Homework 3

Name: 方嘉聪 ID: 2200017849

Problem 1. 在强化学习的模型中, 我们使用 s_t 表示 t 时刻的状态, p 为转移概率, r 为 reward function, γ 为折扣因子. 算子 T 可以将一个 value function 转化为新的 value function. 定义

$$T(V)(s) := \max_{a\in A_s}\left[r(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)V(s')\right].$$

V^π 表示 policy π 对应的 value function:

$$V^\pi(s_t)=\sum_{t'=t}^{\infty}\mathbb{E}_{p,\pi}[\gamma^{t'-t}r(s_{t'},a_{t'})|s_t].$$

最优 policy $\pi^* := \operatorname{argmax}_\pi \mathbb{E}_{s_1}[V^\pi(s_1)]$. 记 $V^* := V^{\pi^*}$.

求证: $V^* = T(V^*)$, 即最优 policy 对应的 value function V^* 为 Bellman optimality operator T 的不动点. ◀

Solution. 注意到 V^π 满足 Bellman expectation equation, 即

$$V^\pi(s)=\mathbb{E}_{a\sim\pi}\left[r(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)V^\pi(s')\right].$$

而在最优策略 π^* 下, 对应的 value function V^* 满足 Bellman optimality equation, 即

$$V^*=V^{\pi^*}=\max_a\left[r(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)V^*(s')\right].$$

因此有 $V^*=T(V^*)$. 从而 V^* 是 Bellman optimality operator T 的不动点. 更严格的数学证明如下:

<u>注: 事实上, 我们还可以证明这一不动点是唯一的.</u>

定义空间中的距离函数为 $d(V_1,V_2)=\|V_1-V_2\|_\infty$. 在课上我们已经证明了如下引理:

Lemma 1. *T is* ***contraction mapping*** *with respect to the infinity norm $\|\cdot\|_\infty$, i.e., for any $V_1,V_2\in\mathbb{R}^{|S|}$, we have*

$$\|T(V_1)-T(V_2)\|_\infty\le\gamma\|V_1-V_2\|_\infty.$$

先证明序列 $\{T^n(V)\}\to V^*$, $n\to\infty$. 记 $d(\cdot)=\|\cdot\|_\infty$. 利用范数的三角不等式, 我们有

$$\begin{aligned}d(T^n(V),T^m(V))&\le d(T^n(V),T^{n+1}(V))+d(T^{n+1}(V),T^{m+1}(V))+d(T^m(V),T^{m+1}(V))\\ \text{(由引理)}\ &\le d(T^n(V),T^{n+1}(V))+\gamma d(T^n(V),T^m(V))+d(T^m(V),T^{m+1}(V))\end{aligned}$$

那么有

$$\begin{aligned}d(T^n(V),T^m(V))&\le\frac{d(T^n(V),T^{n+1}(V))+d(T^m(V),T^{m+1}(V))}{1-\gamma}\\ \text{(由引理)}\ &\le\frac{\gamma^n d(T(V),V)+\gamma^m d(T(V),V)}{1-\gamma}\\ &=\frac{\gamma^n+\gamma^m}{1-\gamma}d(T(V),V).\end{aligned}$$

考虑 $\gamma \in (0,1)$, 当 $n, m \to \infty$ 时, 有 $d(T^n(V), T^m(V)) \to 0$, 即序列 $\{T^n(V)\}$ 是 Cauchy 序列, 从而收敛到某个唯一的极限, 记为

$$V_{\text{opt}} := \lim_{n\to\infty} T^n(V). \implies V_{\text{opt}} = T(V_{\text{opt}}) = \max_a \left[r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) V_{\text{opt}}(s') \right].$$

下面我们来证明 $V_{\text{opt}} = V^*$. 注意 $V^*(s) = \max_\pi V^\pi(s)$. 记

$$T_\pi(V)(s) := \mathbb{E}_{a\sim\pi} \left[r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) V(s') \right].$$

1. 首先证明 $V^* \geq V_{\text{opt}}$. 对于任意 value function V 和 policy π, 有

$$\begin{aligned} d(V_\pi, V) &= \lim_{n\to\infty} d(T_\pi^n(V), V) \\ &\leq \sum_{r=1}^{\infty} d(T_\pi^r(V), T_\pi^{r-1}(V)), \quad \text{（由三角不等式）} \\ &\leq \sum_{r=1}^{\infty} \gamma^{r-1} d(T_\pi(V), V) \\ &= \frac{d(T_\pi(V), V)}{1-\gamma}. \end{aligned}$$

这里 $d(u,v) = \|u-v\|_\infty = \sup_s |u(s) - v(s)|$.

那么令 $V = V_{\text{opt}}$. 且对于任意 $\varepsilon > 0$, 取 $\pi = \pi_\varepsilon$ 使得

$$\begin{aligned} T_{\pi_\varepsilon}(V_{\text{opt}})(s) &= \mathbb{E}_{a\sim\pi_\varepsilon} \left[r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) V_{\text{opt}}(s') \right] \\ &\geq \max_a \left[r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) V_{\text{opt}}(s') \right] - \varepsilon(1-\gamma) \\ &= V_{\text{opt}}(s) - \varepsilon(1-\gamma). \\ \implies V_{\text{opt}}(s) - T_{\pi_\varepsilon}(V_{\text{opt}})(s) &\leq \varepsilon(1-\gamma). \end{aligned}$$

那么有 $d(V_{\pi_\varepsilon}, V_{\text{opt}}) \leq \varepsilon \implies V_{\pi_\varepsilon} \geq V_{\text{opt}} - \varepsilon$. 由 ε 的任意性, 有 $V^* \geq V_{\text{opt}}$.

2. 其次证明 $V^* \leq V_{\text{opt}}$. 这里我们需要用到 T_π 的单调性, 即对于任意两个 value function V_1, V_2 且 $V_1 \leq V_2$, 有 $T_\pi(V_1) \leq T_\pi(V_2), \forall \pi$. 由定义可知, 对于任意 policy π, 有

$$T_\pi(V_{\text{opt}}) \leq T(V_{\text{opt}}) \implies T_\pi^n(V_{\text{opt}}) \leq T_\pi^{n-1}(V_{\text{opt}}) \leq \cdots \leq T_\pi(V_{\text{opt}}) \leq T(V_{\text{opt}}).$$

这里重复使用 n 次 T_π 的单调性. 由 T_π 类似的 contraction mapping 性质, 可以证明

$$\lim_{n\to\infty} T_\pi^n(V_{\text{opt}}) \to V_\pi$$

那么对于任意 π, 有 $V_\pi \leq V_{\text{opt}}$. 故 $V^* \leq V_{\text{opt}}$.

综上, $V^* = V_{\text{opt}}$. 从而 V^* 是 Bellman optimality operator T 的唯一不动点. $\triangleleft$