Abstract:
In Reinforcement Learning (RL), Temporal-Difference (TD) methods are a family of powerful algorithms that apply bootstrapping to efficiently learn various value functions by relating predictions at consecutive time-steps to each other. For TD(0) --- the most fundamental of these --- there exist proofs of convergence when using parametrized function approximators in the on-policy case. TD(0) can also be used for off-policy learning, where data gathered under one behavior policy can be reused to predict the future under a different policy. However, the confluence of approximation, bootstrapping, and off-policy learning --- termed the deadly triad --- runs a risk of divergence. In this thesis, we study how feature normalization can affect the convergence properties of TD(0). Following this analysis, we propose CrossNorm, a variant of Batch Normalization that accounts for the presence of two data distributions in off-policy learning. We demonstrate empirically that CrossNorm improves the stability of the learning process in both policy evaluation and policy improvement. In the Deep RL case, CrossNorm is easily incorporated into methods like DDPG and TD3. For the first time, we demonstrate stable and superior training of DDPG without needing target networks.