Deep RL for the Average Reward Objective

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, Rahul Kulkarni

Average-Reward Reinforcement Learning

- Sequential decision-making problems are typically solved with "discounted" RL (γ)
- However, the **average** evaluation reward is usually the object of interest
- We thus instead directly optimize the "average reward" objective in RL
- Upon adding entropy regularization, connects to free energy formulation
- Gives time-homogeneous, linear framework

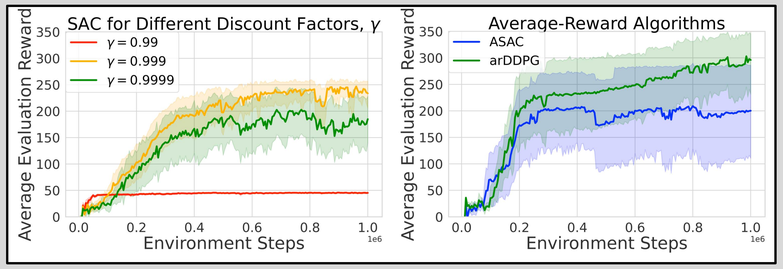
$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p}\, V^*(s')$$
$$V^*(s) = \max_a Q^*(s,a)$$

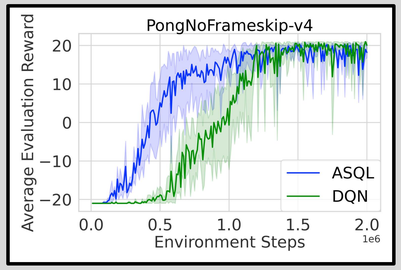
$$Q^*(s,a) = r(s,a) - \theta + \mathbb{E}_{s' \sim p}\, V^*(s')$$
$$V^*(s) = \beta^{-1} \log \mathbb{E}_{a' \sim \pi_0}\, e^{\beta Q(s,a)}$$

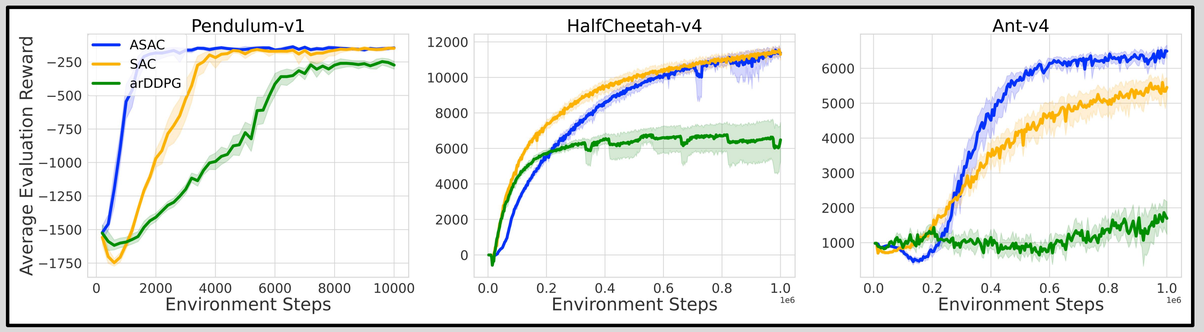
If the **average reward-rate**, θ, is also learned, prior SOTA (SAC, SQL) can be extended to the average-reward framework!

- Expands average rewards literature, especially for deep value-based methods
- We prove PI+PE+convergence of our algo's
- Useful in physical systems!
  - ➢ γ is usually non-physical (quantum ctrl.)

Experiments







- Using known value-based techniques from DQN/SQL/SAC; we demonstrate that the average-reward objective is viable
- The average reward objective is superior to discounting for continuing tasks and can be a competitive algorithm!
- Outperforms the current SOTA for average reward (arDDPG) on several tasks

Theory

We prove policy improvement (PI) in this setting using free energy minimization:

Theorem 1 (ERAR Policy Improvement). *Let a policy π absolutely continuous w.r.t. π_0 and its corresponding differential value $Q_\theta^\pi(s,a)$ be given. Then, the policy*

$$\pi'(a|s) \doteq \frac{\pi_0(a|s) e^{\beta Q_\theta^\pi(s,a)}}{\sum_a \pi_0(a|s) e^{\beta Q_\theta^\pi(s,a)}} \qquad (9)$$

achieves a greater entropy-regularized reward-rate. That is, $\theta^{\pi'} \geq \theta^\pi$, with equality only at convergence, when $\pi' = \pi = \pi^*$.

And exactly characterize the gap between consecutive "improved" policies:

Lemma 2 (ERAR Rate Gap). *Consider two policies π, π' absolutely continuous w.r.t. π_0. Then the gap between their corresponding entropy-regularized reward rates is:*

$$\theta^{\pi'} - \theta^\pi = \mathop{\mathbb{E}}_{\substack{s \sim d_{\pi'} \\ a \sim \pi'}} \left(A_\theta^\pi(s,a) - \frac{1}{\beta} \log \frac{\pi'(a|s)}{\pi_0(a|s)} \right), \qquad (8)$$

where $A_\theta^\pi(s,a) = Q_\theta^\pi(s,a) - V_\theta^\pi(s)$ is the advantage function of policy π and $d_{\pi'}$ is the steady-state distribution induced by π'.

Also offer algo for un-regularized objective:

Algorithm 2 Posterior Policy Iteration (PPI)

Initialize: Prior policy π_0, $\beta > 0$, solve budget.
while $N <$ solve budget **do**
  $\pi_0 \leftarrow \text{Solve}(\pi_0, \beta)$
end while
Output: Deterministic optimal policy $\pi^*_{\beta=\infty} = \pi_0$

Discussion

Deep RL with average-reward objective works!

- First work to combine average-reward & entropy-regularized objectives
- Proved convergence regarding policy evaluation / policy improvement
- Applications to physical problems with time homogeneity
- We also developed successful approaches for un-regularized problem
  - ➢ *Posterior policy iteration* updates the prior policy $\pi_0 \leftarrow \pi^*$ to achieve zero entropic cost

SCAN FOR MORE