Proposal Title: Implementing policy gradient using “difference of values” function

Hossein Najafi, Roham Ghotbi

Background

A classic and conventional policy gradient relies on values assigned to each rollout and each action taken at any given state and is therefore susceptible to a very large dynamic range and hence a very large variance. The “values” assigned are usually either directly computed rewards-to-go, Q(s,a) or V(s), where Q and V are respectively the Q and value function. It has been shown that reducing the magnitude of these values while preserving their information, could yield to a much lower variance and faster training.

Method

We propose to implement “differential training” or training using “difference of values”. This method relies on not the absolute values of Q and V to make a decision, but rather how the different options available at any given state differ from one another in terms of their respective Q, and using this difference and contrast between different actions and state, we can train the policy with a much lower variance. Part of this work will initially focus on gaining a deep enough knowledge about the implementation of this new method, and subsequently create and train an agent that relies on it.

Verification

Using available simulators (Mujoco), we will verify this hypothesis and implement this method on a policy gradient and rate and compare it to its conventional classic policy gradient counterparts. Our end goal will be to demonstrate how much benefit this method will bring to policy gradient methods and provide a comprehensive analysis of the classic and differential policy gradient methods.