CS7011 : Topics in Reinforcement Learning
Summary of Soft Actor-Critic : Haarnoja, Zhou, Abbeel and Levine

Sahana Ramnath : EE15B109

Abstract

This paper (Haarnoja et al., 2018) proposes **Soft Actor-Critic***(SAC)*, an off-policy, actor-critic, deep RL algorithm based on maximum entropy RL framework. Here, the agent tries to maximize return as well as *entropy*, i.e., to succeed while acting as randomly as possible. This method solves two major challenges faced other model-free RL algorithms so far : high sample complexity and brittle convergence properties. This method acieves state-of-the-art performance on many continuous control benchmark tasks. It is also very stable, performing similarly across different random seeds.

1. Maximum Entropy RL

Standard RL maximises expected return $\sum_t \mathrm{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)]$, where ρ_π denotes the marginals of trajectory distribution induced by policy π and the reward r is bounded by $[r_{min}, r_{max}]$.
This model uses a maximum entropy objective which favors stochastic policies by adding to the objective the expected entropy of π over $\rho_\pi(s_t)$: $J(\pi) = \sum_{t=0}^{T} \mathrm{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t) + \alpha\mathcal{H}(\pi(.|s_t))]$.
The temperature parameter α controls the relative importance of the entropy term against the reward and so controls the stochasticity of the optimal policy. α is subsumed into the reward by scaling it by α^{-1}.

2. Advantages of *SAC*

- The policy is incentivized to allow the agent to explore better, while giving up on uncompromising avenues.
- The policy can capture multiple modes of near-optimal behaviour. If multiple actions seem equally attractive, the policy will give equal probability to each of them.
- The off-policy learning reuses past experience.
- It avoids the complexity and potential instability faced in previous algorithms.
- The algorithm extends easily to complex, high-dimensional tasks such as the Humanoid benchmark.

3. Soft Policy Iteration

This is a general algorithm for learning maximum entropy policies that alternates between policy evaluation and policy improvement in the maximum entropy framework. The paper proves the convergence of this method in a tabular setting and extends it to a general continuous setting. The paper proves that soft policy iteration converges to the optimal policy within a set of policies which might correspond for eg:-, to a set of parametrized densities Π. The modified 'soft' Bellman backup operator for this : $\mathcal{T}^\pi Q(s_t,a_t) \triangleq r(s_t,a_t) + \gamma\mathrm{E}_{s_{t+1}\sim p}[V(s_{t+1})]$, where, $V(s_t) = \mathrm{E}_{a_t\sim\pi}[Q(s_t,a_t) - \log\pi(a_t|s_t)]$ is the soft-value function.
In the policy improvement step, the policy is updated towards the exponential of the new Q-value.This results in an improved policy in terms of its soft value. KL-divergence is used for the update projection : $\pi_{new} = \mathrm{argmin}_{\pi'\in\Pi} D_{KL}(\pi'(.|s_t)||\frac{exp(Q^{\pi_{old}}(s_t,.))}{Z^{\pi_{old}}(s_t)})$ where $Z^{\pi_{old}}(s_t)$ normalizes the distribution.

4. Soft Actor-Critic for continuous domain

Function approximators are used both for the Q-function and the policy. The algorithm alternates between optimizing both networks with stochastic gradient descent. They use a parametrized state-value function $V_\psi(s_t)$, soft-Q function $Q_\theta(s_t,a_t)$ and a tractable policy $\pi_\phi(a_t|s_t)$. For example, the value functions can be modelled as expressive neural networks and the policy as a Gaussian with mean and covariance given by neural networks.Squared residual error is minimized for $V_\psi(s_t)$, soft Bellman residual for $Q_\theta(s_t,a_t)$ and KL-divergence for $\pi_\phi(a_t|s_t)$.
Also, two Q-functions are used to mitigate the positive bias in the policy improvement step. The minimum of these two are used in the value-gradient. The method collects experience from the environment in a single step and updates the function approximators in several gradient steps.

5. Experiments

Experiments were conducted to compare stochastic and deterministic policies. *SAC* performs consistently, whereas the deterministic variant exhibits very high variability across seeds indicating worse stability. However during policy evaluation, deterministic evaluation yields better performance. The reward is scaled to account for the temperature. *SAC* is sensitive to reward scaling since it affects the stochasticity. For small magnitudes, the policy becomes nearly uniform and large magnitudes lead to a nearly deterministic policy.

References

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