**Intelligent Control**:

*Implementing Soft Actor-Critic (SAC) from Scratch*

*Description: Implement the SAC algorithm from scratch using PyTorch or JAX and apply it to robotics tasks. Analyze the impact of entropy regularization, target networks, and Q function stability.*

ΜΑRKANTONATOU EIRINI | up1095489

SOLOMONIDI ELEFTHERIA | up1092837

**SEMESTER:** 8th

**1. Introduction**

This project explores a model-free reinforcement learning (RL) approach for training an agent to interact with an environment and learn optimal behavior through experience. The main focus is the implementation of Soft-Actor-Critic (SAC) algorithm, an efficient off-policy algorithm designed for continuous action spaces. A central feature of SAC is entropy regularization which encourages the policy to remain stochastic and exploratory throughout training. The policy is trained to maximize a trade-off between expected return and entropy, a measure of randomness in the policy.

**2. Theoretical Foundations of SAC**

*Limitations of Standard Actor-Critic Methods*

Actor-Critic Algorithm is a type of reinforcement learning algorithm that combines two parts i.e the Actor which selects actions and the Critic which evaluates them. In many environments it works satisfactorily, but there are environments especially those with sparse rewards, the agent may prematurely converge to a suboptimal policy because it fails to adequately explore alternative strategies. This limited exploration leads to the risk of getting stuck in local optima and not explore alternative strategies that lead to better results. Another issue when using the Actor-Critic methods is that they often generalize poorly, as the actor tends to learn a single strategy that performs well during training but are no robust to unseen or slightly different situations.

These problems show why methods like Soft Actor-Critic (SAC), which focus on better exploration and randomness, are helpful in fixing these issues.

*Addressing issues with SAC*

In Soft-Actor-Critic algorithm instead of only asking the Actor to maximize it’s reward when training, the critic will ask the actor to do so while acting as randomly as possible. That is maximizing the entropy of its current policy. “Soft” refers to “entropy regularization”. The entropy of the policy being maximized, the actor is incentivized to explore alternative strategies and to recover from unexpected situations.

Consequently, it introduces a value function, the soft action – value function.

SAC parameterizes a policy and a soft Q- value function .

How does it work?

In SAC we use a policy iteration formulation, in which we evaluate the Q-function of the current policy and update the policy through an off policy gradient rule. In general Actor – Critic methods are derived from policy iteration, which alternates between **policy evaluation** and **policy imporovement** .

*Policy Evaluation*

Policy Evaluation is the initial step in reinforcement learning, involving the determination of the value function for a given policy.

In standard RL the policy evaluation optimizes only for rewards :

*Bellman Optimality Equation*:

**(2.1)**, where:

**(2.2)**

Instead in the policy evaluation step of soft policy iteration the policy is computed using the maximum entropy objective with the modified Bellman backup operator .

*Soft Bellman Equation:*

**(2.3)**, where:

**(2.4)**

Including the entropy term:

**(2.5)**,

the **uncertainty** or **randomness** of the policy’s action choices in a given state is quantified. The higher the entropy, the more spread out or exploratory the policy is.

Can entropy be whatever you want?

No. It is designed to measure uncertainty or randomness in a probability distribution. The term is the probability that the policy chooses action in .

* When is close to 1 is near 0 (high confidence)
* When is close to 0 is a large negative number (low confidence)
* **So is small for confident choices, large for unlikely ones**

Εικόνα που περιέχει κείμενο, γραμμή, γράφημα, διάγραμμα

Το περιεχόμενο που δημιουργείται από AI ενδέχεται να είναι εσφαλμένο.

Figure 1

This figure illustrates how SAC penalizes overconfident decisions to encourage exploration.

In SAC we want policies that balance high reward with high entropy. So penalizes overconfident actions and rewards diverse exploratory behavior.

*Policy Improvement*

Once the value function is evaluated, Policy Improvement follows. Consider a certain action in a state that has a higher expected reward than the current policy's action, then the policy is updated to choose the better action in that state.

In most RL algorithms the policy is improved using the following objective:

**(2.6)**

SAC changes that traditional MDP objective and converges to different solutions. The maximum entropy objective generalizes the standard objective (1) by adding an entropy term, such that the optimal policy additionally aims to maximize its entropy at each visited state:

**(2.7)**

Where a is called “temperature parameter” and controls the balance between reward maximization and exploration in the training process. This term determines the stochasticity of the optimal policy. Specifically:

* A higher α encourages more exploration, since the policy is rewarded for being more random.
* A lower α shifts the focus toward exploitation, where the policy sticks to high-reward actions.

*Iterative Soft Q-Value Optimization and Policy Update in SAC*

To optimize our action-value estimates, we employ the Bellman operator, which iteratively updates the soft Q-values for a fixed policy. Starting from any function and repeatedly applying Equation (2.3).

In the policy improvement step, we update the policy towards the exponential of the new soft Q function. We make actions that have higher soft Q-values more probable (softmax like update).

In reinforcement learning environments, actions can typically be either discrete or continuous. For discrete action spaces, action probabilities can be modeled directly using techniques like the softmax function. In contrast, continuous action spaces require modeling a probability density function, such as a Gaussian distribution.

A Gaussian Distribution is a common way to model continuous stochastic policies:

**(2.8)**,

where s is the current state.

Why are Gaussians tractable?

If you define a Gaussian as shown in Equation (2.8) then sampling an action is:

*(reparameterisation trick)*

Instead of sampling directly from a distribution (which is a non-differentiable operation) we apply the reparameterization trick. This technique expresses the random variable as a deterministic function of fixed random noise and learnable parameters, enabling gradients to flow through the sampling process.

Since in practice we prefer policies that are tractable we will restrict the policy to Gaussians. In the Soft Actor-Critic (SAC) algorithm, Gaussian distributions are used to represent stochastic policies in continuous action spaces.

*Policy Improvement Step*

The partition function Z 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.

Implementation:

3 Neural Networks. We will consider a parameterized soft Q-function, and a tractable policy ). The Q neural network is trained to approximate the **soft Bellman operator** while the policy network can be modeled as a Gaussian with mean and covariance given by neural networks:

: The most likely action

: randomness around that action

Why output instead of ?

* σ must be positive (it is a standard deviation)
* If you output it directly, your network might output negative or zero values, which would break the Gaussian:

3 objective functions:

**3. Environment Setup and Agent Dynamics**

*PointMaze Simulation Setup*

In the context of this project, the Gymnasium library was used, which provides simulation environments for reinforcement learning experiments. Specifically, the PointMaze environment was selected, simulating a maze where an agent must navigate from a randomly initialized starting position to a predetermined goal. The chosen environment is a two-dimensional horizontal maze in which the agent behaves like a point mass free to move along the x and y axes. At each time step, the agent observes its current **state** and selects an **action**, which influences its **next state** and the **reward** it will receive. Those key concepts are explained below:

*State*

The agent's state includes its current position (achieved\_goal) and the position of the desired goal (desired\_goal). The agent needs to know the desired goal at every step (a fixed and specific vector for each episode) because this determines where it should move and what it is trying to achieve. Especially in environments where the goal changes each episode, incorporating the desired goal is necessary for learning a policy that can generalize to different targets.

*Action*

The agent’s action represents the linear force applied to the point mass in the x and y directions, moving it to the next state. At each time step, the agent selects an action corresponding to a continuous change in its position within the maze. Once an action is selected, the agent transitions to the next state, receives a reward, and a signal indicating whether the episode is done.

*Reward*

The reward function used in this environment is based on **sparse rewards**. Specifically, the agent receives a reward if and only if it reaches the goal (the ball is considered to have reached the goal when the Euclidean distance is less than 0.5 m), while all intermediate steps yield zero reward. This makes the learning process more challenging, as the agent does not receive direct feedback on whether it is moving in the right direction or not. Learning is primarily based on a few but crucial experiences where the agent achieves the goal. Despite the use of sparse rewards, the Soft Actor-Critic (SAC) algorithm manages to learn effectively, mainly thanks to the use of **replay buffer**.

*Replay Buffer*

The replay buffer stores all the agent’s **experiences (state, action, reward, next state, done)** at each step of the episode, allowing successful experiences to be reused. During training, the agent samples randomly from the buffer, selecting mini-batches of experiences according to a predefined batch size. This way, the information about "success" is not lost but can reappear multiple times, gradually strengthening the agent’s policy. When an experience includes reaching the goal (i.e., a high reward), the Q-network is updated. Specifically, the Q-value increases because the target equation

Q(sₜ, aₜ) = r + γ ·Q(sₜ+1, π(sₜ+1)) (3.1)

includes a high reward r. The Q-network (the Critic) gradually learns that certain state-action pairs lead to high rewards. At the same time, the policy network (the Actor) is trained to select actions that maximize the Q-value—those most likely to result in successful transitions. As a result, the probability that the agent will reproduce success increases, even in environments with sparse rewards.

*Neural Networks*

This process is supported by the two neural networks mentioned earlier: the **Actor network** and the **Critic network**.

The neural network for the Critic tries to estimate the Q-values, that is, to calculate the expected returns after selecting a specific action from a given state following the policy. Two Q-functions are used to prevent overestimation of the Q-value, meaning to avoid the network mistakenly thinking an action is much better than it really is, which would cause the Actor to prefer it without justification. The critic\_loss measures how well the Critic learns to predict how good or bad a chosen action is; in other words, it evaluates the actions. However, the selection of actions is based on the Actor’s neural network.

The Actor essentially tries to learn a stochastic policy that maximizes expected returns while maintaining high entropy (H), aiming to select good actions. A policy is stochastic when the actions taken at each step are sampled from a probability distribution so that the same action is not always chosen for a given state. Stochasticity thus introduces randomness in action selection. The actor\_loss indicates how well the agent’s policy is being trained. When the actor\_loss is high, it means the Actor proposes actions that, according to the Critic, do not lead to high rewards, signaling that further training is needed to improve the policy.

*Entropy Regularization MORE*

As mentioned earlier in section 2 , besides maximizing the expected reward, the SAC algorithm incorporates another important objective: maintaining high entropy in the policy. **Entropy regularization** is a technique that improves **exploration** by encouraging the policy to maintain a balanced distribution over actions, preventing it from becoming too deterministic too quickly. Without considering entropy regularization, the policy may quickly converge to a small set of actions that yield high rewards but potentially ignore better alternative actions. By introducing the entropy term, we define how stochastic, or how "random," we want the agent’s policy to be when selecting actions. This way, the agent learns to explore the environment by trying different actions instead of getting stuck repeating the same ones.

Specifically, in this work, the entropy is dynamically adjusted during training through the temperature parameter (denoted by α), which controls how much weight entropy has in the training process. If the policy becomes too "confident" (low entropy), α increases to promote more randomness. Conversely, with high entropy, α decreases to make the system more deterministic.

*Agent training*

Everything analyzed so far refers to the data used to train the agent. Training is divided into episodes, and at the beginning of each episode, the environment is reset to an initial state. Each episode ends after a maximum number of steps.

In the first few episodes (the warm-up phase), actions at each step are chosen randomly to allow the agent to explore the environment and fill the replay buffer with some initial experiences. During this phase, the Actor neural network is being trained but is not yet used to select actions.

After enough data has been collected from these episodes, the actual training begins. From this point on, at each step, actions are selected by the Actor network. More specifically, the agent samples past experiences by taking random batches from the buffer and performs updates on the Actor and Critic neural networks.

* *Critic Update*

In the Soft Actor-Critic (SAC) algorithm, two Q-networks and two corresponding target Q-networks are maintained. To train the Critic, the network learns to estimate how good an action is in a given state by computing Q-values. These are compared to the target Q-values computed using the Bellman equation (I), based on the target networks. The objective is to minimize the loss between the prediction and the target. The process works as follows:

From a stored transition (s, a, r, s′) collected by the agent:

The target Q-value is computed as an estimate for the (s, a) pair:

**(3.2)**

This Q\_target(s, a) is the value we want the Critic to learn.

Then, the loss of the Q-networks is computed (let y = Q\_target(s, a)):

**(3.3)**

Each Q-network tries to bring its output as close as possible to *y*. Then, backpropagation is performed and the weights of the Q-networks are updated.

* *Actor Update*

To train the Actor network, once an action is selected for a given state s, both Q1 and Q2 values are computed for that action, and the smaller of the two is used. This serves as a conservative estimate to prevent overestimating the action value.

Then the actor loss is computed using the equation:

**(3.4)**

The goal is to **minimize** this term.

The actor is trying to choose actions that:

* Yield a high Q(s, a) value (i.e., good rewards)
* Have high entropy (low confidence), encouraging exploration

Backpropagation is then performed and the weights of the policy network are updated.

* *Temperature Update*

In addition to updating these two neural networks, the entropy coefficient α is also updated automatically to control how random or confident the policy should be.

For this reason, the temperature loss is computed:

We perform gradient descent on this loss to adjust α:

* If the policy is too predictable → **increase α** to add more randomness
* If the policy is too random → **decrease α** to make it more focused