A Note on Soft Actor-Critic

Yu En Liu
Department of Computer Science
National Chiao Tung University
yu2guang.cs05@g2.nctu.edu.tw

1 Introduction

The Model-free deep reinforcement learning (RL) algorithms always suffer from two major challenges: high sample complexity and brittleness to hyperparameters. Both of them limit the applicability of model-free deep RL to real-world tasks. In the paper (Haarnoja et al. [2018a]), we introduce Soft Actor-Critic (SAC): the off-policy actor-critic algorithm based on the maximum entropy RL framework. In the framework, the actor aims to simultaneously maximize expected return and entropy. It results in improving exploration by acquiring diverse behaviors. Also, the maximum entropy objective considerably improves learning speed over state-of-art methods that optimize the conventional RL objective function.

In maximum entropy RL, the scaling factor has to be compensated by the choice a of suitable temperature; and a sub-optimal temperature can strongly degrade performance (Haarnoja et al. [2018b]). To resolve the issue, we devise an automatic gradient-based temperature tuning method that adjusts the expected entropy over the visited states to match a target value.

We evaluate the method on real-world challenging tasks such as locomotion for a quadrupedal robot and robotic manipulation with a dexterous hand from image observations. The results suggest that SAC is a promising candidate for learning in real-world robotics tasks.

2 Problem Formulation

We first introduce notation and summarize the standard and maximum entropy reinforcement learning frameworks.

2.1 Notation

We will address learning of maximum entropy policies in continuous action spaces, and our RL problem can be defined as policy search in an a Markov decision process (MDP). The MDP $(\mathcal{S}, \mathcal{A}, p, r)$ is specified by

- State space $\mathcal{S}$ (assumed continuous)
- Action space $\mathcal{A}$ (assumed continuous)
- State transition probability $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, \infty)$ (represents the probability density of the next state $\mathbf{s}_{t+1} \in \mathcal{S}$ given the current state $\mathbf{s}_t \in \mathcal{S}$ and action $\mathbf{a}_t \in \mathcal{A}$)
- Reward $r : \mathcal{S} \times \mathcal{A} \to [r_{\min}, r_{\max}]$ (which the environment emits on each transition)
- State marginal $\rho_\pi(\mathbf{s}_t)$ and state-action marginal $\rho_\pi(\mathbf{s}_t, \mathbf{a}_t)$ of the trajectory distribution (induced by a policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$)

Preprint. Under review.

2.2 Maximum Entropy Reinforcement Learning

The standard reinforcement learning objective is the expected sum of rewards $\sum_t \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right)\right]$ and our goal is to learn a policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$ that maximizes that objective. The maximum entropy objective (see e.g. (Ziebart [2010]) generalizes the standard objective by augmenting it with an entropy term, such that the optimal policy additionally aims to maximize its entropy at each visited state:

$$\pi^* = \arg\max_\pi \sum_t \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right)+\alpha\mathcal{H}\left(\pi\left(\cdot\mid\mathbf{s}_t\right)\right)\right], \tag{1}$$

where α is the temperature parameter that determines the relative importance of the entropy term versus the reward, and thus controls the stochasticity of the optimal policy.

If we wish to extend either the conventional or the maximum entropy RL objective to infinite horizon problems, it is convenient to also introduce a discount factor γ to ensure that the sum of expected rewards (and entropies) is finite.

$$J(\pi) = \sum_{t=0}^{\infty} \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[\sum_{l=t}^{\infty}\gamma^{l-t}\mathbb{E}_{\mathbf{s}_l\sim p,\mathbf{a}_l\sim\pi}\left[r\left(\mathbf{s}_l,\mathbf{a}_l\right)+\alpha\mathcal{H}\left(\pi\left(\cdot\mid\mathbf{s}_l\right)\right)\mid\mathbf{s}_t,\mathbf{a}_t\right]\right] \tag{2}$$

The objective corresponds to maximizing the discounted expected reward and entropy for future states originating from every state-action tuple $(\mathbf{s}_t,\mathbf{a}_t)$ weighted by its probability ρ_π under the current policy.

The core idea of maximum entropy is not to drop any useful action or useful trajectory. It has many adventages as showed below:

- **Learned policy can be used as an initialization for more complex and specific tasks.** Because through the maximum entropy, policy not only learns a way to solve the task, but learns any other aspects of the task. Therefore, such policy is more likely to learning new tasks.
- **Stronger exploration capabilities.** It can find better modes more easily under multi-modal rewards.
- **More robust and more generalization.** Since it is necessary to explore various optimal choices in different aspects, it is easier to make adjustments when it comes to interference.

3 Theoretical Analysis

3.1 From Soft Policy Iteration to Soft Actor-Critic

We devise a soft actor-critic algorithm through a policy iteration formulation, where we instead evaluate the Q-function of the current policy and update the policy through an off-policy gradient update. In this section, we treat the temperature as a constant, and later propose an extension to SAC that adjusts the temperature automatically to match an entropy target in expectation.

3.1.1 Soft Policy Iteration

The derivation is based on a tabular setting, to enable theoretical analysis and convergence guarantees, and we extend this method into the general continuous setting later.

Soft Policy Evaluation

In the policy evaluation step of soft policy iteration, we wish to compute the value of a policy π according to the maximum entropy objective. For a fixed policy, the soft Q-value can be computed iteratively, starting from any function $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and repeatedly applying a modified Bellman backup operator $\mathcal{T}^{\pi}$ given by

$$\mathcal{T}^{\pi} Q\left(\mathbf{s}_t, \mathbf{a}_t\right) \triangleq r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V\left(\mathbf{s}_{t+1}\right)\right] \tag{3}$$

where

$$V\left(\mathbf{s}_t\right) = \mathbb{E}_{\mathbf{a}_t \sim \pi}\left[Q\left(\mathbf{s}_t, \mathbf{a}_t\right) - \alpha \log \pi\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right] \tag{4}$$

is the soft state value function. We can obtain the soft Q-function for any policy π by repeatedly applying $\mathcal{T}^{\pi}$ as formalized below.

Lemma 1 (Soft Policy Evaluation) *Consider the soft Bellman backup operator $\mathcal{T}^{\pi}$ in Equation 3 and a mapping $Q^0 : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ 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 $k \rightarrow \infty$.*

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

$$Q\left(\mathbf{s}_t, \mathbf{a}_t\right) \leftarrow r_{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p, \mathbf{a}_{t+1} \sim \pi}\left[Q\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right)\right] \tag{5}$$

and apply the standard convergence results for policy evaluation: (extended proof)

$$\begin{aligned}
\|\mathcal{T}^* Q^k - \mathcal{T}^* Q^{k+1}\|_{\infty} &= \max_{\mathbf{s},\mathbf{a}} |[\mathcal{T}^* Q^k](\mathbf{s},\mathbf{a}) - [\mathcal{T}^* Q^{k+1}](\mathbf{s},\mathbf{a}) \\
&= \max_{\mathbf{s},\mathbf{a}} |(\cancel{r(\mathbf{s},\mathbf{a})} + \gamma \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{s},\mathbf{a}) \max_{\mathbf{a}'} Q^k(\mathbf{s}',\mathbf{a}')) \\
&\qquad - (\cancel{r(\mathbf{s},\mathbf{a})} + \gamma \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{s},\mathbf{a}) \max_{\mathbf{a}^*} Q^{k+1}(\mathbf{s}',\mathbf{a}^*))| \\
&\leq \max_{\mathbf{s}} \max_{\mathbf{a}} |(r(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{s},\mathbf{a}) \max_{\mathbf{a}'} Q^k(\mathbf{s}',\mathbf{a}')) \\
&\qquad - (r(\mathbf{s},\mathbf{a}) + \gamma \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{s},\mathbf{a}) \max_{\mathbf{a}'} Q^{k+1}(\mathbf{s}',\mathbf{a}'))| \\
&\leq \gamma \underline{\|p(\mathbf{s}'|\mathbf{s},\mathbf{a})\|_{\infty}}_{=1} \cdot \|Q^k(\mathbf{s}',\mathbf{a}') - Q^{k+1}(\mathbf{s}',\mathbf{a}')\| \\
&\leq \gamma \|Q^k - Q^{k+1}\|_{\infty}
\end{aligned} \tag{6}$$

The assumption $|\mathcal{A}| < \infty$ is required to guarantee that the entropy augmented reward is bounded.

Soft Policy Improvement

In the policy improvement step, we update the policy towards the exponential of the new soft Q-function. And for each state, we update the policy according to

$$\pi_{\text{new}} = \arg\min_{\pi' \in \Pi} \mathrm{D}_{\mathrm{KL}}\left(\pi'\left(\cdot \mid \mathbf{s}_t\right) \| \frac{\exp\left(\frac{1}{\alpha} Q^{\pi_{\text{old}}}\left(\mathbf{s}_t, \cdot\right)\right)}{Z^{\pi_{\text{old}}}(\mathbf{s}_t)}\right). \tag{7}$$

The partition function $Z^{\pi_{\text{old}}}(\mathbf{s}_t)$ normalizes the distribution, and while it is intractable in general, it does not contribute to the gradient with respect to the new policy and can thus be ignored.

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 7. 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 π_{new} be defined as

$$\begin{aligned}\pi_{\text{new}}\left(\cdot \mid \mathbf{s}_t\right) &= \arg\min_{\pi' \in \Pi} \mathrm{D}_{\mathrm{KL}}\left(\pi'\left(\cdot \mid \mathbf{s}_t\right) \| \exp\left(Q^{\pi_{\text{old}}}\left(\mathbf{s}_t, \cdot\right) - \log Z^{\pi_{\text{old}}}\left(\mathbf{s}_t\right)\right)\right) \\ &= \arg\min_{\pi' \in \Pi} J_{\pi_{\text{old}}}\left(\pi'\left(\cdot \mid \mathbf{s}_t\right)\right)\end{aligned} \tag{8}$$

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

$$\mathbb{E}_{\mathbf{a}_t \sim \pi_{\text{new}}}\left[\log \pi_{\text{new}}\left(\mathbf{a}_t \mid \mathbf{s}_t\right) - Q^{\pi_{\text{old}}}\left(\mathbf{s}_t, \mathbf{a}_t\right) + \cancel{\log Z^{\pi_{\text{old}}}\left(\mathbf{s}_t\right)}\right] \leq \mathbb{E}_{\mathbf{a}_t \sim \pi_{\text{old}}}\left[\log \pi_{\text{old}}\left(\mathbf{a}_t \mid \mathbf{s}_t\right) - Q^{\pi_{\text{old}}}\left(\mathbf{s}_t, \mathbf{a}_t\right) + \cancel{\log Z^{\pi_{\text{old}}}\left(\mathbf{s}_t\right)}\right], \tag{9}$$

and since partition function $Z^{\pi_{\text{old}}}$ depends only on the state and replaced RHS by Equation 4, the inequality reduces to

$$\mathbb{E}_{\mathbf{a}_t \sim \pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}\left(\mathbf{s}_t, \mathbf{a}_t\right) - \log \pi_{\text{new}}\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right] \geq V^{\pi_{\text{old}}}\left(\mathbf{s}_t\right) \tag{10}$$

by multiply minus one at both sides.

Next, consider the soft Bellman equation:

$$\begin{aligned}Q^{\pi_{\text{old}}}\left(\mathbf{s}_t, \mathbf{a}_t\right) &= r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}\right)\right] \\ &\leq r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[\mathbb{E}_{\mathbf{a}_{t+1} \sim \pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right) - \log \pi_{\text{new}}\left(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}\right)\right]\right] \\ &\vdots \\ &\leq Q^{\pi_{\text{new}}\left(\mathbf{s}_t, \mathbf{a}_t\right)}\end{aligned} \tag{11}$$

where we have repeatedly expanded $Q^{\pi_{\text{old}}}$ on the RHS by applying the soft Bellman equation and the bound in Equation 10. Convergence to $Q^{\pi_{\text{new}}}$ follows from Lemma 1.

Soft Policy Iteration

Theorem 1 (Soft Policy Iteration) *Repeated application of soft policy evaluation and soft policy improvement from any $\pi \in \Pi$ converges to a policy π^* such that $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ for all $\pi \in \Pi$ and $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$, assuming $|\mathcal{A}| < \infty$.*

Proof. Let π_i be the policy at iteration i. By Lemma 2, the sequence Q^{π_i} is monotonically increasing. And by Lemma 1, the Q^{π} will converge to some π^*. At convergence, it must be case that $J_{\pi^*}\left(\pi^*\left(\cdot \mid \mathbf{s}_t\right)\right) < J_{\pi^*}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right)$ for all $\pi \in \Pi, \pi \neq \pi^*$. In Lemma 2, we can get $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) > Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$. Hence π^* is optimal in Π.

3.1.2 Soft Actor-Critic

Due to the large continuous domains, we will use function approximators for both the soft Q-function and the policy instead of running evaluation and improvement to convergence. We alternate between optimizing both networks with stochastic gradient descent. We will consider a parameterized soft Q-function $Q_\theta\left(\mathbf{s}_t, \mathbf{a}_t\right)$ and a tractable policy $\pi_\phi\left(\mathbf{a}_t | \mathbf{s}_t\right)$. The parameters of these networks are θ and ϕ.

Soft Q-function (Critic): $Q_\theta\left(\mathbf{s}_t, \mathbf{a}_t\right)$

The soft Q-function parameters can be trained to minimize the soft Bellman residual

$$J_Q(\theta) = \mathbb{E}_{\left(\mathbf{s}_t, \mathbf{a}_t\right) \sim \mathcal{D}}\left[\frac{1}{2}\left(Q_\theta\left(\mathbf{s}_t, \mathbf{a}_t\right) - \left(r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V_{\bar{\theta}}\left(\mathbf{s}_{t+1}\right)\right]\right)\right)^2\right], \tag{12}$$

and it can be optimized with stochastic gradients

$$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta\left(\mathbf{a}_t, \mathbf{s}_t\right)\left(Q_\theta\left(\mathbf{s}_t, \mathbf{a}_t\right) - \left(r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma\left(Q_{\bar{\theta}}\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right) - \alpha \log\left(\pi_\phi\left(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}\right)\right)\right)\right)\right) \tag{13}$$

The update makes use of a target soft Q-function with parameters $\overline{\theta}$ that are obtained as an exponentially moving average of the soft Q-function weights, which has been shown to stabilize training (Mnih et al. [2015]).

(Remarks: exponentially moving average of the weights: $\overline{\theta} \leftarrow \tau\theta + (1-\tau)\overline{\theta}$, smoothing the weight by only update part of the new weights.)

Policy (Actor): $\pi_\phi\left(\mathbf{a}_t | \mathbf{s}_t\right)$

The policy parameters can be learned by directly minimizing the expected KL-divergence in Equation 7(multiplied by α and ignoring the constant log-partition function):

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\mathbb{E}_{\mathbf{a}_t \sim \pi_\phi}\left[\alpha \log\left(\pi_\phi\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right) - Q_\theta\left(\mathbf{s}_t, \mathbf{a}_t\right)\right]\right] \tag{14}$$

The target density is the Q-function, which is represented by a neural network and can be differentiated. It is thus convenient to apply the reparameterization trick instead, resulting in a lower variance estimator. To that end, we reparameterize the policy using a neural network transformation

$$\mathbf{a}_t = f_\phi\left(\epsilon_t; \mathbf{s}_t\right) \tag{15}$$

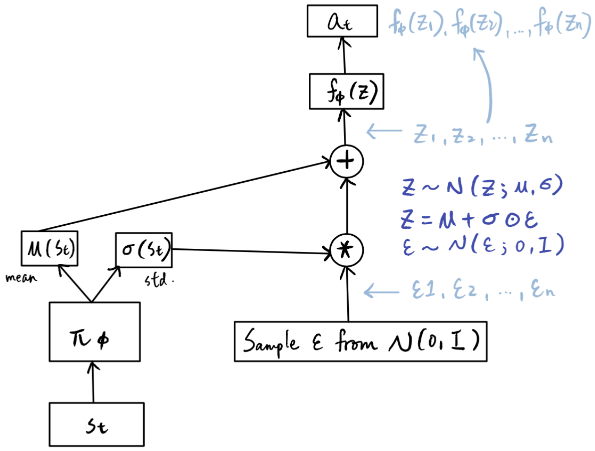
where ϵ_t is an input noise vector, sampled from some fixed distribution. We can now rewrite the objective in Equation 14 as

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}}\left[\alpha \log \pi_\phi\left(f_\phi\left(\epsilon_t; \mathbf{s}_t\right) \mid \mathbf{s}_t\right) - Q_\theta\left(\mathbf{s}_t, f_\phi\left(\epsilon_t; \mathbf{s}_t\right)\right)\right], \tag{16}$$

We can approximate the gradient of Equation 16 with

$$\hat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi \alpha \log\left(\pi_\phi\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right) + \left(\nabla_{\mathbf{a}_t} \alpha \log\left(\pi_\phi\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right) - \nabla_{\mathbf{a}_t} Q\left(\mathbf{s}_t, \mathbf{a}_t\right)\right) \nabla_\phi f_\phi\left(\epsilon_t; \mathbf{s}_t\right) \tag{17}$$

where $\mathbf{a}_t$ is evaluated at $f_\phi\left(\epsilon_t; \mathbf{s}_t\right)$.



a_t
$f_\phi(z_1), f_\phi(z_2), \ldots, f_\phi(z_n)$
$f_\phi(z)$
$\leftarrow z_1, z_2, \ldots, z_n$
$+$
$z \sim N(z; \mu, \sigma)$
$z = \mu + \sigma \odot \varepsilon$
$\varepsilon \sim N(\varepsilon; 0, I)$
$\mu(s_t)$
mean
$\sigma(s_t)$
std.
$*$
$\leftarrow \varepsilon 1, \varepsilon 2, \ldots, \varepsilon n$
π_ϕ
Sample ε from $N(0, I)$
s_t

Fig1. Reparameterization trick (extension)

3.2 Automating Entropy Adjustment for Maximum Entropy RL

Our aim is to find a stochastic policy with maximal expected return that satisfies a minimum expected entropy constraint. Formally, we want to solve the constrained optimization problem

$$\max_{\pi_{0:T}} \mathbb{E}_{\rho_\pi}\left[\sum_{t=0}^{T} r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] \text{ s.t. } \mathbb{E}_{\left(\mathbf{s}_t, \mathbf{a}_t\right)\sim\rho_\pi}\left[-\log\left(\pi_t\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right)\right] \geq \mathcal{H} \;\; \forall t \tag{18}$$

where $\mathcal{H}$ is a desired minimum expected entropy. And we rewrite the objective as an iterated maximization

$$\max_{\pi_0}\left(\mathbb{E}\left[r\left(\mathbf{s}_0, \mathbf{a}_0\right)\right] + \max_{\pi_1}\left(\mathbb{E}[\ldots] + \max_{\pi_T}\mathbb{E}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right]\right)\right), \tag{19}$$

subject to the constraint on entropy. Starting from the last time step, we change the constrained maximization to the dual problem. Subject to $\mathbb{E}_{\left(\mathbf{s}_T, \mathbf{a}_T\right)\sim\rho_\pi}\left[-\log\left(\pi_t\left(\mathbf{a}_T \mid \mathbf{s}_T\right)\right)\right] \geq \mathcal{H}$

$$\max_{\pi_T}\mathbb{E}_{\left(\mathbf{s}_t, \mathbf{a}_t\right)\sim\rho_\pi}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right] = \min_{\alpha_T\geq 0}\max_{\pi_T}\mathbb{E}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right) - \alpha_T\log\pi\left(\mathbf{a}_T \mid \mathbf{s}_T\right)\right] - \alpha_T\mathcal{H}, \tag{20}$$

where α_T is the dual variable. Since the optimal policy is the maximum entropy policy corresponding to temperature $\alpha_T : \pi_T^*\left(\mathbf{a}_T \mid \mathbf{s}_T; \alpha_T\right)$. We can solve for the optimal dual variable α_T^* as

$$\arg\min_{\alpha_T}\mathbb{E}_{\mathbf{s}_t, \mathbf{a}_t\sim\pi_t^*}\left[-\alpha_T\log\pi_T^*\left(\mathbf{a}_T \mid \mathbf{s}_T; \alpha_T\right) - \alpha_T\mathcal{H}\right]. \tag{21}$$

To simplify notation, we make use of the recursive definition of the soft Q-function

$$Q_t^*\left(\mathbf{s}_t, \mathbf{a}_t; \pi_{t+1:T}^*, \alpha_{t+1:T}^*\right) = \mathbb{E}\left[r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] + \mathbb{E}_{\rho_\pi}\left[Q_{t+1}^*\left(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}\right) - \alpha_{t+1}^*\log\pi_{t+1}^*\left(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1}\right)\right], \tag{22}$$

with $Q_T^*\left(\mathbf{s}_T, \mathbf{a}_T\right) = \mathbb{E}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right]$. And we have the dual problem (extended process)

$$\begin{aligned}
&\max_{\pi_{T-1}}\left(\mathbb{E}\left[r\left(\mathbf{s}_{T-1}, \mathbf{a}_{T-1}\right)\right] + \max_{\pi_T}\mathbb{E}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right]\right) \\
&= \max_{\pi_{T-1}}\left(Q_{T-1}^*(\mathbf{s}_{T-1}, \mathbf{a}_{T-1}) - \left(\cancel{\max_{\pi_T}\mathbb{E}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right]} + \alpha_T^*\mathcal{H}\right) + \cancel{\max_{\pi_T}\mathbb{E}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right]}\right) \\
&= \max_{\pi_{T-1}}\left(Q_{T-1}^*\left(\mathbf{s}_{T-1}, \mathbf{a}_{T-1}\right) - \alpha_T^*\mathcal{H}\right) \\
&= \min_{\alpha_{T-1}\geq 0}\max_{\pi_{T-1}}\left(\mathbb{E}\left[Q_{T-1}^*\left(\mathbf{s}_{T-1}, \mathbf{a}_{T-1}\right)\right] - \mathbb{E}\left[\alpha_{T-1}\log\pi\left(\mathbf{a}_{T-1} \mid \mathbf{s}_{T-1}\right)\right] - \alpha_{T-1}\mathcal{H}\right) - \alpha_T^*\mathcal{H}
\end{aligned} \tag{23}$$

subject to the entropy constraints. In this way, we can proceed backwards in time and recursively optimize Equation 18. We can solve the optimal dual variable α_t^* after solving Q_t^* and π_t^*

$$\alpha_t^* = \arg\min_{\alpha_t}\mathbb{E}_{\mathbf{a}_t\sim\pi_t^*}\left[-\alpha_t\log\pi_t^*\left(\mathbf{a}_t \mid \mathbf{s}_t; \alpha_t\right) - \alpha_t\overline{\mathcal{H}}\right]. \tag{24}$$

At the end, we can approximate the dual gradient descent for α with the following objective:

$$J(\alpha) = \mathbb{E}_{\mathbf{a}_t\sim\pi_t}\left[-\alpha\log\pi_t\left(\mathbf{a}_t \mid \mathbf{s}_t\right) - \alpha\overline{\mathcal{H}}\right]. \tag{25}$$

3.3 Practical Algorithm

Algorithm 1 Soft Actor-Critic

```
Input: $\theta_1, \theta_2, \phi$                                              ▷ Initial parameters
  $\bar{\theta}_1 \leftarrow \theta_1, \bar{\theta}_2 \leftarrow \theta_2$                     ▷ Initialize target network weights
  $\mathcal{D} \leftarrow \emptyset$                                          ▷ Initialize an empty replay pool
  for each iteration do
    for each environment step do
      $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$                          ▷ Sample action from the policy
      $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$                ▷ Sample transition from the environment
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$    ▷ Store the transition in the replay pool
    end for
    for each gradient step do
      $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$     ▷ Update the Q-function parameters
      $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$                       ▷ Update policy weights
      $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$                          ▷ Adjust temperature
      $\bar{\theta}_i \leftarrow \tau \theta_i + (1 - \tau)\bar{\theta}_i$ for $i \in \{1, 2\}$        ▷ Update target network weights
    end for
  end for
Output: $\theta_1, \theta_2, \phi$                                             ▷ Optimized parameters
```

4 Conclusion

In the paper, we presented soft actor-critic (SAC), an off-policy maximum entropy deep RL algorithm that provides sample-efficient learning while retaining the benefits of entropy maximization and stability. Unlike the conventional deterministic RL implementation, SAC is more general by fitting multi-modal. Also, the real-world experiments indicate that SAC is robust and sample efficient enough for robotic tasks learned directly in the real world.

References

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018a.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018b.

Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.