The Point to Which Soft Actor-Critic Converges

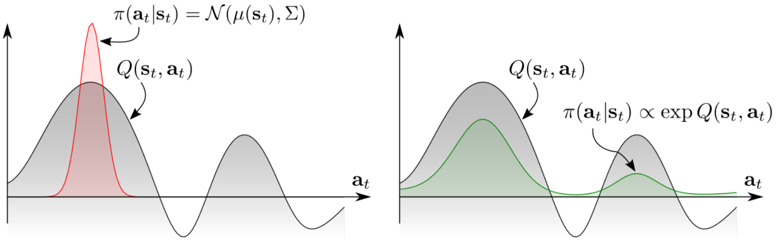
Jianfei Ma

May 5, 2023

Maximum Entropy Reinforcement Learning

Unlike the standard RL formulation, Maximum entropy RL seeks for higher reward region while takes relative importance of the policy entropy into consideration.

$$\pi^*_{\mathrm{MaxEnt}} = \operatorname*{argmax}_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T} r_t + \mathcal{H}(\pi(\cdot|\mathbf{s}_t)) \right] \tag{1}$$



(a) Unimodal Policy (b) Multimodal Policy

Figure: A multimodal Q-function[1]

Algorithms under the Framework

Soft Q-learning:

$$Q(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}\left[r_t + \gamma \ \mathrm{softmax}_{\mathbf{a}} Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})\right]$$

where

$$\mathrm{softmax}_{\mathbf{a}} f(\mathbf{a}) := \log \int \exp f(\mathbf{a}) \, d\mathbf{a}$$

Soft actor-critic:

$$Q(\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}[r_t + \gamma \mathbb{E}_{\mathbf{s}_{t+1}}[V(\mathbf{s}_{t+1})]]$$

where

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi}[Q(\mathbf{s}_t, \mathbf{a}_t) - \eta \log \pi(\mathbf{a}_t | \mathbf{s}_t)]$$

Question

If we repeatedly improve the action-value function, and based on which improve the policy, do SQL and SAC have the same limiting point?

Solution Concept

Define the regularized state-value function as

$$\tilde{V}^{\pi}(s) = \mathbb{E}\Big[\sum_{l=0}^{\infty} \gamma^{l}(r_{t+l} + \eta\Delta_{t+l})|s_0 = s\Big] \tag{2}$$

where η is the temperature parameter, usually positive, determining the relative importance of the regularization term against the reward.
The optimal regularized value function $\tilde{V}^{\star}(s)$ should satisfy the corresponding optimal Bellman equation for all $s \in \mathcal{S}$

$$\tilde{V}^{\star}(s) = \sup_{\pi} \sum_{a \in \mathcal{A}} \pi(a|s)\big[r(s,a) + \eta\Delta(s) + \gamma\mathbb{E}_{s' \sim p}[\tilde{V}^{\star}(s')]\big] \tag{3}$$

Solution Concept

From an optimization perspective, we can transfer the problem into a constraint optimization problem

$$\begin{aligned}\max_{\pi} \mathcal{J}(\pi) &= \sum_{a \in \mathcal{A}} \pi(a|s)\left[r(s,a) + \eta\Delta(s) + \gamma\mathbb{E}_{s' \sim p}[\tilde{V}^{\star}(s')]\right] \\ \text{s.t.} \sum_{a \in \mathcal{A}} \pi(a|s) &= 1\end{aligned} \tag{4}$$

If $\Delta(s) = \mathcal{H}(\pi(\cdot|s))$, we can write out the Lagrangian

$$\mathcal{L}(s;\lambda) = \sum_{a \in \mathcal{A}} \pi(a|s)\left[r(s,a) + \gamma\mathbb{E}_{s' \sim p}[\tilde{V}^{\star}(s')]\right] + \eta\mathcal{H}(s) - \lambda(\sum_{a \in \mathcal{A}} \pi(a|s) - 1) \tag{5}$$

- Objective is linear
- $\mathcal{H}$ is strictly-concave
- Slater condition is satisfied

Solution Concept

Theorem (Optimality)

For all $s \in \mathcal{S}$, the optimal value function $\tilde{V}^\star(s)$ and the optimal policy $\tilde{\pi}^\star(a|s)$, satisfy

$$\begin{aligned}
\tilde{V}^\star(s) &= \eta \log \sum_{a \in \mathcal{A}} \exp \frac{1}{\eta}\Big(r(s,a) + \gamma \mathbb{E}_{s' \sim p}[\tilde{V}^\star(s')]\Big) \\
\tilde{\pi}^\star(a|s) &= \frac{\exp \frac{1}{\eta}\Big(r(s,a) + \gamma \mathbb{E}_{s' \sim p}[\tilde{V}^\star(s')]\Big)}{\sum_{a \in \mathcal{A}} \exp \frac{1}{\eta}\Big(r(s,a) + \gamma \mathbb{E}_{s' \sim p}[\tilde{V}^\star(s')]\Big)}
\end{aligned} \tag{6}$$

Solution Concept

Auxiliary Soft Action-Value Function

$$\tilde{Q}^\star(s, a) \triangleq r(s, a) + \gamma \mathbb{E}_{s' \sim p}[\tilde{V}^\star(s')] \tag{7}$$

Proposition (An Inequality)

For any $V : \mathcal{S} \to \mathbb{R}$ that satisfies $V(s) \leq \tilde{V}^\star(s)$ for all $s \in \mathcal{S}$, then

$$Q(s, a) \triangleq r(s, a) + \gamma \mathbb{E}_{s' \sim p}[V(s')] \leq \tilde{Q}^\star(s, a) \tag{8}$$

Soft Policy Iteration

Soft Bellman Operator

$$\mathcal{T}^{\pi} Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V(s_{t+1})], \tag{9}$$

where

$$V(s_t) = \mathbb{E}_{a_t \sim \pi}[Q(s_t, a_t) - \eta \log \pi(a_t | s_t)] \tag{10}$$

Softmax Operator

$$\mathcal{G}(Q^{\pi}) = \frac{\exp \frac{1}{\eta}(Q^{\pi})}{\sum_{a \in \mathcal{A}} \exp \frac{1}{\eta}(Q^{\pi})} \tag{11}$$

Soft Policy Iteration

- **Soft policy evaluation**: $Q^{k+1} \leftarrow \mathcal{T}^{\pi} Q^{k}, \lim_{k \to \infty} Q^{k+1} = Q^{\pi}$
- **Soft policy improvement**: $\tilde{\pi} \leftarrow \mathcal{G}(Q^{\pi})$

Soft Policy Iteration

Repeated application of soft policy evaluation and soft policy improvement to any $\pi \in \Pi$ converges to a policy $\pi^{\star}$ such that $Q^{\pi^{\star}}(s, a) \geq Q^{\pi}(s, a)$ for all $\pi \in \Pi$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Main Result

Theorem (Convergent Points)

For any initial policy π_0 and corresponding action-value function Q^{π_0}, the convergent points induced by SPI satisfy $Q^{\pi^\star}(s,a) = \tilde{Q}^\star(s,a)$ and $\pi^\star = \tilde{\pi}^\star$.

Proof.

The backward direction is obvious since $\tilde{\pi}^\star \in \Pi$, that is, $Q^{\pi^\star} \geq \tilde{Q}^\star$. We only need to show the other direction. Since $Q^{\pi^\star}$ is the fixed point of the soft Bellman operator $\mathcal{T}^{\pi^\star}$, thus it must satisfy the soft Bellman equation with a value function $V^{\pi^\star}$. And since $\tilde{V}^\star$ is the regularized value function that at most can be obtained, it must have $V^{\pi^\star} \leq \tilde{V}^\star$. By the inequality Proposition, it follows that $Q^{\pi^\star} \leq \tilde{Q}^\star$. And since $\pi^\star \in \Pi$, it immediately follows that $\pi^\star = \tilde{\pi}^\star$. $\square$

Extendibility

If we are interested in constrain our policy w.r.t. some reference policy $\bar{\pi}$, we can set $\Delta(s) = -D_{\mathrm{KL}}(\pi \| \bar{\pi})$ (which is strictly-concave for fixed $\bar{\pi}$).

Conservative Optimal Points

$$V^{\pi^\star}(s) = \eta \log \sum_{a \in \mathcal{A}} \bar{\pi}(a|s) \exp \frac{1}{\eta}\left(r(s,a) + \gamma \mathbb{E}_{s' \sim p}[V^{\pi^\star}(s')]\right)$$
$$\pi^\star(a|s) = \frac{\bar{\pi}(a|s) \exp \frac{1}{\eta}\left(r(s,a) + \gamma \mathbb{E}_{s' \sim p}[V^{\pi^\star}(s')]\right)}{\sum_{a \in \mathcal{A}} \bar{\pi}(a|s) \exp \frac{1}{\eta}\left(r(s,a) + \gamma \mathbb{E}_{s' \sim p}[V^{\pi^\star}(s')]\right)} \tag{12}$$

Conservative Bellman Operator

$$\mathcal{T}^\pi Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V(s_{t+1})],$$
$$V(s_t) = \mathbb{E}_{a_t \sim \pi}[Q(s_t, a_t) - \eta \log \frac{\pi(a_t|s_t)}{\bar{\pi}(a_t|s_t)}] \tag{13}$$

Takeaways

- Translation from the arduous optimization of the LogSumExp to the repeated policy evaluation and improvement is appealing.
- A generalized type of the regularizer such as the KL divergence, can follow another optimization procedure.

References

Haarnoja, T., Tang, H., Abbeel, P. & Levine, S. Reinforcement Learning with Deep Energy-Based Policies. *Proceedings Of The 34th International Conference On Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. **70** pp. 1352-1361 (2017), http://proceedings.mlr.press/v70/haarnoja17a.html

Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *Proceedings Of The 35th International Conference On Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. **80** pp. 1856-1865 (2018), http://proceedings.mlr.press/v80/haarnoja18b.html

Azar, M., Gómez, V. & Kappen, H. Dynamic policy programming. *J. Mach. Learn. Res.*. **13** pp. 3207-3245 (2012), https://dl.acm.org/doi/10.5555/2503308.2503344

Any Qustions?