Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion

Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, Honglak Lee

Background

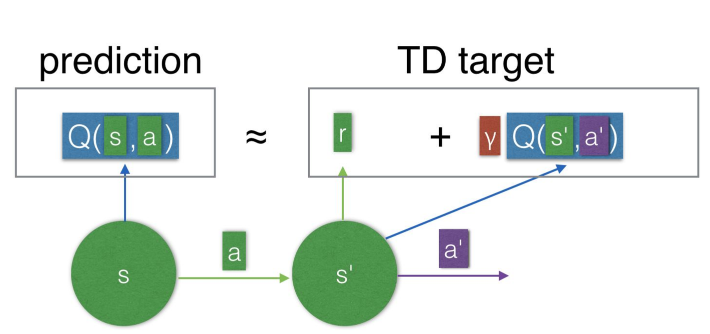
- Goal of reinforcement learning: given the ability to interact with an environment,
  - maximize expected reward
  - minimize number of interactions (*sample efficiency*)
- Algorithms can be split into two categories:
  - Model-free RL: Take actions in the environment, and learn a policy which generalizes and the most successful actions
  - Model-based RL: Learn a dynamics model of the environment, then learn a policy that succeeds in the modeled environment (*planning*)
- Since dynamics is a much richer signal than reward, model-based RL is typically more sample-efficient. But relying on a model comes with many challenges:
  - Approximation error: the learned model puts an upper bound on performance
  - Exploiting inaccuracies: the planner is adversarial to the model
  - Accumulating errors: small modeling errors accumulate quickly
  - Mode collapse: lack of trajectory diversity causes blind spots
- Uncertainty-aware hybrid model-free/model-based approaches show promise in remedying these issues

Preliminaries: Q-Learning and MVE

Q-Learning

$$L_\theta = \mathbb{E}_{(s,a,r,s')}\left[(\hat{Q}^\pi_\theta(s,a) - \mathcal{T}^{TD}(r,s'))^2\right]$$

$$\mathcal{T}^{TD}(r,s') = r + \gamma \hat{Q}^\pi_{\theta^-}(s', \pi(s'))$$

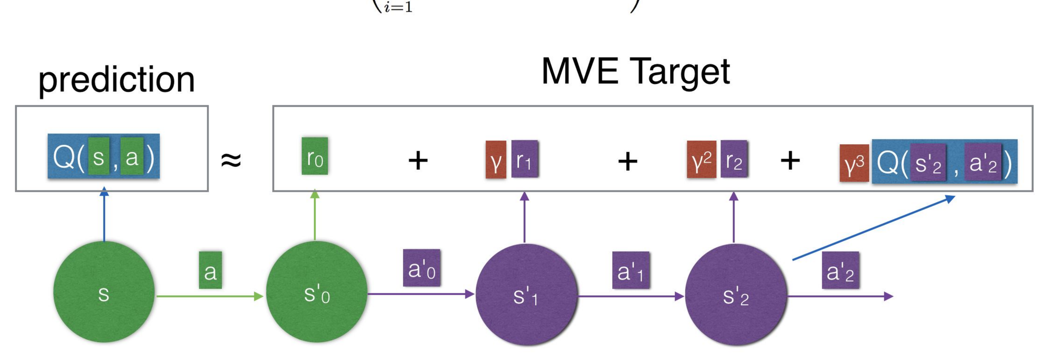


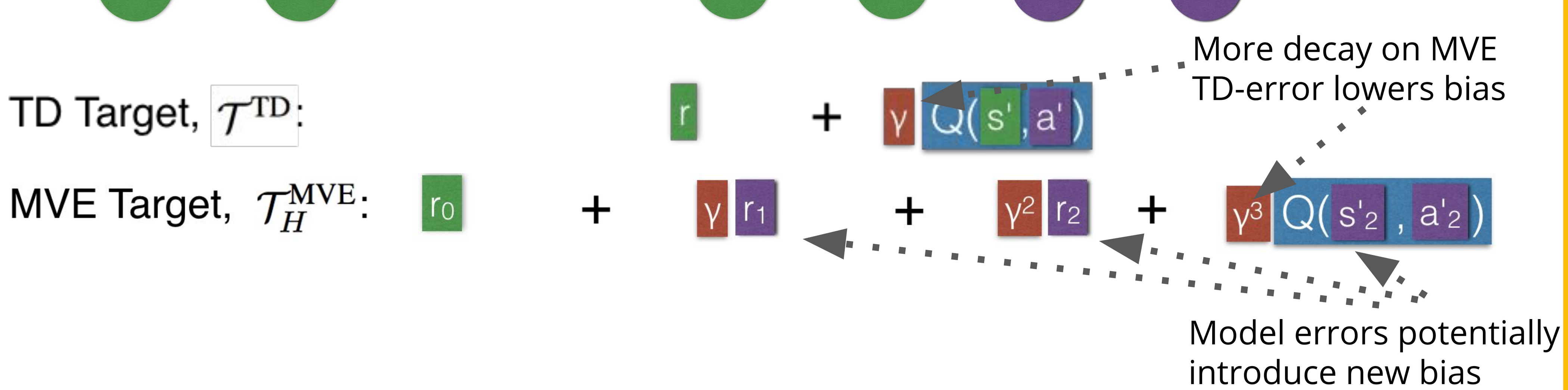
Model-based Value Expansion

(Feinberg et al. 2018)

$$s'_0 = s', \quad a'_i = \pi_\phi(s'_i), \quad s'_i = \hat{T}_\xi(s'_{i-1}, a'_{i-1}), \quad D^i = \prod_{j=0}^{i}(1 - d(s'_j))$$

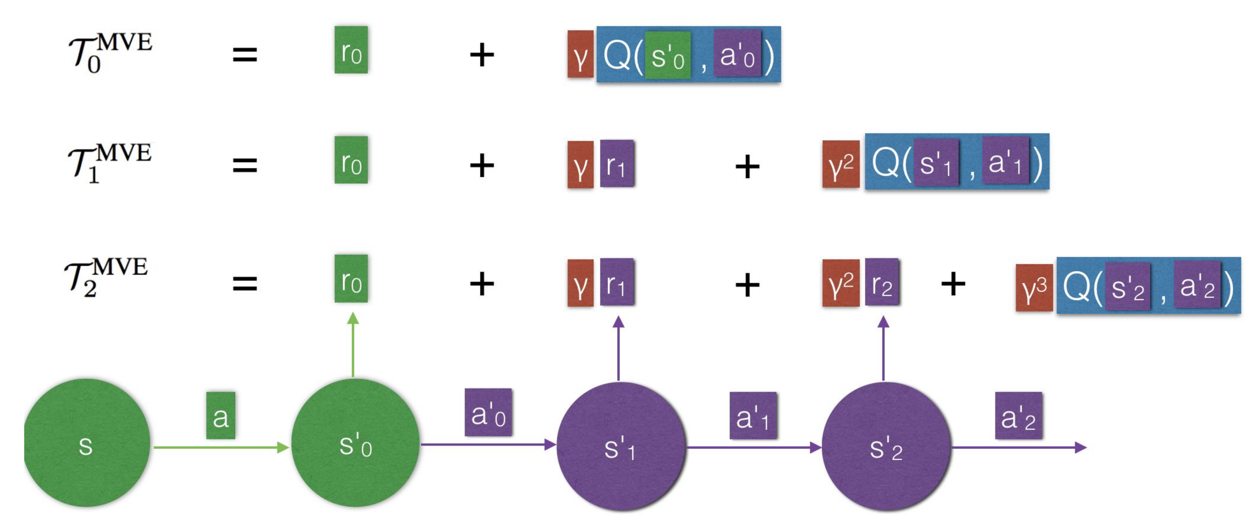
$$\mathcal{T}^{\mathrm{MVE}}_H(r,s') = r + \left(\sum_{i=1}^{H} D^i \gamma^i \hat{r}_\psi(s'_{i-1}, a'_{i-1}, s'_i)\right) + D^{H+1}\gamma^{H+1}\hat{Q}^\pi_{\theta^-}(s'_H, a'_H).$$



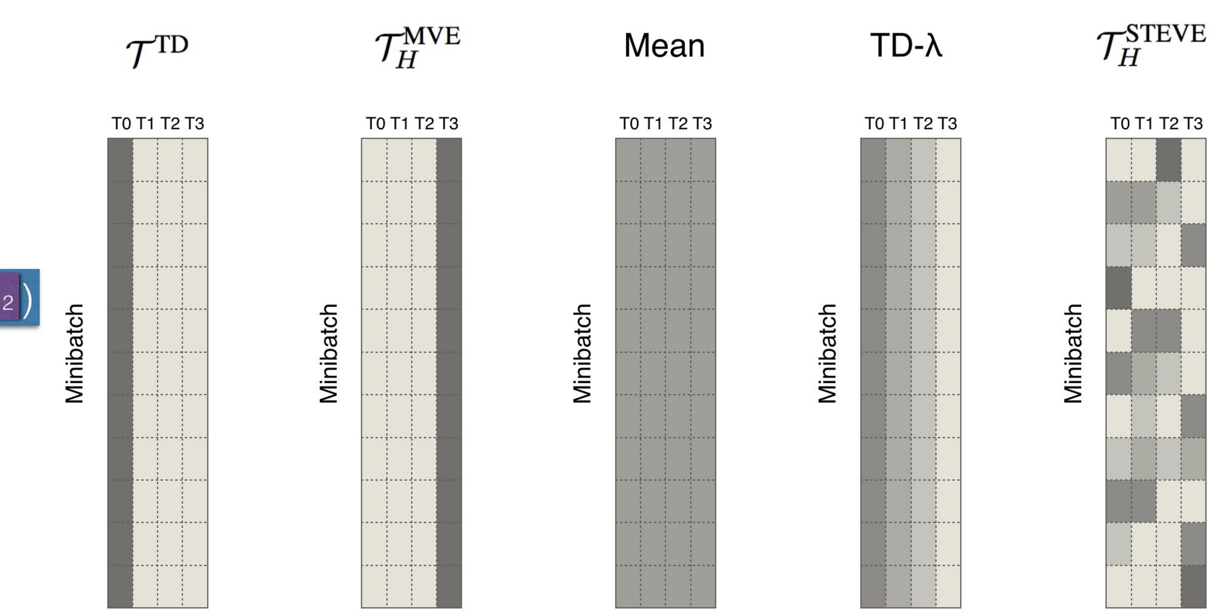


Stochastic Ensemble Value Expansion (STEVE)

For a *maximum* horizon *H*, we actually have *H*+1 distinct candidate targets: $\mathcal{T}^{\mathrm{MVE}}_0, \mathcal{T}^{\mathrm{MVE}}_1, \mathcal{T}^{\mathrm{MVE}}_2, \ldots, \mathcal{T}^{\mathrm{MVE}}_H$



STEVE dynamically adjusts model usage based on uncertainty; other options are fixed and inflexible



Estimate uncertainty by variance under an ensemble, and construct a target by an inverse-variance weighted sum of candidates:

$$\mathcal{T}^{\mathrm{STEVE}}_H = \sum_{i=0}^{H} w_i \mathcal{T}^{\mathrm{MVE}}_i = \sum_{i=0}^{H} \frac{\tilde{w}_i}{\sum_j \tilde{w}_j}\mathcal{T}^\mu_i, \qquad \tilde{w}_i^{-1} = \mathcal{T}^{\sigma^2}_i$$

Results