Soft Q learning 公式证明

Git-123-Hub

2021-12-22

论文 Reinforcement Learning with Deep Energy-Based Policies 在最大熵的意义下，给出了基于能量的策略、soft Q 以及 soft V 的定义，这三者的定义和性质之间有关联的部分，同时给出了 soft Bellman 公式以及收敛保证，本文对以上内容进行公式推导。

标准的强化学习目标是找到可以使得折扣期望收益最大的策略 π:

$$\pi^*_{std} = \arg\max_\pi \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)] \tag{1}$$

论文中提到的方法不仅希望最大化策略的收益，同时希望可以最大化策略的熵，以此来鼓励探索（exploitation）。定义在最大熵意义下的策略为：

$$\pi^*_{MaxEnt} = \arg\max_\pi \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))] \tag{2}$$

1. soft Q 定义及性质

在标准的 Q 函数的基础上，将策略的熵考虑在内，定义 soft Q（原文中式 15）：

$$Q^\pi_{soft}(s,a) \triangleq r_0 + \mathbb{E}_{\tau\sim\pi,s_0=s,a_0=a}\left[\sum_{t=1}^{\infty}\gamma^t\left(r_t + \mathcal{H}\left(\pi\left(\cdot|s_t\right)\right)\right)\right] \tag{3}$$

我们可以从中得到一些它的简单性质：

$$\begin{aligned}
Q^\pi_{soft}(s_0,a_0) &= r_0 + \gamma[r_1 + \mathcal{H}(\pi(\cdot|s_1))] + \gamma^2[r_2 + \mathcal{H}(\pi(\cdot|s_2))] + \cdots \quad \text{将(3)逐项展开} \\
&= r_0 + \gamma\mathcal{H}(\pi(\cdot|s_1)) + \gamma(r_1 + \gamma(r_2 + \mathcal{H}(\pi(\cdot|s_2)) + \cdots)) \\
&\quad \text{将 } r_1 \text{ 合并到后一项中，注意红色括号中的内容符合的式(3)定义} \\
&= r_0 + \gamma(\mathcal{H}(\pi(\cdot|s_1)) + Q^\pi_{soft}(s_1,a_1))
\end{aligned}$$

由此我们可以得到 soft Q 的递归表达式：

$$Q^\pi_{soft}(s,a) = r + \gamma(\mathcal{H}(\pi(\cdot|s')) + Q^\pi_{soft}(s',a')) \tag{4}$$

式(4)表明：当前 s，a 的 soft Q 值等于奖励 r 加上下一状态中策略熵与 soft Q 值之和的折扣值，考虑到我们在最大熵策略式(2)以及 soft Q 的定义式(3)中引入了熵，这个公式看起来很自然。

2. 策略定义及其等价形式

原文式 3 给出了基于能量的策略 (energy-based model)

$$\pi(a_t|s_t) \propto exp(-\mathcal{E}(s_t,a_t)) \tag{5}$$

同时使用上文中定义的 soft Q 作为能量函数，即 $\mathcal{E}(s_t,a_t)=-\frac{1}{\alpha}Q_{soft}(s_t,a_t)$。**注：在下文的证明中一律省略常数项** α

我们先考虑一个简单例子，假如在某一状态 s 下，选择三个动作 a_1，a_2，a_3 的概率分别正比于 $\exp(Q_1)$，$\exp(Q_2)$，$\exp(Q_3)$，那么我们可以据此计算选择每个动作的概率：

$$\pi(a_i|s)=\frac{\exp(Q_i)}{\sum\limits_{t=1}^{3}\exp(Q_t)},\quad i=1,2,3 \tag{6}$$

由于其中能量函数的 exp，这看起来就和 softmax 差不多。

将这种想法扩展到连续动作，结合式(5)，并将求和改为积分，我们可以得到下式：

$$\pi(a_i|s_t)=\frac{\exp(Q^{\pi}_{soft}(s_t,a_i))}{\int_{\mathcal{A}}\exp(Q^{\pi}_{soft}(s_t,a'))\,da'} \tag{7}$$

其中分子部分为配分项（归一化常数）。

同时注意到原文式 5 给出的 soft V 的定义：

$$V^{\pi}_{soft}(s)\triangleq\log\int_{\mathcal{A}}\exp(Q^{\pi}_{soft}(s,a))\,da \tag{8}$$

两边同时取指数可以得到：

$$\exp(V^{\pi}_{soft}(s_t))=\int_{\mathcal{A}}\exp(Q^{\pi}_{soft}(s_t,a))\,da \tag{9}$$

将式(9)代入式(7)的分子部分可以得到：

$$\begin{aligned}\pi(a_i|s_t)&=\frac{\exp(Q^{\pi}_{soft}(s_t,a_i))}{\exp(V^{\pi}_{soft}(s_t))}\\&=\exp(Q^{\pi}_{soft}(s_t,a_i)-V^{\pi}_{soft}(s_t))\end{aligned} \tag{10}$$

由此我们可以通过 soft Q 和 soft V 将策略 π 表达出来，即式(5)和式(10)是等价的。更近一步，我们可以对式(10)两边同时取对数，即可得到：

$$\log\pi(a|s)=Q^{\pi}_{soft}(s,a)-V^{\pi}_{soft}(s) \tag{11}$$

3. 策略提升定理

策略提升定理：给定一个策略 π，定义新策略 $\tilde{\pi}$：

$$\tilde{\pi}(\cdot|s)\propto exp\left(Q^{\pi}_{soft}(s,\cdot)\right),\quad\forall s \tag{12}$$

假设在计算过程中无论是 π 还是 $\tilde{\pi}$，对于任何的 s，Q 和 $\int\exp(Q(s,a))\,da$ 的值都是有界的，那么

$$Q^{\tilde{\pi}}_{soft}(s,a)\geq Q^{\pi}_{soft}(s,a)\quad\forall s,a \tag{13}$$

其含义很明确：对于一个给定的策略 π，我们可以计算出其 soft Q 的值，并将其作为能量函数，按照式(5)构造新的策略 $\tilde{\pi}$，通过这种方法构造新策略，就可以保证 soft Q 意义下，$\tilde{\pi}$ 一定优于 π。

证明：根据前文中 soft Q 的性质，我们考虑式(4)中的递归项，并对其进行一些变换：

$$
\begin{aligned}
\mathcal{H}(\pi(\cdot|s)) + \mathbb{E}_{a\sim\pi}[Q^{\pi}_{soft}(s,a)] &= -\int \pi(a|s)\log\pi(a|s)\,da + \int \pi(a|s) Q^{\pi}_{soft}(a|s)\,da \\
&\text{利用式(11)将 } Q \text{ 替换为 } \log\tilde{\pi} + V\text{，注意这里是 } \tilde{\pi} \text{ 服从 } \exp(Q) \\
&= -\int \pi(a|s)\log\pi(a|s)\,da + \int \pi(a|s)\left(\log\tilde{\pi}(a|s) + V^{\pi}_{soft}(s)\right)\,da \\
&= -\int \pi(a|s)\left(\log\pi(a|s) - \log\tilde{\pi}(a|s)\right)\,da + \int \pi(a|s) V^{\pi}_{soft}(s)\,da \\
&= -\int \pi(a|s)\log(\frac{\pi(a|s)}{\tilde{\pi}(a|s)}) + V^{\pi}_{soft}(s)\int \pi(a|s)\,da \\
&\text{前项符合 } KL \text{ 散度的定义，后项积分为 } 1 \\
&= -D_{KL}\left(\pi(\cdot|s)||\tilde{\pi}(\cdot|s)\right) + V^{\pi}_{soft}(s) \quad \text{替换的 } softV \text{ 的定义} \\
&= -D_{KL}\left(\pi(\cdot|s)||\tilde{\pi}(\cdot|s)\right) + \log\int \exp(Q^{\pi}_{soft}(s,a)\,da
\end{aligned} \tag{14}
$$

同理可得：

$$
\begin{aligned}
\mathcal{H}(\tilde{\pi}(\cdot|s)) + \mathbb{E}_{a\sim\tilde{\pi}}[Q^{\pi}_{soft}(s,a)] &= -D_{KL}\left(\tilde{\pi}(\cdot|s)||\tilde{\pi}(\cdot|s)\right) + \log\int \exp(Q^{\pi}_{soft}(s,a)\,da \\
&= -0 + \log\int \exp(Q^{\pi}_{soft}(s,a))\,da
\end{aligned} \tag{15}
$$

显然：

$$
\mathcal{H}(\pi(\cdot|s)) + \mathbb{E}_{a\sim\pi}[Q^{\pi}_{soft}(s,a)] \le \mathcal{H}(\tilde{\pi}(\cdot|s)) + \mathbb{E}_{a\sim\tilde{\pi}}[Q^{\pi}_{soft}(s,a)] \tag{16}
$$

其意义也很直观，因为我们的目标就是希望策略的熵以及 Q 值能够越大越好，而公式(16)告诉我们，只要按照式(12)的方式构造新策略，我们确实可以保证这一点。

将式(16)代入之前介绍的 soft Q 的性质式(4)中，重复展开、代入，可以得到：

$$
\begin{aligned}
Q^{\pi}_{soft}(s,a) &= \mathbb{E}_{s_1}[r_0 + \gamma(\mathcal{H}(\pi(\cdot|s_1)) + \mathbb{E}_{a_1\sim\pi}[Q^{\pi}_{soft}(s_1,a_1)])] \\
&\le \mathbb{E}_{s_1}[r_0 + \gamma(\mathcal{H}(\tilde{\pi}(\cdot|s_1)) + \mathbb{E}_{a_1\sim\tilde{\pi}}[Q^{\pi}_{soft}(s_1,a_1)])] \\
&= \mathbb{E}_{s_1}[r_0 + \gamma(\mathcal{H}(\tilde{\pi}(\cdot|s_1)) + \mathbb{E}_{a_1\sim\tilde{\pi}}[r_1 + \gamma(\mathcal{H}(\pi(\cdot|s_2)) + \mathbb{E}_{a_2\sim\pi}[Q^{\pi}_{soft}(s_2,a_2)])])] \\
&= \mathbb{E}_{s_1}[r_0 + \gamma(\mathcal{H}(\tilde{\pi}(\cdot|s_1)) + r_1)] + \gamma^2\mathbb{E}_{s_2}[\mathcal{H}(\pi(\cdot|s_2)) + \mathbb{E}_{a_2\sim\pi}[Q^{\pi}_{soft}(s_2,a_2)]] \\
&\le \mathbb{E}_{s_1}[r_0 + \gamma(\mathcal{H}(\tilde{\pi}(\cdot|s_1)) + r_1)] + \gamma^2\mathbb{E}_{s_2}[\mathcal{H}(\tilde{\pi}(\cdot|s_2)) + \mathbb{E}_{a_2\sim\tilde{\pi}}[Q^{\pi}_{soft}(s_2,a_2)]] \\
&\vdots \\
&\le \mathbb{E}_{\tau\sim\tilde{\pi}}[r_0 + \sum_{t=1}^{\infty}\gamma^t(\mathcal{H}(\tilde{\pi}(\cdot|s_t)) + r_t)] \\
&= Q^{\tilde{\pi}}_{soft}(s,a)
\end{aligned} \tag{17}
$$

因此，如果我们通过以下方式进行策略迭代：

$$
\pi_{i+1}(\cdot|s) \propto \exp(Q^{\pi_i}_{soft}(s,\cdot)) \tag{18}
$$

那么 $Q^{\pi}_{soft}(s,a)$ 会单调增加直至收敛。

证毕。

4. Soft Bellman 方程

为了得到 soft Bellman 方程，我们还是对 soft Q 性质式(4)中的迭代项进行处理：

$$\begin{aligned}
\mathcal{H}(\pi(\cdot|s)) + \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^{\pi}_{soft}(s,a)] &= -\int \pi(a|s)\log\pi(a|s)\,da + \int \pi(a|s)Q^{\pi}_{soft}(a|s)\,da \\
&\quad \text{利用式(10)替换 } \log\pi\text{，注意这里是当前策略 } \pi \\
&= -\int \pi(a|s)(Q^{\pi}_{soft}(s,a) - V^{\pi}_{soft}(s))\,da + \int \pi(a|s)Q^{\pi}_{soft}(a|s)\,da \\
&= \int \pi(a|s)V^{\pi}_{soft}(s)\,da \\
&= V^{\pi}_{soft}(s)\int \pi(a|s)\,da \\
&= V^{\pi}_{soft}(s)
\end{aligned} \tag{19}$$

代入原来的式(4)中可得：

$$\begin{aligned}
Q^{\pi}_{soft}(s,a) &= r + \gamma(\mathcal{H}(\pi(\cdot|s')) + Q^{\pi}_{soft}(s',a')) \\
&= r + \gamma(V^{\pi}_{soft}(s'))
\end{aligned} \tag{20}$$

在 soft Bellman 方程式(20)的基础上，把其中的 soft V 按照定义式(8)展开，我们可以定义 soft value 迭代 $\mathcal{T}$：

$$\mathcal{T}Q(s,a) \triangleq r(s,a) + \gamma\mathbb{E}_{s'\sim p_s}\left[\log\int \exp Q(s',a')\,da'\right] \tag{21}$$

接下来我们证明这种迭代方法是可以使 soft Q 收敛的。

首先针对 soft Q 定义一个范数：$||Q_1 - Q_2|| \triangleq \max_{s,a}|Q_1(s,a) - Q_2(s,a)|$ 并令 $\epsilon = ||Q_1 - Q_2||$

$$\begin{aligned}
\log\int \exp Q_1(s',a')\,da' &\leq \log\int \exp(Q_2(s',a') + \epsilon)\,da' \\
&\quad \epsilon\text{ 是 } Q_1\text{，}Q_2\text{ 两者之中的大值，将其加到一个上一定会大于另外一个} \\
&= \log\left(\exp(\epsilon)\int \exp(Q_2(s',a'))\,da'\right) \\
&= \epsilon + \log\int \exp(Q_2(s',a'))\,da'
\end{aligned} \tag{22}$$

类似地，如果其中一个减去 ϵ（大值）结果一定小于另一个，我们可以得到：

$$\log\int \exp Q_1(s',a')\,da' \geq -\epsilon + \log\int \exp(Q_2(s',a'))\,da' \tag{23}$$

因此：$||\mathcal{T}Q_1 - \mathcal{T}Q_2|| \leq \gamma\epsilon = \gamma||Q_1 - Q_2||$ 即：运算符 $\mathcal{T}$ 会使 Q 值之间的差距缩小，最终收敛至最优的那一个。证毕。