Model-Mixed Value Estimation

Vladimir Feinberg, Michael I. Jordan, Ion Stoica, Joseph Gonzalez, Sergey Levine
University of California, Berkeley

Motivation

- Sample complexity reduction for model-free function-approximated value-based RL.
- Currently, the effect of multiple-step value learning is poorly understood.
- Focus on continuous control: we may only have sparse but known rewards and a rich transition signal.

Definitions

Fully observable Markov Decision Process over continuous spaces:

- States $s \in \mathcal{S}$
- Actions $a \in \mathcal{A}$
- Known reward $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, bounded.
- Dynamics are deterministic and feasible: we have access to a model class $\{f_\theta\}_\theta$ such that for any policy $\pi : \mathcal{S} \to \mathcal{A}$ there exists a θ_π such that $f_{\theta_\pi} : \mathcal{S} \times \mathcal{A} \to$ satisfies $f_{\theta_\pi}(s, a) = s'$, where $a = \pi(s)$, s is a state that arises from π's marginal state distribution, and s' is the true resulting state. We do not know θ_π.

All stochasticity is currently from the initial state. Objective is to maximize infinite-horizon discounted reward:

$$\max_\pi \mathbb{E}_{s_0} \sum_{t=0}^{\infty} \gamma^t r_t, \quad r_t \triangleq r(s_t, a_t, s_{t+1})$$

For a fixed $\pi : \mathcal{S} \to \mathcal{A}$ in context:

$$Q^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t r_t, \text{ where } s_0 = s, a_0 = a$$

Denote estimates with a hat. E.g., $\hat{r}_1 = r(s_0, a_0, \hat{s}_1)$ where $\hat{s}_1 = f_{\hat{\theta}}(s_0, a_0)$. Define the h-step estimator:

$$\hat{Q}_h(s, a) = \sum_{t=0}^{h} \gamma^t \hat{r}_t + \gamma^h \hat{Q}(\hat{s}_h, \pi(\hat{s}_h))$$

Theoretical Gaps

(Szepesvári 2009) Multi-step Q isn't obviously convergent, even with perfect dynamics in the tabular case. Defining:

$$\mathcal{T}_h^\pi(Q)(s, a) = \sum_{t=0}^{h-1} \gamma^t r_t + \max_{a'} \gamma^h Q(s_h, a'), a_t = \pi(s_t)$$

If $\pi = \text{argmax}\, Q$, $\mathcal{T}_h$ won't contract more than γ, even if $h > 1$. If π is fixed then $\mathcal{T}_h$ contracts by γ^h, but does the analysis hold as we update π?

Contributions (Ongoing Work)

- Extensive empirical validation of (van Seijen 2016) claim (improved Q^π estimator accelerates training).
- Flexible generalization of existing joint model-based and model-free methods (Dyna, Stochastic Value Gradients, ME-TRPO, NAF): $\tilde{Q} = \sum_h w(h) \hat{Q}_h$.

(van Seijen 2016) Lower-bias estimates of Q^π still help in practice. With perfect dynamics, $\hat{Q}_h$ exponentially improves Q^π MSE throughout training:

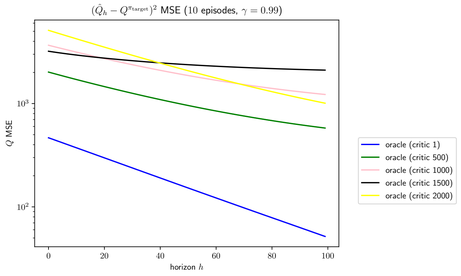

Fig. 1: DDPG-based mixture estimator MSE at various points during training on the `HalfCheetah` task.

Impact. If successful, this can demonstrate utility of a new class of joint model-free and learned-model-based algorithms. Applications to robotic control are immediate.

Model-Mixed DDPG

Initialize $\pi, \hat{\theta}, \hat{Q}$. Loop:

1. Collect a sample playing π, add to replay buffer B.
2. Update $\hat{\theta}$ with loss $\mathbb{E} \left\| s' - f_{\hat{\theta}}(s, a) \right\|^2$, $(s, a, s') \sim B$
3. Estimate $w(h)$ used in $\tilde{Q}$ (via sampling, holdout, or delta method)
4. Update $\hat{Q}$ with loss $(\hat{Q} - \tilde{Q})^2$ sampling from B.
5. Update π to maximize $\hat{Q}(s, \pi(s))$ for $s \in B$.

Results Preview

Fix ideal dynamics $\theta = \theta_\pi$ and $w(h) = \mathbb{1}\{h = 5\}$.

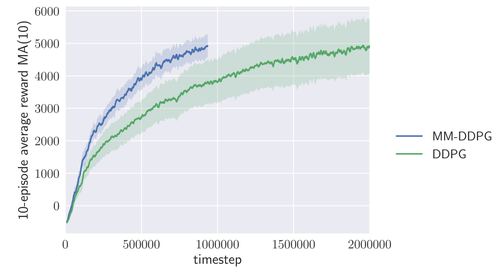

Fig. 2: Comparison of MM-DDPG to DDPG on `HalfCheetah`.

Future Work

- Apply model-mixtures to other actor-critic algorithms, even policy-gradient ones.
- Theoretical results in probabilistic tabular setting with iterate-based analysis (Szepesvári 1997).
- Extend to constrained model-predictive control by learning f which generalizes to actions near π, not just on π.