A Brief Overview of the Actor-Critic Method

A policy $\pi_\theta(A|S)$ gives the probability of taking action A given S. The goal then is to maximze the objective function

$$C_{t_0}(\theta) = E[\sum_{t=t_0}^{\infty} \log(\pi_\theta(a_t|s_t)) * r(a_t|s_t)],$$

where $r(A|S)$ represents the reward from taking action A at state S. We then define

$$Q(a_t|s_t) = C_t(\theta|a_t),$$

representing the q-value. The q-value of taking an action A at state S is equal to the expected sum of the discounted rewards, given that the action taken at time t is a_t. We now define two agents: an actor, $\pi_\theta(A|S)$, which predicts the probability of taking action A at state S, and critic $q_\theta(a_t|s_t)$, which predics the value of $Q(a_t|s_t)$. Now, as the actor trains on the environment, the critic can be optimized by looking at the actor's total reward after an episode, and the actor is able to learn using the critic's predicted q-values. [TODO MAKE MORE DETAILED]

Title about my algorithm

The agent will be trained on low-level, general tasks, such as walking forward and turning. After these low-level skills have been developed, we can map movements into a latent space using an autoencoder then, another agent will be able to learn to perform specific tasks using generalized skills by predicting actions on the latent space instead of raw movements.