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])] \le -\delta]) \le \exp(-\frac{2n\delta^2}{(b-a)^2})$$

作答

引入马尔可夫不等式：

$$P(X \ge \varepsilon) \le \frac{E(X)}{\varepsilon}$$

则

$$P(\frac{1}{n}[\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])]\geq\delta])\leq\frac{E[\frac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])]}{\delta}$$

又因为

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i]) &= \sum_i\frac{Z_i}{n}-\sum_i\mathbb{E}[\frac{Z_i}{n}] \\ &= \sum_i\frac{Z_i}{n}-\mathbb{E}[\sum_i\frac{Z_i}{n}] \\ &= Z-\mathbb{E}(Z)\end{aligned}$$

则

$$P(\frac{1}{n}[\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])]=P((Z-\mathbb{E}(Z))\geq\delta])\leq\frac{E[Z-\mathbb{E}(Z)]}{\delta}$$

等价变形为

$$\begin{aligned}&P((Z-\mathbb{E}(Z))\geq\delta]) \\ &= P(exp(s[Z-\mathbb{E}[Z]])\geq exp(s\delta))\leq exp(-s\delta)E(exp(s[Z-\mathbb{E}[Z]])\end{aligned}$$

这里需要用到霍夫丁不等式引理，不加证明地引入到这里：

对于随机变量，如果满足$P(X\in[a,b])=1$和$E(X)=0$，有

$$E(exp(sX))\leq exp(\frac{s^2(b-a)^2}{8})$$

由于 $E(\frac{Z_i-\mathbb{E}[Z_i]}{n})=0$， 且 $\frac{Z_i-\mathbb{E}[Z_i]}{n}\in(\frac{a-\mathbb{E}[Z_i]}{n},\frac{b-\mathbb{E}[Z_i]}{n})$ 有界，则有

$$E(exp(s\frac{Z_i-\mathbb{E}[Z_i]}{n}))\leq exp(\frac{s^2(b-a)^2}{8n^2})$$

又因

$$
\begin{aligned}
E(exp(s[Z-\mathbb{E}[Z]]) &= E(exp(s(\sum_i \frac{Z_i}{n} - \mathbb{E}[\sum_i \frac{Z_i}{n}]))) \\
&= E(exp(s(\sum_i \frac{Z_i}{n} - \sum_i \mathbb{E}[\frac{Z_i}{n}]))) \\
&= \prod_i E(exp(s(\frac{Z_i-\mathbb{E}[Z_i]}{n})))
\end{aligned}
$$

则

$$
E(exp(s[Z-\mathbb{E}[Z]]) \leq exp(\frac{s^2(b-a)^2}{8n})
$$

则

$$
P((Z-\mathbb{E}(Z)) \geq \delta]) \leq exp(-s\delta)exp(\frac{s^2(b-a)^2}{8n})
$$

求右边最小项地过程略，最终可得

$$
P(\frac{1}{n}[\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])] = P((Z-\mathbb{E}(Z)) \geq \delta]) \leq exp(-\frac{2n\delta^2}{(b-a)^2})
$$

代码实践

> 已经在 20230312 提交了pr
>
> https://github.com/opendilab/PPOxFamily/pull/50

在本题目中，我们将尝试一个简化后的 RND 奖励模型训练实践任务。具体来说，我们收集了Minigrid [5] 环境（具体为“MiniGrid-Empty-8x8-v0”）强化学习训练过程中产生的部分数据，现在请你为ran法**尝试使用不同超参数设置**的 RND 预测网络，用于回归一个固定大小的目标网络的输出，并将预测网络和目标网络的输出的预测差值作为内在探索奖励的大小，观察 Tensorboard 中相关的数据记录结果，分析探索奖励的数值在预测网络各个训练阶段的变化，评估其奖励数值的大小和范围与模型本身的**欠拟合或过拟合**等因素之间的关联。

题目2（应用实践）

在课程第四讲（解密稀疏奖励空间）几个应用中任选一个

- minigrid 迷宫（奖励的稀疏性）
- metadrive 自动驾驶（奖励的多尺度变化）

根据课程组给出的示例代码，训练得到相应的智能体。最终提交需要上传相关训练代码、日志截图或最终所得的智能体效果视频（replay），具体样式可以参考第四讲的示例 ISSUE。

- 代码实践题提交方式：

  PPO × Family 官方GitHub 上发起 Pull Request

  - 地址：**PPOxFamily**/tree/main/**chapter4_reward**/hw_submission
  - PR示例：https://github.com/opendilab/PPOxFamily/pull/5
  - 命名规范：

    hw_submission(学生名称): add hw4（第几节课）+作业提交日期

    示例：hw_submission(nyz): add hw3_20230104

提交**截止时间为 2023.03.16 23:59 (GMT +8)**，逾期作业将不会计入证书考量。

Reference

机器学习——霍夫丁（Hoeffding）不等式证明 – 颀周 – 博客园

知 霍夫丁（Hoeffding）不等式证明

https://blog.csdn.net/qqiseeu/article/details/46293457

高级算法 | Chernoff bound

https://en.wikipedia.org/wiki/Wasserstein_metric

知 Wasserstein距离