Research Paper Ideas

Trust Region Policy Optimization

Reducing Approximations

- How does the replacement of the penalty term with a constraint affect the monotonic improvement guarantee, and if the guarantee no longer holds, how can we (perhaps dynamically) adjust δ in the constraint to better ensure monotonic improvement?
  - This will likely require knowledge of constraint optimization algorithms.
- Instead of replacing D_{KL}^{max} with the expected KL-divergence $\bar{D}_{KL}^{\rho}$, can we replace it with something that better approximates D_{KL}^{max}?
- Considering the theory behind equation (14) and the single path and vine implementations of TRPO, it seems like the authors implicitly approximate the normalized discounted visitation frequencies $(1-\gamma)p_{\pi_{\theta_{old}}}(s) = (1-\gamma)(P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \dots)$ with the on-policy distribution of states under $\pi_{\theta_{old}}$. To what extent can this be justified? It seems that they are not necessarily equal, because the former depends on the discount factor γ, while the latter does not.

Improving Sample Efficiency

- To what extent can samples from previous runs (e.g. from runs using values of θ older than θ_{old}) be re-used in the current iteration? Perhaps importance sampling can help here?

Generalizations

- How can we generalize the vine method beyond simulated environments? Can resetting an uncontrolled environment to a state that is "similar enough" to the branch point ever yield competent performance?

Proximal Policy Optimization

Improving Clipped Approximation Function

- Is there a way to robustly choose a clipping parameter ϵ? Can this parameter be dynamic?
- The sample advantage $\hat{A}_t$ expanded in equation (10) seems inherently biased towards 0 because it makes use of a T-step return that bootstraps off of $V(s_T)$. How can this bias be addressed?

Investigating Other Potential Approximation Functions

- The clipping version of PPO only imposes a penalty (reduction in approximated policy value) for a given sign of $\hat{A}$ if the policies diverge in one direction (where "direction" refers to whether $r_t(\theta) > 1$ or $r_t(\theta) < 1$), while the theory behind TRPO would suggest imposing a penalty in both directions. Only penalizing one direction could be helpful for increasing step size, but doesn't this come at the cost of the monotonic improvement guarantee? Is there a more precise way of modulating the tradeoff between step size and monotonic improvement?
- The original paper on TRPO replaced the penalty term with a constraint because the penalty coefficient $C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$ caused the step sizes to be "too small." By what standard is "too small" defined, and, after determining what this is, can we alter β in equation (8) to make sure we are taking a step size that is as large as possible without being too big? Section 4 presents a techique for modifying β, but there doesn't seem to be much theory behind it, so can that be improved?