RL Objective

1. BELLMAN EQUATION

Because Bellman equation is

$$T_\pi V(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$$

and

$$V_\pi = T_\pi V_\pi.$$

The target of reinforcement learning by using Bellman equation is

$$\begin{cases} \max_\pi \sum_s p_1(s) V(s) \\ \min_V \sum_s p_2(s) \{\sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] - V(s)\}^2 \\ V(s) = \sum_a \pi(a|s) Q(s,a) \end{cases}$$

1.1 Tabular Algorithm

$$\begin{cases} l_1(\pi, V) = -\sum_s p_1(s) V(s) \\ l_2(\pi, V) = \sum_s p_2(s) \{\sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] - V(s)\}^2 \end{cases}$$

Or

$$\begin{cases} l_1(\pi, Q) & = -\sum_s p_1(s) \sum_a \pi(a|s) Q(s,a) \\ l_2(\pi, Q) & = \sum_s p_2(s) \{\sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \sum_{a'} \pi(a'|s') Q(s',a')] - \sum_a \pi(a|s) Q(s,a)\}^2 \\ & = \sum_s p_2(s) \{\sum_a \pi(a|s) \left[\sum_{s'} p(s'|s,a)[r(s,a,s') + \sum_{a'} \pi(a'|s') Q(s',a')] - Q(s,a)\right]\}^2 \end{cases}$$

$$\begin{aligned} \frac{\partial}{\partial \pi(s,a)} l_1(\pi, Q) &= -p_1(s) Q(s,a) = -\mathbb{E}_{p_1,\pi}\left[\frac{Q(s,a)}{\pi(a|s)}\right] \\ \frac{\partial}{\partial Q(s,a)} l_1(\pi, Q) &= -p_1(s) \pi(s,a) \\ \frac{\partial}{\partial \pi(a''|s'')} l_2(\pi, Q) =& 2 \sum_s p_2(s) \left\{\sum_a \pi(a|s) \left[\sum_{s'} p(s'|s,a)[r(s,a,s') + \sum_{a'} \pi(a'|s') Q(s',a')] - Q(s,a)\right]\right\} \\ & \cdot \left\{\sum_a \pi(a|s) p(s''|s,a) Q(s'',a'')\right. \\ & \left. - 1\{s = s''\} \left[\sum_{s'} p(s'|s'',a'')[r(s'',a'',s') + \sum_{a'} \pi(a'|s') Q(s',a')] - Q(s'',a'')\right]\right\} \\ \frac{\partial}{\partial Q(s'',a'')} l_2(\pi, Q) =& 2 \sum_s p_2(s) \left\{\sum_a \pi(a|s) \left[\sum_s p(s'|s,a)[r(s,a,s') + \sum_{a'} \pi(a'|s') Q(s',a')] - Q(s,a)\right]\right\} \\ & \cdot \left\{\sum_a \pi(a|s) p(s''|s,a) \pi(a''|s'') - 1\{s = s''\} \pi(a''|s'')\right\} \end{aligned}$$

1.2 Approximation Algorithm

$$\begin{cases} \theta_\pi = \arg\max_{\theta_\pi} \sum_s p_1(s) \sum_a \pi(a|s;\theta_\pi) Q(s,a|\theta_Q) \\ \theta_Q = \arg\min_{\theta_Q} \sum_s p_2(s) \{\sum_a \pi(a|s;\theta_\pi) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s';\theta_\pi,\theta_Q)] - V(s;\theta_\pi,\theta_Q)\}^2 \\ V(s;\theta_\pi,\theta_Q) = \sum_a \pi(a|s;\theta_\pi) Q(s,a|\theta_Q) \end{cases}$$

$$\begin{cases} l_1 = -\sum_s p_1(s) \sum_a \pi(a|s;\theta_\pi) Q(s,a|\theta_Q) \\ l_2 = \sum_s p_2(s) \{\sum_a \pi(a|s;\theta_\pi) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s';\theta_\pi,\theta_Q)] - V(s;\theta_\pi,\theta_Q)\}^2 \\ V(s;\theta_\pi,\theta_Q) = \sum_a \pi(a|s;\theta_\pi) Q(s,a|\theta_Q) \end{cases}$$

2. OPTIMAL BELLMAN EQUATION

Because optimal Bellman equation is

$$TV(s) = \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')], \forall V \in \mathbb{R}^{|S|}$$

therefore the target becomes

$$\min_V \sum_s p_2(s) \{TV(s) - V(s)\}^2.$$

The target of reinforcement learning by using Optimal equation is

$$\min_V \sum_s p_2(s) \left\{ \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] - V(s) \right\}^2$$

We make further exploration:

$$\begin{aligned} &\min_V \sum_s p(s) \left\{ \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] - V(s) \right\}^2 \\ =&\min_Q \sum_s p(s) \left\{ \max_{\pi} \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_{a'} \pi(a'|s') Q(s',a') \right] \right. \\ &\qquad \left. - \sum_a \pi(a|s) Q(s,a) \right\}^2 \\ =&\min_{\theta_Q} \sum_s p(s) \left\{ \max_{\theta_\pi} \sum_a \pi(a|s;\theta_\pi) \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_{a'} \pi(a'|s';\theta_\pi) Q(s',a';\theta_Q) \right] \right. \\ &\qquad \left. - \sum_a \pi(a|s;\theta_\pi) Q(s,a;\theta_Q) \right\}^2 \end{aligned}$$

2.1 V-Based-Loss Function

$$L(V) = \sum_s p(s) \left\{ \max_a \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')] - V(s) \right\}^2$$

2.2 Q-Based-Loss Function

2.2.1 On-policy

Let $\pi_Q(a|s) = 1\{a = \arg\max_{a'} Q(s,a')\}$:

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

(Hint: from smoothed Bellman equation, we have $\pi(a|s) = \lim_{\lambda \to 0} \pi_\lambda(a|s)$.)

2.2.2 Q-Learning

$$
\pi_{Q,\epsilon} = (1-\epsilon)\pi_Q + \epsilon \pi_{uniform}
$$

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

2.2.3 SARSA

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

2.2.3 Q-Learning with Replay Buffer

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

2.3 Q-Loss with Function Approximation

2.3.1 On-policy with function approximation

Let $\pi_Q(a|s;\theta_Q) = 1\{a = \arg\max_{a'} Q(s,a';\theta_Q)\}$

$$
\begin{aligned}
L(\theta_Q) &= \sum_s p(s) \sum_a \pi_Q(a|s;\theta_Q) \\
& \quad \cdot \left\{ \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_a \pi_Q(a'|s';\theta_Q) Q(s',a';\theta_Q) \right] - Q(s,a;\theta_Q) \right\}^2, \\
\nabla_{\theta_Q} L(\theta_Q) &= \sum_s p(s) \sum_a \pi_Q(a|s;\theta_Q) \\
& \quad \cdot \left\{ \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_a \pi_Q(a'|s';\theta_Q) Q(s',a';\theta_Q) \right] - Q(s,a;\theta_Q) \right\} \\
& \quad \cdot \left\{ \gamma \sum_{s'} p(s'|s,a) \nabla_{\theta_Q} \sum_a \pi_Q(a'|s';\theta_Q) Q(s',a';\theta_Q) - \nabla_{\theta_Q} Q(s,a;\theta_Q) \right\}
\end{aligned}
$$

2.3.2 Q-learning with function approximation

$$
\begin{aligned}
L(\theta_Q) = &\sum_s p(s) \sum_a \pi_{Q,\epsilon}(a|s;\theta_Q) \\
&\cdot \left\{ \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_a \pi_Q(a|s;\theta_Q) Q(s',a';\theta_Q) \right] - Q(s,a;\theta_Q) \right\}^2 \\
\nabla_{\theta_Q} L(\theta_Q) = &\sum_s p(s) \sum_a \pi_{Q,\epsilon}(a|s;\theta_Q) \\
&\cdot \left\{ \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_a \pi_Q(a'|s';\theta_Q) Q(s',a';\theta_Q) \right] - Q(s,a;\theta_Q) \right\} \\
&\cdot \left\{ \gamma \sum_{s'} p(s'|s,a) \nabla_{\theta_Q} \sum_a \pi_Q(a'|s';\theta_Q) Q(s',a';\theta_Q) - \nabla_{\theta_Q} Q(s,a;\theta_Q) \right\}
\end{aligned}
$$

2.3.3 SARSA with function approximation

$$
\begin{aligned}
L(\theta_Q) = &\sum_s p(s) \sum_a \pi_{Q,\epsilon}(a|s;\theta_Q) \\
&\cdot \left\{ \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_{a'} \pi_{Q,\epsilon}(a'|s';\theta_Q) Q(s',a';\theta_Q) \right] - Q(s,a;\theta_Q) \right\}^2 \\
\nabla_{\theta_Q} L(\theta_Q) = &\sum_s p(s) \sum_a \pi_{Q,\epsilon}(a|s;\theta_Q) \\
&\cdot \left\{ \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \sum_a \pi_{Q,\epsilon}(a'|s';\theta_Q) Q(s',a';\theta_Q) \right] - Q(s,a;\theta_Q) \right\} \\
&\cdot \left\{ \gamma \sum_{s'} p(s'|s,a) \nabla_{\theta_Q} \sum_a \pi_{Q,\epsilon}(a'|s';\theta_Q) Q(s',a';\theta_Q) - \nabla_{\theta_Q} Q(s,a;\theta_Q) \right\}
\end{aligned}
$$

2.3.4 DQN

The loss of DQN:

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

The derivative of the loss:

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

The update rule of DQN:

$$
\begin{cases}
\theta_Q = \theta_Q - \alpha_1 \nabla_{\theta_Q} L_Q(\theta_Q, \theta_{target}) \\
\theta_{Q_{target}} = \theta_{Q_{target}} - \alpha_2 (\theta_{Q_{target}} - \theta_Q) (\textbf{polyak averaging})
\end{cases}
$$

3. SMOOTHED BELLMAN EQUATION

3.1 Preliminaries

3.1.1 Normal Reinforcement Learning Target

$$\max_\pi \sum_{s_0} p_0(s_0) \sum_{a_0} \pi(a_0|s_0) \sum_{s_1} p(s_1|s_0,a_0) \Bigg(r(s_0,a_0,s_1) + \sum_{a_1} \pi(a_1|s_1) \sum_{s_2} p(s_2|s_1,a_1)\gamma \Bigg(r(s_1,a_1,s_2) + + \sum_{a_2} \pi(a_2|s_2) \sum_{s_3} p(s_3|s_2,a_2)\gamma^2 (r(s_2,a_2,s_3) + \cdots) \Bigg)\Bigg)$$

3.1.2 Regularization Based Reinforcement Learning Target

$$\max_\pi \sum_{s_0} p_0(s_0) \sum_{a_0} \pi(a_0|s_0) \sum_{s_1} p(s_1|s_0,a_0) \Bigg(r(s_0,a_0,s_1) + \mathcal{H}(\pi(a_0|s_0)) + \sum_{a_1} \pi(a_1|s_1) \sum_{s_2} p(s_2|s_1,a_1)\gamma \Bigg(r(s_1,a_1,s_2) + \mathcal{H}(\pi(a_1|s_1)) + \sum_{a_2} \pi(a_2|s_2) \sum_{s_3} p(s_3|s_2,a_2)\gamma^2 (\cdots) \Bigg)\Bigg)$$

3.2 Policy Based

For all $\pi \in \Pi$, we have $J(\pi)$, $Q^\pi_{soft}(s_0,a_0)$ and $V^\pi_{soft}(s_0)$ defined below:

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

$$Q^\pi_{soft}(s_0,a_0) = \sum_{s_1} p(s_1|s_0,a_0) \Bigg(r(s_0,a_0,s_1) + \mathcal{H}(\pi(a_0|s_0)) + \sum_{a_1} \pi(a_1|s_1) \sum_{s_2} p(s_2|s_1,a_1)\gamma \Bigg(r(s_1,a_1,s_2) + \mathcal{H}(\pi(a_1|s_1)) + \sum_{a_2} \pi(a_2|s_2) \sum_{s_3} p(s_3|s_2,a_2)\gamma^2 (\cdots) \Bigg)\Bigg)$$

$$V^\pi_{soft}(s_0) = \sum_{a_0} \pi(a_0|s_0) \sum_{s_1} p(s_1|s_0,a_0) \Bigg(r(s_0,a_0,s_1) + \mathcal{H}(\pi(a_0|s_0)) + \sum_{a_1} \pi(a_1|s_1) \sum_{s_2} p(s_2|s_1,a_1)\gamma \Bigg(r(s_1,a_1,s_2) + \mathcal{H}(\pi(a_1|s_1)) + \sum_{a_2} \pi(a_2|s_2) \sum_{s_3} p(s_3|s_2,a_2)\gamma^2 (\cdots) \Bigg)\Bigg)$$

> **Lemma**:
>
> $$J(\pi) = \sum_{s_0} p_0(s_0) V^\pi_{soft}(s_0) = \sum_{s_0} p_0(s_0) \sum_{a_0} \pi(a_0|s_0) Q^\pi_{soft}(s_0).$$

3.3 Value Based

For all $V \in \mathbb{R}^{|S|}$, we have the following things.

Definition Smoothed Bellman equation: $\forall V, \pi$:

$$T^{\pi}_{soft} V(s) = \sum_a \pi(a|s) \left(\sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma V(s') \right] \right) + H(\pi(\cdot|s)).$$

Definition Smoothed optimal Bellman equation:

$$T_{soft} V(s) = \max_{\pi(\cdot|s)} T^{\pi}_{soft} V(s) = \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \left(\sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma V(s') \right] \right) + H(\pi(\cdot|s)).$$

If $H(\pi(\cdot|s)) = -\lambda \sum_a \pi(a|s) \log(\pi(a|s))$, then

$$T_{soft} V(s) = \lambda \log \left\{ \sum_a \exp \left[\frac{1}{\lambda} \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s')) \right] \right\}$$

$$\begin{aligned} &\pi_{V,soft}(a|s) = \arg\max_{\pi} T^{\pi} V(s) \\ =& \frac{\exp\{\frac{1}{\lambda} \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma V(s') \right]\}}{\sum_{a'} \exp\{\frac{1}{\lambda} \sum_{s'} p(s'|s,a') \left[r(s,a',s') + \gamma V(s') \right]\}} = \frac{\exp\{\frac{1}{\lambda} Q_V(s,a)\}}{\sum_{a'} \exp\{\frac{1}{\lambda} Q_V(s,a')\}} \end{aligned}$$

where we define $Q_V(s,a) = \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma V(s') \right]$.

We construct the target:

$$\begin{aligned} &\min_V \sum_s p(s) \{ T_\lambda V(s) - V(s) \}^2 \\ =& \min_V \sum_s p(s) \left\{ \lambda \log \left\{ \sum_a \exp \left[\frac{1}{\lambda} \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s')) \right] \right\} - V(s) \right\}^2 \end{aligned}$$

We have the loss:

$$L(V) = \sum_s p(s) \left\{ \lambda \log \left\{ \sum_a \exp \left[\frac{1}{\lambda} \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s')) \right] \right\} - V(s) \right\}^2.$$

$$L(\theta_V) = \sum_s p(s) \left\{ \lambda \log \left\{ \sum_a \exp \left[\frac{1}{\lambda} \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s';\theta_V)) \right] \right\} - V(s;\theta_V) \right\}^2$$

3.3 Q-Loss Function

$$\begin{aligned} L(Q) = \sum_s p(s) \Bigg\{ \lambda \log \Bigg\{ \sum_a \exp \Bigg[\frac{1}{\lambda} \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma \frac{\sum_{a'} Q(s',a') \exp\{\frac{1}{\lambda} Q(s',a')\}}{\sum_{a'} \exp\{\frac{1}{\lambda} Q(s',a')\}}) \Bigg] \Bigg\} \\ - \frac{\sum_{a'} Q(s,a') \exp\{\frac{1}{\lambda} Q(s,a')\}}{\sum_{a'} \exp\{\frac{1}{\lambda} Q(s,a')\}} \Bigg\}^2, \quad where \quad \pi_\lambda(a|s) = \frac{\exp\{\frac{1}{\lambda} Q(s,a)\}}{\sum_{a'} \exp\{\frac{1}{\lambda} Q(s,a')\}} \end{aligned}$$

$$\begin{aligned} L(\tilde{Q}) = \sum_s p(s) \Bigg\{ \lambda \log \Bigg\{ \sum_a \exp \Bigg[\frac{1}{\lambda} \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma\lambda \sum_{a'} \pi(a'|s') \log(\tilde{Q}(s',a')) \Bigg] \Bigg\} \\ - \lambda \sum_a \pi(a|s) \log(\tilde{Q}(s,a)) \Bigg\}^2, \qquad \pi(a|s) = \frac{\tilde{Q}(s,a)}{\sum_a \tilde{Q}(s,a)} \end{aligned}$$

3.4 Q-Loss Function with Function Approximation

$$L(\theta) = \sum_s p(s)\left\{\lambda \log\left\{\sum_a \exp\left[\frac{1}{\lambda}\sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma\frac{\sum_{a'} Q(s',a';\theta)\exp\{\frac{1}{\lambda}Q(s',a';\theta)\}}{\sum_{a'}\exp\{\frac{1}{\lambda}Q(s',a';\theta)\}})\right]\right\}\right.$$

$$\left. - \frac{\sum_{a'} Q(s,a';\theta)\exp\{\frac{1}{\lambda}Q(s,a';\theta)\}}{\sum_{a'}\exp\{\frac{1}{\lambda}Q(s,a';\theta)\}}\right\}^2, \quad where \quad \pi_\lambda(a|s;\theta) = \frac{\exp\{\frac{1}{\lambda}Q(s,a;\theta)\}}{\sum_{a'}\exp\{\frac{1}{\lambda}Q(s,a';\theta)\}}$$

This is a little difficult to take the derivative.

$$L(\theta) = \sum_s p(s)\left\{\lambda \log\left\{\sum_a \exp\left[\frac{1}{\lambda}\sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma\lambda\sum_{a'}\pi(a'|s';\theta)\log\tilde{Q}(s',a';\theta))\right]\right\}\right.$$

$$\left. - \lambda\sum_a \pi(a|s;\theta)\log\tilde{Q}(s,a;\theta))\right\}^2, \quad \pi(a|s;\theta) = \frac{\tilde{Q}(s,a;\theta)}{\sum_a \tilde{Q}(s,a;\theta)}$$

3.5 SBEED Loss Function

$$\begin{aligned}
&\min_V \sum_s p(s)\{T_\lambda V(s) - V(s)\}^2 \\
&= \min_V \sum_s p(s)\left\{\lambda\log\left\{\sum_a \exp\left[\frac{1}{\lambda}\sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s'))\right]\right\} - V(s)\right\}^2 \\
&?\min_V \sum_s p(s)\sum_a \pi_\lambda(a|s)\left\{\lambda\log\left\{\frac{\exp\left[\frac{1}{\lambda}\sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s'))\right]}{\pi_\lambda(a|s)}\right\} - V(s)\right\}^2 \\
&= \min_V \sum_s p(s)\sum_a \pi_\lambda(a|s)\left\{\sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s')) - \lambda\log\pi_\lambda(a|s) - V(s))\right\}^2
\end{aligned}$$

$$\min_{V,\pi} l(V,\pi) = \mathbb{E}_{s,a}\left\{\mathbb{E}_{s'|s,a}[r(s,a,s') + \gamma V(s')] - \lambda\log(\pi(a|s)) - V(s)\right\}^2$$

$$\pi_\lambda(a|s;\theta) = \frac{\exp\{\frac{1}{\lambda}Q(s,a;\theta)\}}{\sum_{a'}\exp\{\frac{1}{\lambda}Q(s,a';\theta)\}}$$

$$V(s;\theta) = \sum_a Q(s,a;\theta)\pi_\lambda(a|s;\theta)$$

$$L(\theta) = \mathbb{E}_{s,a}\left\{\mathbb{E}_{s'|s,a}[r(s,a,s') + \gamma V(s';\theta)] - \lambda\log(\pi(a|s;\theta)) - V(s;\theta)\right\}^2$$

$$\begin{aligned}
\min_\theta L(\theta) &= \min_\theta \mathbb{E}_{s,a}\left\{\mathbb{E}_{s'|s,a}[r(s,a,s') + \gamma V(s';\theta)] - \lambda\log(\pi(a|s;\theta)) - V(s;\theta)\right\}^2 \\
&= \min_\theta \mathbb{E}_{s,a}\left\{\mathbb{E}_{s'|s,a}[\delta(s,a,s';\theta) - V(s;\theta)]\right\}^2, \\
&\quad \delta(s,a,s';\theta) = r(s,a,s') + \gamma V(s';\theta) - \lambda\log(\pi(a|s;\theta))
\end{aligned}$$

$$\begin{aligned}
\min_\theta L(\theta) &= \min_\theta\max_w 2\mathbb{E}_{s,a,s'}\{\nu(s,a;w)[\delta(s,a,s';\theta) - V(s;\theta)]\} - \mathbb{E}_{s,a,s'}\{\nu^2(s,a;w)\} \\
&= \min_\theta\max_w \mathbb{E}_{s,a,s'}\{\delta(s,a,s';\theta) - V(s,a;\theta)\}^2 - \mathbb{E}_{s,a,s'}\{\delta(s,a,s';\theta) - \nu(s,a;w)\}^2
\end{aligned}$$

$$\nabla_\theta L(\theta) = 2\nu(s,a;w)[\gamma\nabla_\theta V(s';\theta) - \lambda\nabla_\theta\log(\pi(a|s;\theta)) - \nabla_\theta V(s;\theta)]$$

Appendix: Soft Optimal Bellman equation

Smoothed Bellman equation:

$$T_\lambda V(s) = \max_{\pi(\cdot|s)}\sum_a \pi(a|s)\left(\sum_{s'} p(s'|s,a)\left[r(s,a,s') + \gamma V(s')\right]\right) + \lambda H(\pi(\cdot|s)).$$

If $H(\pi(\cdot|s)) = -\sum_a \pi(a|s)\log(\pi(a|s))$, then

$$T_\lambda V(s) = \lambda \log \left\{ \sum_a \exp \left[\frac{1}{\lambda} \sum_{s'} p(s'|s,a)(r(s,a,s') + \gamma V(s')) \right] \right\}$$

proof:

$$\max_{\pi(\cdot|s) \in \Pi(s)} \sum_a \pi(a|s) \left(\sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma V(s') \right] - \lambda \log \pi(a|s) \right),$$
$$s.t. \sum_a \pi(a|s) = 1.$$

$$\begin{aligned}
&\max_{\pi(\cdot|s) \succeq 0} \min_{k_s \neq 0} \sum_a \pi(a|s) \left(Q(s,a) - \lambda \log \pi(a|s) \right) + k_s \left(1 - \sum_a \pi(a|s) \right) \\
\leq &\min_{k_s \neq 0} \max_{\pi(\cdot|s) \succeq 0} \sum_a \pi(a|s) \left(Q(s,a) - \lambda \log \pi(a|s) \right) + k_s \left(1 - \sum_a \pi(a|s) \right)
\end{aligned}$$

We solve the dual problem:

$$\begin{aligned}
&Q(s,a) - \lambda(1 + \log \pi(a|s)) - k_s = 0 \\
\Rightarrow &\pi(a|s) exp(1 + k_s/\lambda) = \exp \left\{ \frac{1}{\lambda} Q(s,a) \right\} \\
\Rightarrow &exp(1 + k_s/\lambda) = \sum_a \exp \left\{ \frac{1}{\lambda} Q(s,a) \right\} \\
\Rightarrow &1 + k_s/\lambda = \log \left\{ \sum_a \exp \left[\frac{1}{\lambda} Q(s,a) \right] \right\} \\
\Rightarrow &\pi(a|s) = \frac{\exp\{\frac{1}{\lambda} Q(s,a)\}}{\sum_a \exp\{\frac{1}{\lambda} Q(s,a)\}}
\end{aligned}$$

$$\sum_a \pi(a|s) \left[Q(s,a) - \lambda(1 + \log \pi(a|s)) - k_s \right] = 0$$

$$\begin{aligned}
\Rightarrow &\min_{k_s \neq 0} \max_{\pi(\cdot|s) \succeq 0} \sum_a \pi(a|s) \left(Q(s,a) - \lambda \log \pi(a|s) \right) + k_s \left(1 - \sum_a \pi(a|s) \right) \\
&= k_s + \lambda \sum_a \pi(a|s) = k_s + \lambda = \lambda \log \left\{ \sum_a \exp \left[\frac{1}{\lambda} Q(s,a) \right] \right\}
\end{aligned}$$