Value Gap

1. Occupancy Measure定义

首先定义一个后续分析value gap中一个关键的定义：occupancy measure。occupancy measure分为state occupancy measure 和 state-action occupancy measure。

State Occupancy Measure

形式化定义为：

$$d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathrm{Pr}(s_t = s|\pi, d_0)$$

其中d_0为初始状态分布，上述定义表示的是一个指定的状态s在每个时间步出现的概率的累计折扣和。为什么需要折扣？因为价值函数的定义通常是累计折扣奖励，后面我们会看到这样定义的occupancy measure更方便对价值函数进行表示。为什么在最前面乘以系数$(1-\gamma)$？因为$\sum_{t=0}^{\infty}\gamma^t \mathrm{Pr}(s_t = s|\pi, d_0)$对$s$进行求和的结果等于$\frac{1}{1-\gamma}$，所以为了使$d_\pi$作为一个概率分布，需要对其进行归一化，所以需要乘以$(1-\gamma)$。

进一步，定义转移矩阵$P_\pi(s'|s) = \sum_a M^*(s'|s,a)\pi(a|s)$，其中$M^*(s'|s,a)$为环境转移矩阵，那么可以进一步简化定义：

$$d_\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t (P_\pi^t d_0)(s).$$

State-Action Occupancy Measure

类似于state occupancy measure，其定义为：

$$\begin{aligned}\rho_\pi(s,a) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathrm{Pr}(s_t = s, a_t = a|\pi, d_0) \\ &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \pi(a|s)\mathrm{Pr}(s_t = s|\pi, d_0) \\ &= \pi(a|s)d_\pi(s).\end{aligned}$$

Formalize Value with State-Action Occupancy Measure

$$
\begin{aligned}
V^\pi &= \mathbb{E}_{s_0\sim d_0, a_t\sim\pi(\cdot|s_t), s_{t+1}\sim P(\cdot|s_t,a_t)}[r(s_0,a_0)+\gamma r(s_1,a_1)+\gamma^2 r(s_2,a_2)+...] \\
&= \mathbb{E}_{s_0\sim d_0, a_0\sim\pi(\cdot|s_0)}[r(s_0,a_0)]+\gamma\mathbb{E}_{s_1\sim P_\pi, a_1\sim\pi(\cdot|s_1)}[r(s_1,a_1)]+\gamma^2\mathbb{E}_{s_2\sim P_\pi^2, a_2\sim\pi(\cdot|s_2)}[r(s_2,a_2)]+... \\
&= \sum_{s,a}\sum_{t=0}^{\infty}\gamma^t\pi(a|s)(P_\pi^t d_0)(s)\cdot r(s,a) \\
&= \sum_{s,a}\sum_{t=0}^{\infty}\gamma^t\pi(a|s)\Pr(s_t=s|\pi,d_0)\cdot r(s,a) \\
&= \frac{1}{1-\gamma}\sum_{s,a}\pi(a|s)d_\pi(s)\cdot r(s,a) \\
&= \frac{1}{1-\gamma}\sum_{s,a}\rho_\pi(s,a)\cdot r(s,a) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim\rho_\pi(s,a)}[r(s,a)].
\end{aligned}
$$

2 Value Gap under Different Policies

考虑两个不同的策略π, π_E，他们value gap的bound为：

$$
\begin{aligned}
|V^\pi - V^{\pi_E}| &= \left|\frac{1}{1-\gamma}\mathbb{E}_{s,a\sim\rho_\pi}[r(s,a)]-\frac{1}{1-\gamma}\mathbb{E}_{s,a\sim\rho_{\pi_E}(s,a)}[r(s,a)]\right| \\
&\le \frac{1}{1-\gamma}\sum_{s,a}|(\rho_\pi(s,a)-\rho_{\pi_E}(s,a))r(s,a)| \\
&\le \frac{2R_{\max}}{1-\gamma}D_{\mathrm{TV}}(\rho_\pi,\rho_{\pi_E}).
\end{aligned} \tag{1}
$$

我们注意到此时的bound中含有$D_{\mathrm{TV}}(\rho_\pi,\rho_{\pi_E})$，这一项我们一般很难估计，所以我们需要接着去放缩这一项。

放缩$D_{\mathrm{TV}}(d_\pi, d_{\pi_E})$

考虑到ρ_π可以由d_π表示，所以我们可以从$D_{\mathrm{TV}}(d_\pi, d_{\pi_E})$入手。

首先注意到d_π中的求和项有着简洁的解析形式：

$$
\sum_{t=0}^{\infty}\gamma^t P_\pi^t d_0 = (I-\gamma P_\pi)^{-1}d_0.
$$

接着令$M_\pi = (I-\gamma P_\pi)^{-1}, M_{\pi_E} = (I-\gamma P_{\pi_E})^{-1}$，那么：

$$
d_\pi - d_{\pi_E} = (1-\gamma)(M_\pi - M_{\pi_E})d_0.
$$

对于$M_\pi - M_{\pi_E}$，我们有：

$$M_\pi - M_{\pi_E} = M_\pi (M_{\pi_E}^{-1} - M_\pi^{-1}) M_{\pi_E} = \gamma M_\pi (P_\pi - P_{\pi_E}) M_{\pi_E}.$$

所以：

$$\begin{aligned} d_\pi - d_{\pi_\mathrm{E}} &= (1-\gamma)\gamma M_\pi \left(P_\pi - P_{\pi_\mathrm{E}}\right) M_{\pi_\mathrm{E}} d_0 \\ &= \gamma M_\pi \left(P_\pi - P_{\pi_\mathrm{E}}\right) d_{\pi_\mathrm{E}} \end{aligned} \tag{2}$$

因此：

$$\begin{aligned} D_{\mathrm{TV}}(d_\pi, d_{\pi_E}) &= \frac{\gamma}{2} \| M_\pi (P_\pi - P_{\pi_E}) d_{\pi_E} \|_1 \\ &\leq \frac{\gamma}{2} \| M_\pi \|_1 \| (P_\pi - P_{\pi_E}) d_{\pi_E} \|_1 . (\text{Cauchy-Schwarz inequality}) \end{aligned} \tag{3}$$

第一项M_π的上界为：

$$\| M_\pi \|_1 = \| \sum_{t=0}^{\infty} \gamma^t P_\pi^t \|_1 \leq \sum_{t=0}^{\infty} \gamma^t \| P_\pi \|_1^t \leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}.$$

第二项$\| (P_\pi - P_{\pi_E}) d_{\pi_E} \|_1$的上界为：

$$\begin{aligned} \| (P_\pi - P_{\pi_E}) d_{\pi_E} \|_1 &\leq \sum_{s,s'} | P_\pi(s'|s) - P_{\pi_E}(s'|s) | d_{\pi_E}(s) \\ &= \sum_{s,s'} \left| \sum_a M^*(s'|s,a)(\pi(a|s) - \pi_E(a|s)) \right| d_{\pi_E}(s) \\ &\leq \sum_{s,a,s'} M^*(s'|s,a) |\pi(a|s) - \pi_E(a|s)| d_{\pi_E}(s) \\ &= \sum_s d_{\pi_E}(s) \sum_a |\pi(a|s) - \pi_E(a|s)| \\ &= 2 \mathbb{E}_{s \sim d_{\pi_E}} [D_{\mathrm{TV}}(\pi_E(\cdot|s), \pi(\cdot|s))]. \end{aligned}$$

将这两个bound带入式(3)，得到：

$$D_{\mathrm{TV}}(d_\pi, d_{\pi_E}) \leq \frac{\gamma}{1-\gamma} \mathbb{E}_{s \sim d_{\pi_E}} [D_{\mathrm{TV}}(\pi(\cdot|s), \pi_E(\cdot|s))]. \tag{4}$$

放缩$D_{\mathrm{TV}}(\rho_\pi, \rho_{\pi_E})$

考虑到$\rho_\pi(s,a) = \pi(a|s) d_\pi(s)$，于是有：

$$
\begin{aligned}
&D_{\mathrm{TV}}\left(\rho_{\pi}, \rho_{\pi_{\mathrm{E}}}\right) \\
=&\frac{1}{2} \sum_{(s, a)}\left|\left[\pi_{\mathrm{E}}(a \mid s)-\pi(a \mid s)\right] d_{\pi_{\mathrm{E}}}(s)+\left[d_{\pi_{\mathrm{E}}}(s)-d_{\pi}(s)\right] \pi(a \mid s)\right| \\
\leq&\frac{1}{2} \sum_{(s, a)}\left|\pi_{\mathrm{E}}(a \mid s)-\pi(a \mid s)\right| d_{\pi_{\mathrm{E}}}(s)+\frac{1}{2} \sum_{(s, a)} \pi(a \mid s) \mid d_{\pi_{\mathrm{E}}}(s)-d_{\pi}(s) \\
=&\mathbb{E}_{s \sim d_{\pi_{\mathrm{E}}}}\left[D_{\mathrm{TV}}\left(\pi(\cdot \mid s), \pi_{\mathrm{E}}(\cdot \mid s)\right)\right]+D_{\mathrm{TV}}\left(d_{\pi}, d_{\pi_{\mathrm{E}}}\right) \\
\leq&\frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\pi_{\mathrm{E}}}}\left[D_{\mathrm{TV}}\left(\pi(\cdot \mid s), \pi_{\mathrm{E}}(\cdot \mid s)\right)\right] .
\end{aligned}
$$

带入式(1)可得：

$$
\begin{aligned}
\left|V^{\pi}-V^{\pi_{E}}\right| & \leq \frac{2 R_{\max }}{1-\gamma} D_{\mathrm{TV}}\left(\rho_{\pi}, \rho_{\pi_{E}}\right) \\
& \leq \frac{2 R_{\max }}{(1-\gamma)^{2}} \mathbb{E}_{s \sim d_{\pi_{E}}}\left[D_{\mathrm{TV}}\left(\pi(\cdot \mid s), \pi_{E}(\cdot \mid s)\right)\right] .
\end{aligned} \tag{5}
$$

进一步我们可以将此上界由TV散度转化为KL散度，由Pinsker's不等式，$D_{\mathrm{TV}}(\mu, \nu) \leq \sqrt{2 D_{\mathrm{KL}}(\mu, \nu)}$，于是有：

$$
\begin{aligned}
\left|V^{\pi}-V^{\pi_{E}}\right| & \leq \frac{2 R_{\max }}{(1-\gamma)^{2}} \mathbb{E}_{s \sim d_{\pi_{E}}}\left[\sqrt{2 D_{\mathrm{KL}}\left(\pi(\cdot \mid s), \pi_{E}(\cdot \mid s)\right)}\right] \\
& \leq \frac{2 \sqrt{2} R_{\max }}{(1-\gamma)^{2}} \sqrt{\mathbb{E}_{s \sim d_{\pi_{E}}}\left[D_{\mathrm{KL}}\left(\pi(\cdot \mid s), \pi_{E}(\cdot s)\right)\right]} .(\text{Jensen's inequality})
\end{aligned} \tag{6}
$$

3 Value Gap under Different Models

考虑model based的setting，即我们训练一个modelM_θ去拟合真实的环境转移模型M^*（假设奖励函数已知，我们只需要拟合状态的转移），此时一个自然的问题是：对于同一个policy，我们分别在真实环境和虚拟环境下去评估该policy，评估出的价值差距有多大呢，即$|V_{M_\theta}^\pi-V_M^\pi|$多大？

首先我们可以很容易得到：

$$
\begin{aligned}
\left|V_{M_{\theta}}^{\pi}-V_{M^{*}}^{\pi}\right| & =\frac{1}{1-\gamma} \sum_{s, a}\left(\rho_{\pi}^{M_{\theta}}(s, a)-\rho_{\pi}^{M^{*}}(s, a)\right) r(s, a) \\
& \leq \frac{R_{\max }}{1-\gamma} \sum_{s, a}\left|\rho_{\pi}^{M_{\theta}}(s, a)-\rho_{\pi}^{M^{*}}(s, a)\right| \\
& \leq \frac{R_{\max }}{1-\gamma} \sum_{s, a}\left|d_{\pi}^{M_{\theta}}(s, a)-d_{\pi}^{M^{*}}(s, a)\right| \sum_{a} \pi(a \mid s) \\
& =\frac{2 R_{\max }}{1-\gamma} D_{\mathrm{TV}}\left(d_{\pi}^{M_{\theta}}, d_{\pi}^{M^{*}}\right) .
\end{aligned} \tag{7}
$$

接下来我们放缩上式中的TV散度。类似于上一节推导中的式(2)，我们有：

$$d_\pi^{\pi_\theta} - d_\pi^{M^*} = \gamma G_\theta (P_\theta - P^*) d_\pi^{M^*},$$

其中，$P_\theta(s'|s) = \sum_a M_\theta(s'|s,a)\pi(a|s), G_\theta = (I - \gamma P_\theta)^{-1}$。于是TV散度可放缩成：

$$D_{\mathrm{TV}}(d_\pi^{M_\theta}, d_\pi^{M^*}) = \frac{\gamma}{2} \|G_\theta (P_\theta - P^*) d_\pi^{M^*}\|_1 \le \frac{\gamma}{2} \|G_\theta\|_1 \|(P_\theta - P^*) d_\pi^{M^*}\|_1.$$

第一项$\|G_\theta\|_1$可放缩成：

$$\|G_\theta\|_1 = \left\|\sum_{t=0}^{\infty} \gamma^t P_\theta^t\right\|_1 \le \sum_{t=0}^{\infty} \gamma^t \|P_\theta\|_1^t \le \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}.$$

第二项$\|(P_\theta - P^*) d_\pi^{M^*}\|_1$可放缩成：

$$\begin{aligned}
\left\|(P_\theta - P^*)\, d_\pi^{M^*}\right\|_1 &\le \sum_{s',s} \left|P_\theta(s'|s) - P^*(s'|s)\right| d_\pi^{M^*}(s) \\
&\le \sum_{s',s,a} \left|M_\theta(s'|s,a) - M^*(s'|s,a)\right| \pi(a|s) d_\pi^{M^*}(s) \\
&= 2\mathbb{E}_{(s,a)\sim\rho_\pi^{M^*}} \left[D_{\mathrm{TV}}\left(M_\theta(\cdot|s,a), M^*(\cdot|s,a)\right)\right].
\end{aligned}$$

类似于上一节的式6，再根据Pinsker和Jensen不等式：

$$D_{\mathrm{TV}}(d_\pi^{M_\theta}, d_\pi^{M^*}) \le \frac{\sqrt{2}\gamma}{2(1-\gamma)} \sqrt{\mathbb{E}_{(s,a)\sim\rho_\pi^{M^*}} \left[D_{\mathrm{KL}}\left(M^*(\cdot|s,a), M_\theta(\cdot|s,a)\right)\right]}.$$

带入式7可得：

$$|V_{M_\theta}^\pi - V_{M^*}^\pi| \le \frac{\sqrt{2} R_{\max}\gamma}{(1-\gamma)^2} \sqrt{\mathbb{E}_{(s,a)\sim\rho_\pi^{M^*}} \left[D_{\mathrm{KL}}\left(M^*(\cdot|s,a), M_\theta(\cdot|s,a)\right)\right]}.$$