Soft Policies

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\,\cdot\,|\mathbf{s}_t)) \right].$$

How does the Bellman equation change?

Soft Policies

Lemma 1 (Soft Policy Evaluation)**.** *Consider the soft Bellman backup operator $\mathcal{T}^\pi$ in Equation 2 and a mapping $Q^0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $|\mathcal{A}| < \infty$, and define $Q^{k+1} = \mathcal{T}^\pi Q^k$. Then the sequence Q^k will converge to the soft Q-value of π as $k \to \infty$.*

Proof

Lemma 1 (Soft Policy Evaluation). *Consider the soft Bellman backup operator $\mathcal{T}^\pi$ in Equation 2 and a mapping $Q^0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $|\mathcal{A}| < \infty$, and define $Q^{k+1} = \mathcal{T}^\pi Q^k$. Then the sequence Q^k will converge to the soft Q-value of π as $k \to \infty$.*

Proof. Define the entropy augmented reward as $r_\pi(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[\mathcal{H}\left(\pi(\,\cdot\,|\mathbf{s}_{t+1})\right)\right]$ and rewrite the update rule as

$$Q(\mathbf{s}_t, \mathbf{a}_t) \leftarrow r_\pi(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1} \sim p, \mathbf{a}_{t+1} \sim \pi}\left[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})\right] \tag{15}$$

and apply the standard convergence results for policy evaluation (Sutton & Barto, 1998). The assumption $|\mathcal{A}| < \infty$ is required to guarantee that the entropy augmented reward is bounded. $\square$

Optimal Soft Policy

- The optimal soft policy (optimizing entropy augment objective) is:

$$\pi^*(a|s) = \frac{\exp Q(s,a)}{\sum_{a'} \exp Q(s,a')}$$

Soft actor-critic

1. Q-function update

Update Q-function to evaluate current policy:

$$Q(\mathbf{s},\mathbf{a}) \leftarrow r(\mathbf{s},\mathbf{a}) + \mathbb{E}_{\mathbf{s}' \sim p_{\mathbf{s}},\ \mathbf{a}' \sim \pi}\left[Q(\mathbf{s}',\mathbf{a}') - \log \pi(\mathbf{a}'|\mathbf{s}')\right]$$

This converges to Q^{π}.

2. Update policy

Update the policy with gradient of information projection:

$$\pi_{\text{new}} = \arg\min_{\pi'} \mathrm{D}_{\mathrm{KL}}\left(\pi'(\,\cdot\,|\mathbf{s}) \,\middle\|\, \frac{1}{Z}\exp Q^{\pi_{\text{old}}}(\mathbf{s},\,\cdot\,)\right)$$

In practice, only take one gradient step on this objective

3. Interact with the world, collect more data

Haarnoja, et al. **Soft Actor-Critic Algorithms and Applications**. ‘18

Soft Policy Improvement

Lemma 2 (Soft Policy Improvement). *Let $\pi_{\text{old}} \in \Pi$ and let π_{new} be the optimizer of the minimization problem defined in Equation 4. Then $Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t)$ for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$ with $|\mathcal{A}| < \infty$.*

Proof. Let $\pi_{\text{old}} \in \Pi$ and let $Q^{\pi_{\text{old}}}$ and $V^{\pi_{\text{old}}}$ be the corresponding soft state-action value and soft state value, and let π_{new} be defined as

$$\begin{aligned} \pi_{\text{new}}(\,\cdot\,|\mathbf{s}_t) &= \arg\min_{\pi' \in \Pi} \mathrm{D_{KL}}\left(\pi'(\,\cdot\,|\mathbf{s}_t) \;\middle\|\; \exp\left(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \,\cdot\,) - \log Z^{\pi_{\text{old}}}(\mathbf{s}_t)\right)\right) \\ &= \arg\min_{\pi' \in \Pi} J_{\pi_{\text{old}}}(\pi'(\,\cdot\,|\mathbf{s}_t)). \end{aligned} \tag{16}$$

It must be the case that $J_{\pi_{\text{old}}}(\pi_{\text{new}}(\,\cdot\,|\mathbf{s}_t)) \leq J_{\pi_{\text{old}}}(\pi_{\text{old}}(\,\cdot\,|\mathbf{s}_t))$, since we can always choose $\pi_{\text{new}} = \pi_{\text{old}} \in \Pi$. Hence

$$\mathbb{E}_{\mathbf{a}_t \sim \pi_{\text{new}}}\left[\log \pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t) - Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + \log Z^{\pi_{\text{old}}}(\mathbf{s}_t)\right] \leq \mathbb{E}_{\mathbf{a}_t \sim \pi_{\text{old}}}\left[\log \pi_{\text{old}}(\mathbf{a}_t|\mathbf{s}_t) - Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + \log Z^{\pi_{\text{old}}}(\mathbf{s}_t)\right], \tag{17}$$

Soft Policy Improvement

and since partition function $Z^{\pi_{\text{old}}}$ depends only on the state, the inequality reduces to

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_t,\mathbf{a}_t)-\log\pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t)\right]\geq V^{\pi_{\text{old}}}(\mathbf{s}_t). \tag{18}$$

Next, consider the soft Bellman equation:

$$\begin{aligned}
Q^{\pi_{\text{old}}}(\mathbf{s}_t,\mathbf{a}_t) &= r(\mathbf{s}_t,\mathbf{a}_t)+\gamma\,\mathbb{E}_{\mathbf{s}_{t+1}\sim p}\left[V^{\pi_{\text{old}}}(\mathbf{s}_{t+1})\right]\\
&\leq r(\mathbf{s}_t,\mathbf{a}_t)+\gamma\,\mathbb{E}_{\mathbf{s}_{t+1}\sim p}\left[\mathbb{E}_{\mathbf{a}_{t+1}\sim\pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_{t+1},\mathbf{a}_{t+1})-\log\pi_{\text{new}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})\right]\right]\\
&\;\;\vdots\\
&\leq Q^{\pi_{\text{new}}}(\mathbf{s}_t,\mathbf{a}_t),
\end{aligned} \tag{19}$$

where we have repeatedly expanded $Q^{\pi_{\text{old}}}$ on the RHS by applying the soft Bellman equation and the bound in Equation 18. Convergence to $Q^{\pi_{\text{new}}}$ follows from Lemma 1. □

Soft actor-critic

Algorithm 1 Soft Actor-Critic

Inputs: The learning rates, λ_π, λ_Q, and λ_V for functions π_θ, Q_w, and V_ψ respectively; the weighting factor τ for exponential moving average.

```
Initialize parameters θ, w, ψ, and ψ̄.
for each iteration do
    (In practice, a combination of a single environment step and multiple
    gradient steps is found to work best.)
    for each environment setup do
        a_t ~ π_θ(a_t|s_t)
        s_{t+1} ~ ρ_π(s_{t+1}|s_t, a_t)
        D ← D ∪ {(s_t, a_t, r(s_t, a_t), s_{t+1}}
    for each gradient update step do
        ψ ← ψ − λ_V ∇_ψ J_V(ψ).
        w ← w − λ_Q ∇_w J_Q(w).
        θ ← θ − λ_π ∇_θ J_π(θ).
        ψ̄ ← τψ + (1 − τ)ψ̄).
```