Soft Actor-Critic (SAC)

CloudWolf Live Lab 9

February 8, 2024

Abstract

Soft Actor-Critic (SAC) is a state-of-the-art algorithm in reinforcement learning that combines the strengths of off-policy learning with the stability of actor-critic methods. Known for its sample efficiency, stability, and ability to solve complex, high-dimensional control tasks, SAC incorporates entropy into the reward structure, promoting exploration by favoring stochastic policies. This document aims to elucidate the principles behind SAC, offering insights into its mathematical foundation and practical applications.

Contents

1 Introduction

Soft Actor-Critic (SAC) represents a pivotal development in the reinforcement learning landscape, particularly in continuous action spaces. By prioritizing exploration through entropy maximization, SAC achieves robust performance across a diverse range of environments.

2 Background and Motivation

2.1 Reinforcement Learning Overview

Reinforcement learning involves an agent learning to make decisions by interacting with an environment to maximize cumulative rewards. The challenges of exploration and exploitation are central to RL.

2.2 Actor-Critic Methods

Actor-critic methods utilize two models: the actor, which proposes actions given states, and the critic, which evaluates these actions. SAC enhances this framework with an entropy-based approach for improved exploration.

3 The Soft Actor-Critic Algorithm

SAC is an off-policy algorithm that optimizes a stochastic policy in an actor-critic setting, emphasizing entropy in the objective function to encourage exploration and prevent premature convergence to suboptimal policies.

3.1 Entropy-Augmented Reward

The key innovation of SAC is the incorporation of entropy into the reward signal, promoting policy diversity and exploration. The augmented objective function is defined as:

$$J(\pi) = \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[\sum_{t=0}^{T}\gamma^t\left(r(s_t,a_t)+\alpha\mathcal{H}(\pi(\cdot|s_t))\right)\right], \tag{1}$$

where $\mathcal{H}$ denotes the entropy of the policy π, and α is a temperature parameter that balances the trade-off between entropy and reward.

3.2 Soft Policy Iteration

SAC utilizes soft policy iteration, alternating between soft policy evaluation and improvement steps to converge to the optimal policy that maximizes the entropy-augmented objective.

3.2.1 Soft Policy Evaluation

Given a policy π, the soft Q-function is updated to minimize the difference between the left-hand side and the right-hand side of the soft Bellman equation:

$$Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p} \left[V(s_{t+1}) \right], \tag{2}$$

where $V(s_t)$ is the soft value function, representing the expected return of following π from state s_t, including entropy bonuses.

3.2.2 Soft Policy Improvement

The policy is updated to maximize the expected return plus entropy, effectively improving by acting greedily with respect to the current soft Q-function.

3.3 Practical Implementation

Practical SAC implementations parameterize the actor and critic with neural networks, using reparameterization tricks for stable policy updates and employing automatic temperature adjustment to find the optimal α.

4 Advantages of SAC

SAC's emphasis on entropy not only encourages thorough exploration but also leads to more robust policies that are less sensitive to perturbations, making it well-suited for a wide range of tasks, from robotics to video game AI.

5 Conclusion

Soft Actor-Critic stands as a testament to the power of integrating entropy into reinforcement learning. Its balance between exploration and exploitation, efficiency in sample use, and applicability to complex problems make it a valuable tool in the RL toolkit.