Maximization Bias Example

刘经宇

2023 年 4 月 24 日

这次作业是关于书中的例 6.7: **Maximization Bias Example** 的, 其 MDP 模型和相关的数值结果如图 1 所示.

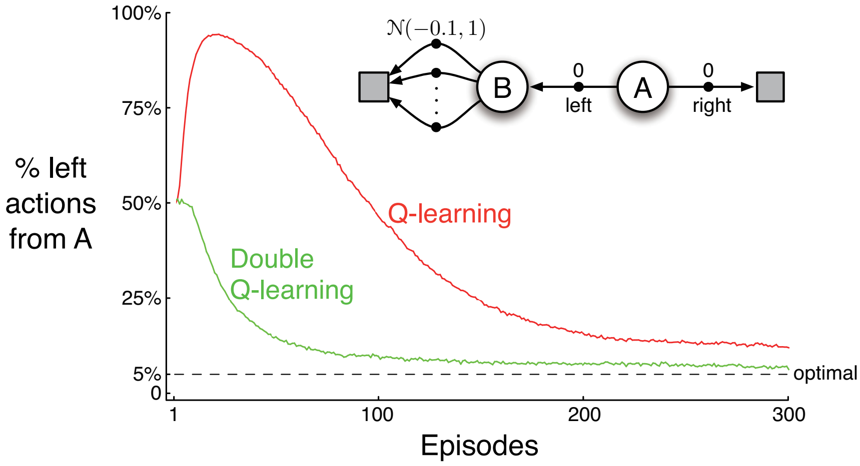

图 1: Maximization bias.

为了复现这一结果和做进一步的探究, 我们假设当智能体到达状态 B 后, 通过玩一个多臂老虎机 (multi-armed bandit) 到达终止状态, 我们假设这个多臂老虎机的臂的数量为 m, 每个臂 a_i 的收益 ξ_i 是独立同分布的随机变量, 它们都服从正态分布 $\xi_i \sim N(\mu, \sigma^2)$.

为了方便, 我们在这里也介绍这篇报告中的其他记号. 状态集 $\mathcal{S} = \{A, B, C\}$, 其中 C 是终止状态. 动作集 $\mathcal{A}(A) = \{\text{left}, \text{right}\}$ (在状态 A 可以选择左或者右), $\mathcal{A}(B) = \{a_1, \ldots, a_m\}$ (在状态 B 可以选择臂 $\{a_i\}_{i=1}^m$). ε-greedy 策略的参数为 $\varepsilon \in [0,1]$, 学习步长为 $\alpha \in (0,1]$, 关于未来收益的 discount 为 $\gamma \in [0,1]$. 在书中的例子里, 部分参数的设置为 $\mu = -0.1$, $\sigma^2 = 1$, $\varepsilon = 0.1$, $\alpha = 0.1$, $\gamma = 1$.

在复现中, 我们取 $m = 10$, 运行 10000 次程序后取平均 (之后的结果也都是在运行 10000 次后取平均得到的). 其结果如图 2 所示. 在最后一个 episode 中, Q-learning 和 doule Q-learning 向左的动作比例分别为 12.17% 和 6.53%.

我们先解释为什么在上述参数设定下, 最优策略中向左的动作比例仍存在 5%. 考虑最优策略下

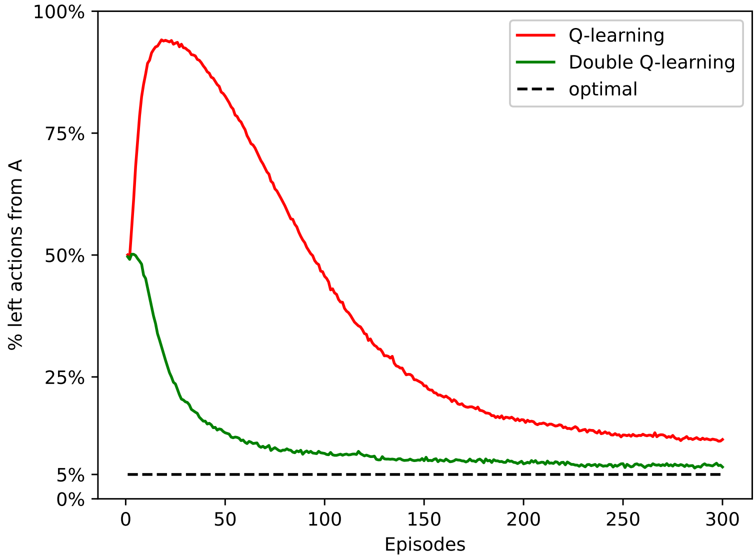

图 2: 复现的 Maximization bias.

关于 q_* 的 Bellman 方程

$$
\begin{aligned}
q_*(A, \mathrm{left}) &= \mathbb{E}[R_{t+1} + \gamma \max_a q_*(S_{t+1}, a) \mid S_t = A, A_t = \mathrm{left}] \\
&= \gamma \mathbb{E}[\max_{1 \le i \le n} q_*(B, a_i)], \\
q_*(A, \mathrm{right}) &= \mathbb{E}[R_{t+1} + \gamma \max_a q_*(S_{t+1}, a) \mid S_t = A, A_t = \mathrm{right}] \\
&= 0, \\
q_*(B, a_i) &= \mathbb{E}[R_{t+1} + \gamma \max_a q_*(S_{t+1}, a) \mid S_t = B, A_t = a_i] \\
&= \mathbb{E}[\xi_i] \\
&= \mu.
\end{aligned} \tag{1}
$$

通过求解 Bellman 方程 (1), 我们得到

$$
\begin{aligned}
q_*(A, \mathrm{left}) &= \gamma\mu, \\
q_*(A, \mathrm{right}) &= 0, \\
q_*(B, a_i) &= \mu.
\end{aligned} \tag{2}
$$

当 $\mu < 0$ 的时候 (这就是例子中的情况), 我们的最优策略应该是在状态 A 时永远选择向右. 然而, 由于我们是在 ε-greedy 下进行学习, 因此最终我们只会以 $1 - \varepsilon$ 的概率选择向右, 而以 ε 的概率随机选择向右或者向左, 这样, 最终选择向左的概率就变成了 $\varepsilon/2$. 取 $\varepsilon = 0.1$ 就得到最终选择向左的概率为 $0.05 = 5\%$.

还有一种考虑方法. 我们把 ε-greedy 不再看成是我们采取的策略, 而看作环境的一部分. 具体来说, 假设当我们在状态 A 选择好一个动作后, 以 $1 - \varepsilon$ 的概率, 环境会让智能体执行我们选择的动作, 但以 ε 的概率去随机选择向左或者向右, 这种描述和 ε-greedy 是完全等价的. 在这种情况下, 最

优策略关于的 Bellman 方程满足:

$$
\begin{aligned}
\tilde{q}_*(A,\mathrm{left}) &= \mathbb{E}[R_{t+1}+\gamma \max_a \tilde{q}_*(S_{t+1},a) \mid S_t = A, A_t = \mathrm{left}] \\
&= (1-\varepsilon+\frac{\varepsilon}{2})\gamma\mathbb{E}[\max_{1\le i\le n} \tilde{q}_*(B,a_i)], \\
\tilde{q}_*(A,\mathrm{right}) &= \mathbb{E}[R_{t+1}+\gamma \max_a \tilde{q}_*(S_{t+1},a) \mid S_t = A, A_t = \mathrm{right}] \\
&= \frac{\varepsilon}{2}\gamma\mathbb{E}[\max_{1\le i\le n} \tilde{q}_*(B,a_i)], \\
\tilde{q}_*(B,a_i) &= \mathbb{E}[R_{t+1}+\gamma \max_a \tilde{q}_*(S_{t+1},a) \mid S_t = B, A_t = a_i] \\
&= \mathbb{E}[\xi_i] \\
&= \mu.
\end{aligned} \tag{3}
$$

求解 (3) 得到

$$
\begin{aligned}
\tilde{q}_*(A,\mathrm{left}) &= (1-\frac{\varepsilon}{2})\gamma\mu, \\
\tilde{q}_*(A,\mathrm{right}) &= \frac{\varepsilon}{2}\gamma\mu, \\
\tilde{q}_*(B,a_i) &= \mu.
\end{aligned} \tag{4}
$$

这样, 当 $\mu < 0$ 且 $0 < \varepsilon < 1$, 我们就有 $\tilde{q}_*(A,\mathrm{left}) < \tilde{q}_*(A,\mathrm{right})$, 从而最优策略也是在状态 A 时永远选择向右. 同样, 由于我们假定了在 A 时环境会以 ε 的概率去随机选择向左或者向右, 最终选择向左的概率还是 $\varepsilon/2$.

Q-learning 向左的动作比例真的会像图 1 中所展现的那样, 比 double Q-learning 多大约 5% 吗? 我们对此的答案是否定的! 我们已经知道, Q-learning 产生的 q 最终会收敛到按 ε-greedy 意义下的最优, 这表明从理论上来说, 与 double Q-learning 一样, Q-learning 的向左的动作比例也会是 5%. 图 1 产生这种结果的原因是数值实验的 episodes 的不够大, 为了证实这一点, 我们把 episodes 数增大到 2000, 结果见图 3. 在这种情况下, 最后一个 episode 中, double Q-learning 向左的动作比例为 5.03%, 而 Q-learning 向左的动作比例为 8.49%. 进一步的实验表明, 当 episodes 数为 3000 的时候, Q-learning 向左的动作比例约为 8%. 这表明, 随着 episodes 数继续增大, Q-learning 的向左的动作比例会趋向于 5%.

考虑如何让向左动作的概率趋于 0. 这需要在我们的 ε-greedy 中, 让参数 ε 随着 episodes 的增加而下降. 一种可行的方法是, 每过 100 个 episode, 就让 ε 变为原来的 η 倍. 我们取 $\eta = 0.8$, 这个条件下的数值结果见图 4. 此时在最后一个 episode 中, Q-learning 和 double Q-learning 向左的动作比例别为 0.14% 和 0.10%, 与最优策略已经非常接近了.

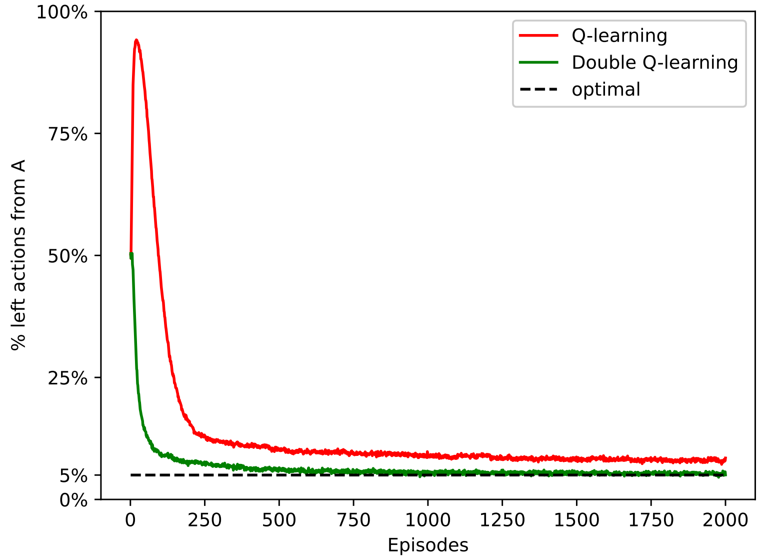

图 3: Maximization bias on long episodes.

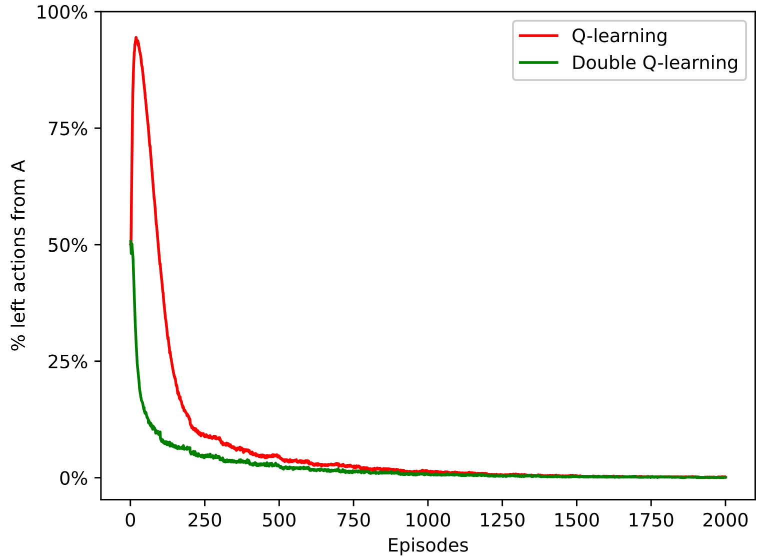

图 4: Maximization bias with refinement, $\eta = 0.8$.