1. MDP

$$\forall a \in \mathcal{A}, s \in \mathcal{S} : \pi \xrightarrow{p(s'|s,a)} MC = \left\{ \tau = (s_0, a_0, s_1, \ldots, s_{T-1}, a_{T-1}, s_T) \sim p(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t) \right\}$$
$$\xrightarrow{r(s,a,s'),\gamma} V(s_0; \pi) = \mathbb{E}^{\pi} \left\{ \sum_{t=0}^{T-1} \gamma^t r(s_t, a_t, s_{t+1}) \Big| s_0 \in \mathcal{S} \right\} \xrightarrow{p_0(s)} J(\pi) = \sum_s p_0(s) V(s; \pi).$$

1.1 Some Definitions

- The loss of policy π:

$$\begin{aligned} J(\pi) = \sum_{s_0} p_0(s_0) \sum_{a_0} \pi(a_0|s_0) \sum_{s_1} p(s_1|s_0, a_0) \Bigg(& r(s_0, a_0, s_1) \\ & + \sum_{a_1} \pi(a_1|s_1) \sum_{s_2} p(s_2|s_1, a_1) \gamma \Bigg(r(s_1, a_1, s_2) \\ & + \sum_{a_2} \pi(a_2|s_2) \sum_{s_3} p(s_3|s_2, a_2) \gamma^2 (\cdots) \Bigg) \Bigg); \end{aligned}$$

- The value function:

$$\begin{aligned} V(s_0; \pi) = \sum_{a_0} \pi(a_0|s_0) \sum_{s_1} p(s_1|s_0, a_0) \Bigg(& r(s_0, a_0, s_1) \\ & + \sum_{a_1} \pi(a_1|s_1) \sum_{s_2} p(s_2|s_1, a_1) \gamma \Bigg(r(s_1, a_1, s_2) \\ & + \sum_{a_2} \pi(a_2|s_2) \sum_{s_3} p(s_3|s_2, a_2) \gamma^2 (\cdots) \Bigg) \Bigg); \end{aligned}$$

- The state-action function (Q function):

$$\begin{aligned} Q(s_0, a_0; \pi) = \sum_{s_1} p(s_1|s_0, a_0) \Bigg(& r(s_0, a_0, s_1) \\ & + \sum_{a_1} \pi(a_1|s_1) \sum_{s_2} p(s_2|s_1, a_1) \gamma \Bigg(r(s_1, a_1, s_2) \\ & + \sum_{a_2} \pi(a_2|s_2) \sum_{s_3} p(s_3|s_2, a_2) \gamma^2 (\cdots) \Bigg) \Bigg); \end{aligned}$$

- The relationships in infinite horizon MDP:

$$\begin{cases} J(\pi) = \sum_s p_0(s) V(s; \pi), \\ V(s; \pi) = \sum_a \pi(a|s) Q(s, a; \pi), \\ Q(s, a; \pi) = \sum_s p(s'|s, a)(r(s, a, s') + \gamma V(s'; \pi)); \end{cases}$$

- $V^{\pi} = \begin{bmatrix} | \\ V(s; \pi) \\ | \end{bmatrix};$
- $\pi^* = \arg\max_{\pi} J(\pi), \quad V^* = V^{\pi^*}.$

1.2 Bellman Equation

> **Definition**: (Bellman equation) $\forall V \in \mathbb{R}^{|\mathcal{S}|}$,
>
> $$T^{\pi}V(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')].$$

> **Lemma**:
>
> $$V^{\pi} = T^{\pi}V^{\pi}.$$

1.2.1 Policy Based Algorithm

$$\theta_{\pi} \to \pi \to J(\pi) \to J(\theta_{\pi}).$$

The key tool of policy based algorithm is policy gradient: $\frac{dJ(\theta_{\pi})}{d\theta_{\pi}}$:

$$\theta_{\pi} = \theta_{\pi} + \alpha \frac{dJ(\theta_{\pi})}{d\theta_{\pi}}.$$

1.3 Optimal Bellman Equation and Algorithms

> **Definition**: (Optimal Bellman equation) $\forall V \in \mathbb{R}^{|S|}$,
>
> $$TV(s) = \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')].$$

> **Lemma**:
>
> $$V^{*} = TV^{*}.$$

1.3.1 Value Based Loss Function

$$\begin{aligned} L(V) =& \frac{1}{2}\|TV - V\|^2_{s\sim p_0} \\ =& \frac{1}{2}\sum_s p_0(s)\left\{\max_{\pi(\cdot|s)} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] - V(s)\right\}^2 \end{aligned}$$

1.3.2 Q Based Loss Function

> **Definition**: (Q-policy)
>
> $$\pi_Q(a|s) = 1\{a = \arg\max_{a'} Q(s,a')\}.$$

We notice that $\forall Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, there is a $V_Q = \sum_a \pi_Q(a|s)Q(s,a) \in \mathbb{R}^{|\mathcal{S}|}$.

Then the value based loss function becomes

$$L_1(Q) = \frac{1}{2}\sum_s p_0(s)\left\{\max_{\pi(\cdot|s)} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma \sum_{a'} \pi_Q(a'|s')Q(s',a')] - \sum_a \pi_Q(a|s)Q(s,a)\right\}^2.$$

We construct another loss function

$$
\begin{aligned}
L_2(Q) =& \frac{1}{2}\sum_s p_0(s)\left\{\max_\pi \sum_a \pi(a|s)\sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma\sum_{a'}\pi(a'|s')Q(s',a')] - \sum_a \pi(a|s)Q(s,a)\right\}^2 \\
=& \frac{1}{2}\sum_s p_0(s)\left\{\max_\pi \sum_a \pi(a|s)\left[\sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma\sum_{a'}\pi(a'|s')Q(s',a')] - Q(s,a)\right]\right\}^2 \\
=& \frac{1}{2}\sum_s p_0(s)\left\{\sum_a \pi_Q(a|s)\left[\sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma\sum_{a'}\pi_Q(a'|s')Q(s',a')] - Q(s,a)\right]\right\}^2 \\
=& \frac{1}{2}\sum_s p_0(s)\sum_a \pi_Q(a|s)\left\{\sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma\sum_{a'}\pi_Q(a'|s')Q(s',a')] - Q(s,a)\right\}^2 \\
& (\text{The property of } \pi_Q)
\end{aligned}
$$

1.3.3 The Algorithms

> **Definition**: (Q-ϵ policy)
>
> $$\pi_{Q,\epsilon} = (1-\epsilon)\pi_Q + \epsilon\pi_{uniform}.$$

- Q-learning algorithm:

$$L_3(Q) = \sum_s p(s)\sum_a \pi_{Q,\epsilon}(a|s)\left\{\sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma\sum_{a'}\pi_Q(a'|s')Q(s',a')\right] - Q(s,a)\right\}^2$$

- SARSA:

$$L_3(Q) = \sum_s p(s)\sum_a \pi_{Q,\epsilon}(a|s)\left\{\sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma\sum_{a'}\pi_{Q,\epsilon}(a'|s')Q(s',a')\right] - Q(s,a)\right\}^2$$

- DQN:

$$
\begin{aligned}
L_Q(\theta_Q) =& \frac{1}{2}\sum_s p_{replay}(s)\sum_a \pi_{replay}(a|s) \\
& \left\{\sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma\sum_{a'}\pi_Q(a'|s')Q(s',a';\theta_{Q_{target}})\right] - Q(s,a;\theta_Q)\right\}^2 \\
L_{Q_{target}}(\theta_{Q_{target}}) =& \frac{1}{2}\|\theta_{Q_{target}} - \theta_Q\|_2^2
\end{aligned}
$$

2. Soft Actor Critic

2.1 Some Definitions

- The loss of policy π:

$$
\begin{aligned}
J_{soft}(\pi) = \sum_{s_0} p_0(s_0)\sum_{a_0}\pi(a_0|s_0)\sum_{s_1} p(s_1|s_0,a_0)\Bigg(& r(s_0,a_0,s_1) - \alpha\log\pi(a_0|s_0) \\
& + \sum_{a_1}\pi(a_1|s_1)\sum_{s_2} p(s_2|s_1,a_1)\gamma\bigg(r(s_1,a_1,s_2) - \alpha\log\pi(a_1|s_1) \\
& + \sum_{a_2}\pi(a_2|s_2)\sum_{s_3} p(s_3|s_2,a_2)\gamma^2(\cdots)\bigg)\Bigg)
\end{aligned}
$$

- The value function:

$$
\begin{aligned}
V_{soft}(s_0;\pi) = \sum_{a_0}\pi(a_0|s_0)\sum_{s_1}p(s_1|s_0,a_0)\Bigg(r(s_0,a_0,s_1) - \alpha\log\pi(a_0|s_0) \\
+\sum_{a_1}\pi(a_1|s_1)\sum_{s_2}p(s_2|s_1,a_1)\gamma\Bigg(r(s_1,a_1,s_2) - \alpha\log\pi(a_1|s_1) \\
+\sum_{a_2}\pi(a_2|s_2)\sum_{s_3}p(s_3|s_2,a_2)\gamma^2(\cdots)\Bigg)\Bigg)
\end{aligned}
$$

- The state-value function:

$$
\begin{aligned}
Q_{soft}(s_0,a_0;\pi) = \sum_{s_1}p(s_1|s_0,a_0)\Bigg(r(s_0,a_0,s_1) - \alpha\log\pi(a_0|s_0) \\
+\sum_{a_1}\pi(a_1|s_1)\sum_{s_2}p(s_2|s_1,a_1)\gamma\Bigg(r(s_1,a_1,s_2) - \alpha\log\pi(a_1|s_1) \\
+\sum_{a_2}\pi(a_2|s_2)\sum_{s_3}p(s_3|s_2,a_2)\gamma^2(\cdots)\Bigg)\Bigg)
\end{aligned}
$$

- The relationships in infinite horizon MDP:

$$
\begin{cases}
J_{soft}(\pi) = \sum_s p_0(s_0)V_{soft}(s;\pi) \\
V_{soft}(s;\pi) = \sum_a \pi(a|s)Q_{soft}(s,a;\pi) \\
Q_{soft}(s,a;\pi) = \sum_s p(s'|s,a)(r(s,a,s') - \alpha\log\pi(a|s) + V_{soft}(s';\pi)) \\
V_{soft}(s;\pi) = \sum_a \pi(a|s)\sum_s p(s'|s,a)(r(s,a,s') + V_{soft}(s';\pi)) - \alpha\sum_a \pi(a|s)\log\pi(a|s)
\end{cases}
$$

- $V^\pi_{soft} = \begin{bmatrix} | \\ V_{soft}(s;\pi) \\ | \end{bmatrix}$;
- $\pi^*_{soft} = \arg\max_\pi J_{soft}(\pi), \quad V^*_{soft} = V^{\pi^*}_{soft}$.

2.2 Soft Bellman Equation

> **Definition**: (Soft Bellman equation) $\forall V \in \mathbb{R}^{|\mathcal{S}|}$,
>
> $$T^\pi_{soft}V(s) = \sum_a \pi(a|s)\sum_{s'}p(s'|s,a)[r(s,a,s') - \alpha\log\pi(a|s) + \gamma V(s')].$$

If we denote $Q_V(s,a) = \sum_{s'} p(s'|s,a)\,(r(s,a,s') + V(s'))$, then

$$Q_{V,soft}(s,a) = \sum_{s'}p(s'|s,a)\,(r(s,a,s') - \alpha\log\pi(a|s) + V(s')) = Q_V(s,a) - \alpha\log\pi(a|s).$$

We remind Bellman Equation in here:

$$T^\pi V(s) = \sum_a \pi(a|s)\sum_{s'}p(s'|s,a)\,(r(s,a,s') + V(s')) = \langle\pi(\cdot|s), Q_V(s,\cdot)\rangle,$$

and we notice that

$$T^\pi_{soft}V(s) = T^\pi V(s) - \alpha\sum_a \pi(a|s)\log\pi(a|s)$$

2.3 Soft Optimal Bellman Equation

> **Definition**: (Soft optimal Bellman equation) $\forall V \in \mathbb{R}^{|S|}$,
>
> $$T_{soft}V(s) = \max_{\pi(\cdot|s)}\sum_a \pi(a|s)\sum_{s'}p(s'|s,a)[r(s,a,s') - \alpha\log\pi(a|s) + \gamma V(s')].$$

Note that

$$
\begin{aligned}
\pi_{V,soft}(\cdot|s) :=&\arg\max_{\pi(\cdot|s)} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') - \alpha \log \pi(a|s) + \gamma V(s')] \\
=&\frac{\exp\{\frac{1}{\alpha}\sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma V(s')\right]\}}{\sum_{a'} \exp\{\frac{1}{\alpha}\sum_{s'} p(s'|s,a')\left[r(s,a',s') + \gamma V(s')\right]\}} = \frac{\exp\{\frac{1}{\alpha}Q_V(s,a)\}}{\sum_{a'}\exp\{\frac{1}{\alpha}Q_V(s,a')\}},
\end{aligned}
$$

and

$$
T_{soft}V(s) = \alpha \log\left\{\sum_a \exp\left[\frac{1}{\alpha}Q_V(s,a)\right]\right\}.
$$

> **Lemma**:
>
> $$V^*_{soft} = T_{soft}V^*_{soft}.$$

2.3.1 V Based Loss Function

$$
\begin{aligned}
L(V) =&\frac{1}{2}\|T_{soft}V - V\|^2_{s\sim p_0} \\
=&\frac{1}{2}\sum_s p_0(s)\left\{\max_{\pi(\cdot|s)}\sum_a \pi(a|s)\sum_{s'} p(s'|s,a)[r(s,a,s') - \alpha\log\pi(a|s) + \gamma V(s')] - V(s)\right\}^2
\end{aligned}
$$

2.3.2 The idea of SAC

The algorithm is a V-based method.

- First we have three net: $V(s;\theta_V)$, $Q(s,a;\theta_Q)$ and $\pi(a|s;\theta_\pi)$.
- We want $Q(s,a;\theta_Q) = Q_V = \sum_{s'} p(s'|s,a)\left(r(s,a,s') + \gamma V(s';\theta_V)\right)$;

$$
L_1(\theta_Q) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left\{\frac{1}{2}(\sum_{s'} p(s'|s,a)\left(r(s,a,s') + \gamma V(s';\theta_V)\right) - Q(s,a;\theta_Q))^2\right\};
$$

- We want $\pi(a|s;\theta_\pi) = \pi^*_{V,soft}(a|s) = \frac{exp(\frac{1}{\alpha}Q(s,a;\theta_Q))}{\sum_{a'}\exp(\frac{1}{\alpha}Q(s,a';\theta_Q))}$;

$$
\begin{aligned}
L_2(\theta_\pi) =&\mathbb{E}_{s\sim\mathcal{D}}\left\{D_{KL}\left(\pi(\cdot|s;\theta_\pi)\|\pi^*_{Q,soft}(s,\cdot)\right)\right\} \\
=&\mathbb{E}_{s\sim\mathcal{D},a\sim\pi(\cdot|s;\theta_\pi)}\left\{\log(\pi(a|s;\theta_\pi)) - log(\pi^*_{Q,soft}(a|s))\right\} \\
=&\mathbb{E}_{s\sim\mathcal{D},a\sim\pi(\cdot|s;\theta_\pi)}\left\{\log(\pi(a|s;\theta_\pi)) - \frac{1}{\alpha}Q(s,a;\theta_Q) + \log\left(\sum_a \exp\left\{\frac{1}{\alpha}Q(s,a;\theta_Q)\right\}\right)\right\}
\end{aligned}
$$

$$
\begin{aligned}
&D_{KL}\left(\pi(\cdot|s;\theta_\pi)\middle\|\frac{\exp(\frac{1}{\alpha}Q(\cdot|s;\theta_Q))}{Z(s;\theta_Q)}\right) \\
=&\int_a \pi(a|s;\theta_\pi)\left(\ln\pi(a|s;\theta_\pi) - \frac{1}{\alpha}Q(a|s;\theta_Q) + \ln Z(s;\theta_Q)\right)da \\
=&\int_\epsilon \pi(\tanh(\sigma_{\theta_\pi}\epsilon + \mu_{\theta_\pi})|s;\theta_\pi)\left(\ln\pi(\tanh(\sigma_{\theta_\pi}\epsilon + \mu_{\theta_\pi})|s;\theta_\pi)\right. \\
&\left.-\frac{1}{\alpha}Q(\tanh(\sigma_{\theta_\pi}\epsilon + \mu_{\theta_\pi})|s;\theta_Q) + \ln Z(s;\theta_Q)\right)d\tanh(\sigma_{\theta_\pi}\epsilon + \mu_{\theta_\pi}) \\
=&\int_\epsilon p(\epsilon)\left(\ln\pi(\tanh(\sigma_{\theta_\pi}\epsilon + \mu_{\theta_\pi})|s;\theta_\pi) - \frac{1}{\alpha}Q(\tanh(\sigma_{\theta_\pi}\epsilon + \mu_{\theta_\pi})|s;\theta_Q) + \ln Z(s;\theta_Q)\right)d\epsilon,
\end{aligned}
$$

- We want $V(s;\theta_V) = T^\pi_{soft}V(s;\theta_V) = \mathbb{E}_{a\sim\pi(\cdot|s;\theta_\pi)}\left[Q(s,a;\theta_Q) - \alpha\log(\pi(a|s;\theta_\pi))\right]$;

$$L_3(\theta_V) = \mathbb{E}_{s\sim\mathcal{D}}\left\{\frac{1}{2}\left(V(s;\theta_V) - \mathbb{E}_{a\sim\pi(\cdot|s;\theta_\pi)}\left[Q(s,a;\theta_Q) - \alpha\log(\pi(a|s;\theta_\pi))\right]\right)^2\right\}$$

Overall, the algorithm of SAC combining two techniques, target net and double q net, becomes:

$$\begin{cases} L_1(\theta_{Q_1}) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left\{\frac{1}{2}(\sum_{s'} p(s'|s,a)\left(r(s,a,s') + \gamma V(s';\theta_{V_{target}})\right) - Q_1(s,a;\theta_{Q_1}))^2\right\} \\ L_2(\theta_{Q_2}) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left\{\frac{1}{2}(\sum_{s'} p(s'|s,a)\left(r(s,a,s') + \gamma V(s';\theta_{V_{target}})\right) - Q_2(s,a;\theta_{Q_2}))^2\right\} \\ L_3(\theta_\pi) = \mathbb{E}_{s\sim\mathcal{D},\epsilon\sim\mathcal{N}(0,1)}\left\{\log(\pi(f(s;\epsilon,\theta_\pi)|s)) - \frac{1}{\alpha}\min\{Q_1(s,f(s;\epsilon,\theta_\pi);\theta_{Q_1}),Q_2(s,f(s;\epsilon,\theta_\pi);\theta_{Q_2})\}\right\} \\ J_4(\theta_V) = \mathbb{E}_{s\sim\mathcal{D}}\left\{\frac{1}{2}\left(V(s;\theta_V) - \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[Q(s,f(s;\epsilon,\theta_\pi);\theta_Q) - \alpha\log(\pi(f(s;\epsilon,\theta_\pi)|s;\theta_\pi))\right]\right)^2\right\} \\ J_5(\theta_{V_{target}}) = \frac{1}{2}\|\theta_V - \theta_{V_{target}}\|^2 \end{cases}$$

3. A Two Players MDPs Game

- One player MDP:

$$\begin{aligned} \forall a\in\mathcal{A}, s\in\mathcal{S}: \pi &\xrightarrow{p(s'|s,a)} MC = \left\{\tau = (s_0,a_0,s_1,\ldots,s_{T-1},a_{T-1},s_T) \sim p(s_0)\prod_{t=0}^{T-1}\pi(a_t|s_t)p(s_{t+1}|s_t,a_t)\right\} \\ &\xrightarrow{r(s,a,s'),\gamma} V(s_0;\pi) = \mathbb{E}^\pi\left\{\sum_{t=0}^{T-1}\gamma^t r(s_t,a_t,s_{t+1})\Big|s_0\in\mathcal{S}\right\} \\ &\xrightarrow{p_0(s)} J(\pi) = \sum_s p_0(s)V(s;\pi). \end{aligned}$$

- Two players MDPs:

$$\begin{aligned} \forall a,b\in\mathcal{A}, s\in\mathcal{S}: \pi_a,\pi_b &\xrightarrow{p(s'|s,a,b)} MC = \Bigg\{\tau = (s_0,a_0,b_0,s_1,\ldots,s_{T-1},a_{T-1},b_{T-1},s_T) \\ &\qquad \sim p(s_0)\prod_{t=0}^{T-1}\pi_a(a_t|s_t)\pi_b(b_t|s_t,a_t)p(s_{t+1}|s_t,a_t,b_t)\Bigg\} \\ &\xrightarrow{r(s,a,b,s'),\gamma} V^\pi(s_0) = \mathbb{E}^\pi\left\{\sum_{t=0}^{T-1}\gamma^t r(s_t,a_t,b_t,s_{t+1})\Big|s_0\in\mathcal{S}\right\} \\ &\xrightarrow{p_0(s)} J(\pi) = \sum_s p_0(s)V^\pi(s;\pi). \end{aligned}$$

3.1 Some Definitions

- The loss of policies:

$$\begin{aligned} J(\pi_a,\pi_b) = \sum_{s_0}p_0(s_0)\sum_{a_0}\pi_a(a_0|s_0)\sum_{b_0}\pi_b(b_0|s_0,a_0)\sum_{s_1}p(s_1|s_0,a_0,b_0)\Bigg(r(s_0,a_0,b_0,s_1) + \\ \sum_{a_1}\pi_a(a_1|s_1)\sum_{b_1}\pi_b(b_1|s_1,a_1)\sum_{s_2}p(s_2|s_1,a_1,b_1)\gamma\Bigg(r(s_1,a_1,b_1,s_2) + \\ \sum_{a_2}\pi_a(a_2|s_2)\sum_{b_2}\pi_b(b_2|s_2,a_2)\sum_{s_3}p(s_3|s_2,a_2,b_2)\gamma^2(\cdots)\Bigg)\Bigg) \end{aligned}$$

- Value function of player a:

$$V(s_0|\pi_a,\pi_b)=\sum_{a_0}\pi_a(a_0|s_0)\sum_{b_0}\pi_b(b_0|s_0,a_0)\sum_{s_1}p(s_1|s_0,a_0,b_0)\Bigg(r(s_0,a_0,b_0,s_1)+$$

$$\sum_{a_1}\pi_a(a_1|s_1)\sum_{b_1}\pi_b(b_1|s_1,a_1)\sum_{s_2}p(s_2|s_1,a_1,b_1)\gamma\Bigg(r(s_1,a_1,b_1,s_2)+$$

$$\sum_{a_2}\pi_a(a_2|s_2)\sum_{b_2}\pi_b(b_2|s_2,a_2)\sum_{s_3}p(s_3|s_2,a_2,b_2)\gamma^2(\cdots)\Bigg)\Bigg)$$

- Q function of player a, and value function of player b:

$$QV(s_0,a_0|\pi_a,\pi_b)=\sum_{b_0}\pi_b(b_0|s_0,a_0)\sum_{s_1}p(s_1|s_0,a_0,b_0)\Bigg(r(s_0,a_0,b_0,s_1)+$$

$$\sum_{a_1}\pi_a(a_1|s_1)\sum_{b_1}\pi_b(b_1|s_1,a_1)\sum_{s_2}p(s_2|s_1,a_1,b_1)\gamma\Bigg(r(s_1,a_1,b_1,s_2)+$$

$$\sum_{a_2}\pi_a(a_2|s_2)\sum_{b_2}\pi_b(b_2|s_2,a_2)\sum_{s_3}p(s_3|s_2,a_2,b_2)\gamma^2(\cdots)\Bigg)\Bigg)$$

- Q function of player b:

$$Q(s_0,a_0,b_0|\pi_a,\pi_b)=\sum_{s_1}p(s_1|s_0,a_0,b_0)\Bigg(r(s_0,a_0,b_0,s_1)+$$

$$\sum_{a_1}\pi_a(a_1|s_1)\sum_{b_1}\pi_b(b_1|s_1,a_1)\sum_{s_2}p(s_2|s_1,a_1,b_1)\gamma\Bigg(r(s_1,a_1,b_1,s_2)+$$

$$\sum_{a_2}\pi_a(a_2|s_2)\sum_{b_2}\pi_b(b_2|s_2,a_2)\sum_{s_3}p(s_3|s_2,a_2,b_2)\gamma^2(\cdots)\Bigg)\Bigg)$$

- The relationships in infinite horizon MDP:

$$\begin{cases} J(\pi_a,\pi_b)=\sum_{s\in\mathcal{S}}p_0(s)V(s|\pi_a,\pi_b) \\ V(s|\pi_a,\pi_b)=\sum_a\pi_a(a|s)QV(s,a|\pi_a,\pi_b) \\ QV(s,a|\pi_a,\pi_b)=\sum_b\pi_b(b|s,a)Q(s,a,b|\pi_a,\pi_b) \\ Q(s,a,b|\pi_a,\pi_b)=\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s')) \end{cases}$$

3.2 Bellman Equation and Multi-task Learning

For all $V\in\mathbb{R}^{|\mathcal{S}|}$,

$$T_{\pi_b}V(s)=\max_{\pi_a(\cdot|s)}\sum_a\pi_a(a|s)\sum_b\pi_b(b|s,a)\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s')).$$

In original optimal Bellman equation,

$$V_{*,\pi_b}=T_{\pi_b}V_{*,\pi_b}.$$

We denote $\pi_a^*(\pi_b)$ that satisfies $V_{\pi_a^*(\pi_b),\pi_b}=V_{*,\pi_b}$.

Multi-task learning problem: can we get $\pi_a^*(\pi_b)$ more effectively?

But this model is too general to solve the problem.

3.3 Optimal Bellman Equation

3.3.1 Cooperative Game

> **Definition** (Cooperative optimal Bellman equation). $\forall V\in\mathbb{R}^{|\mathcal{S}|}$,

> $$TV(s) = \max_{\pi_a(\cdot|s),\pi_b(\cdot|s,\cdot)} \sum_a \pi_a(a|s) \sum_b \pi_b(b|s,a) \sum_{s'} p(s'|s,a,b)(r(s,a,b,s') + \gamma V(s')).$$

Let us talk about the loss of algorithms:

$$\begin{aligned} L_1(Q) = &\sum_s p(s) \Bigg\{ \max_{\pi_a(\cdot|s),\pi_b(\cdot|s,\cdot)} \sum_a \pi_a(a|s) \sum_b \pi_b(b|s,a) \Bigg[\sum_{s'} p(s'|s,a,b)(r(s,a,b,s') \\ &+ \gamma \sum_{a'} \pi_a(a'|s') \sum_b \pi_b(b'|s',a') Q(s',a',b')) - Q(s,a,b) \Bigg] \Bigg\}^2 \\ = &\sum_s p(s) \sum_a \pi_{a,Q}(a|s) \sum_b \pi_{b,Q}(b|s,a) \Bigg\{ \sum_{s'} p(s'|s,a,b)(r(s,a,b,s') \\ &+ \gamma \sum_{a'} \pi_a(a'|s') \sum_b \pi_b(b'|s',a') Q(s',a',b')) - Q(s,a,b) \Bigg\}^2. \end{aligned}$$

The q-learning algorithm's loss function is

$$\begin{aligned} L_2(Q) = &\sum_s p(s) \sum_a \pi_{a,Q,\epsilon}(a|s) \sum_b \pi_{b,Q,\epsilon}(b|s,a) \Bigg\{ \sum_{s'} p(s'|s,a,b)(r(s,a,b,s') \\ &+ \gamma \sum_{a'} \pi_a(a'|s') \sum_b \pi_b(b'|s',a') Q(s',a',b')) - Q(s,a,b) \Bigg\}^2. \end{aligned}$$

If we already get the Q function:

$$Q(s,\cdot,\cdot) = \begin{bmatrix} Q(s,a_1,b_1) & Q(s,a_1,b_2) & \cdots & Q(s,a_1,b_n) \\ Q(s,a_2,b_1) & Q(s,a_2,b_2) & \cdots & Q(s,a_2,b_n) \\ \vdots & \vdots & \ddots & \vdots \\ Q(s,a_m,b_1) & Q(s,a_m,b_2) & \cdots & Q(s,a_m,b_n) \end{bmatrix},$$

then we can denote that

$$\pi_{b,Q}(s,a_i) = \arg\max_b Q(s,a_i,b)$$

and

$$\pi_{a,Q}(s) = \arg\max_a Q(s,a,\pi_{b,Q}).$$

3.3.2 Zero-Sum Game

> **Definition** (Zero-sum optimal Bellman equation). $\forall V \in \mathbb{R}^{|S|}$,
>
> $$TV(s) = \max_{\pi_a(\cdot|s)} \min_{\pi_b(\cdot|s,\cdot)} \sum_a \pi_a(a|s) \sum_b \pi_b(b|s,a) \sum_{s'} p(s'|s,a,b)(r(s,a,b,s') + \gamma V(s')).$$

We still can use L_2:

$$\begin{aligned} L_2(Q) = &\sum_s p(s) \sum_a \pi_{a,Q,\epsilon}(a|s) \sum_b \pi_{b,Q,\epsilon}(b|s,a) \Bigg\{ \sum_{s'} p(s'|s,a,b)(r(s,a,b,s') \\ &+ \gamma \sum_{a'} \pi_a(a'|s') \sum_b \pi_b(b'|s',a') Q(s',a',b')) - Q(s,a,b) \Bigg\}^2. \end{aligned}$$

If we already get the Q function:

$$Q(s,\cdot,\cdot)=\begin{bmatrix} Q(s,a_1,b_1) & Q(s,a_1,b_2) & \cdots & Q(s,a_1,b_n) \\ Q(s,a_2,b_1) & Q(s,a_2,b_2) & \cdots & Q(s,a_2,b_n) \\ \vdots & \vdots & \ddots & \vdots \\ Q(s,a_m,b_1) & Q(s,a_m,b_2) & \cdots & Q(s,a_m,b_n) \end{bmatrix},$$

then we can denote that

$$\pi_{b,Q}(s,a_i)=\arg\min_b Q(s,a_i,b)$$

and

$$\pi_{a,Q}(s)=\arg\max_a Q(s,a,\pi_{b,Q}).$$

3.4 Soft Optimal Bellman Equation

3.4.1 Cooperative Game

> **Definition** (Cooperative optimal Bellman equation). $\forall V\in\mathbb{R}^{|\mathcal{S}|}$,
>
> $$\begin{aligned} TV(s)=\max_{\pi_a(\cdot|s),\pi_b(\cdot|s,\cdot)}\sum_a\pi_a(a|s)\sum_b\pi_b(b|s,a)\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s') \\ -\alpha\log\pi_a(a|s)-\beta\log\pi_b(b|s,a)). \end{aligned}$$

$$\pi_{b,V}(b|s,a)=\frac{\exp\left\{\frac{1}{\beta}\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s'))\right\}}{\sum_{b'}\exp\left\{\frac{1}{\beta}\sum_{s'}p(s'|s,a,b')(r(s,a,b',s')+\gamma V(s'))\right\}}$$

$$\pi_{a,V}(a|s)=\frac{\exp\{\frac{1}{\alpha}\sum_b\pi_{b,V}(b|s,a)\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s'))\}}{\sum_{a'}\exp\{\frac{1}{\alpha}\sum_b\pi_{b,V}(b|s,a')\sum_{s'}p(s'|s,a',b)(r(s,a',b,s')+\gamma V(s'))\}}$$

$$\begin{aligned} Q_V(s,\cdot,\cdot)&=\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s')) \\ &=\begin{bmatrix} Q_V(s,a_1,b_1) & Q_V(s,a_1,b_2) & \cdots & Q_V(s,a_1,b_n) \\ Q_V(s,a_2,b_1) & Q_V(s,a_2,b_2) & \cdots & Q_V(s,a_2,b_n) \\ \vdots & \vdots & \ddots & \vdots \\ Q_V(s,a_m,b_1) & Q_V(s,a_m,b_2) & \cdots & Q_V(s,a_m,b_n) \end{bmatrix} \end{aligned}$$

$$\pi_{b,V}(v|s,a)=\frac{\exp\left\{\frac{1}{\beta}Q_V(s,a,b)\right\}}{\sum_{b'}\exp\left\{\frac{1}{\beta}Q_V(s,a,b')\right\}}$$

$$\pi_{a,V}(a|s)=\frac{\exp\{\frac{1}{\alpha}\sum_b\pi_{b,V}(b|s,a)Q_V(s,a,b)\}}{\sum_{a'}\exp\{\frac{1}{\alpha}\sum_b\pi_{b,V}(b|s,a')Q_V(s,a,b')\}}$$

3.4.2 Zero-sum Game

> **Definition** (Zero-sum optimal Bellman equation). $\forall V\in\mathbb{R}^{|\mathcal{S}|}$,
>
> $$\begin{aligned} TV(s)=\max_{\pi_a(\cdot|s)}\min_{\pi_b(\cdot|s,\cdot)}\sum_a\pi_a(a|s)\sum_b\pi_b(b|s,a)\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s') \\ -\alpha\log\pi_a(a|s)+\beta\log\pi_b(b|s,a)). \end{aligned}$$

$$\pi_{b,V}(b|s,a)=\frac{\exp\left\{-\frac{1}{\beta}\sum_{s'}p(s'|s,a,b)(r(s,a,b,s')+\gamma V(s'))\right\}}{\sum_{b'}\exp\left\{-\frac{1}{\beta}\sum_{s'}p(s'|s,a,b')(r(s,a,b',s')+\gamma V(s'))\right\}}$$

$$
\pi_{a,V}(a|s) = \frac{\exp\{\frac{1}{\alpha}\sum_b \pi_{b,V}(b|s,a)\sum_{s'} p(s'|s,a,b)(r(s,a,b,s') + \gamma V(s'))\}}{\sum_{a'} \exp\{\frac{1}{\alpha}\sum_b \pi_{b,V}(b|s,a')\sum_{s'} p(s'|s,a',b)(r(s,a',b,s') + \gamma V(s'))\}}
$$

$$
\begin{aligned}
Q_V(s,\cdot,\cdot) &= \sum_{s'} p(s'|s,a,b)(r(s,a,b,s') + \gamma V(s')) \\
&= \begin{bmatrix} Q_V(s,a_1,b_1) & Q_V(s,a_1,b_2) & \cdots & Q_V(s,a_1,b_n) \\ Q_V(s,a_2,b_1) & Q_V(s,a_2,b_2) & \cdots & Q_V(s,a_2,b_n) \\ \vdots & \vdots & \ddots & \vdots \\ Q_V(s,a_m,b_1) & Q_V(s,a_m,b_2) & \cdots & Q_V(s,a_m,b_n) \end{bmatrix}
\end{aligned}
$$

$$
\pi_{b,V}(v|s,a) = \frac{\exp\left\{-\frac{1}{\beta}Q_V(s,a,b)\right\}}{\sum_{b'} \exp\left\{-\frac{1}{\beta}Q_V(s,a,b')\right\}}
$$

$$
\pi_{a, V}(a \vert s) = \frac{\exp\left\{\frac{1}{\alpha} \sum_b \pi_{b,V} (b \vert s, a) Q_V(s, a, b)\right\}}{\sum_{a'}\exp\left\{\frac{1}{\alpha} \sum_b \pi_{b,V} (b \vert s, a') Q_V(s, a, b')\right\}}
$$

$$
\pi_{a,V}(a|s) = \frac{\exp\{\frac{1}{\alpha}\sum_b \pi_{b,V}(b|s,a)Q_V(s,a,b)\}}{\sum_{a'} \exp\{\frac{1}{\alpha}\sum_b \pi_{b,V}(b|s,a')Q_V(s,a,b')\}}
$$

$$
\pi_{a,V}(a|s) = \frac{\exp\{\frac{1}{\alpha}\sum_b \pi_{b,V}(b|s,a)Q_V(s,a,b)\}}{\sum_{a'} \exp\{\frac{1}{\alpha}\sum_b \pi_{b,V}(b|s,a')Q_V(s,a,b')\}}
$$