Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion

presented by Jason TOKO

背景与动机

- Model-free方法需要大量样本，实际应用中数据采集开销过大。
- Model-based方法可通过各种途径来提高采样效率（略），但是存在一个缺陷：
  - 在复杂、有噪声的环境下，难以学习到准确的模型，不准确的模型会导致策略学习错误，影响算法表现。

- 动机：
  - 需要找到一种model-free和model-based结合的方法，使得模型的误差不会影响算法表现。

MVE

- Model-based Value Expansion（MVE）依托于其他强化学习算法（如DDPG等），是算法中的一环。
- 核心思想：在学习的模型上roll out，将得到轨迹用于计算Target
- TD Target：

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

- MVE Target：

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

MVE

- 模型组成：
  - Transition function：$\hat{T}_\xi(s, a)$
  - Termination function：$\hat{d}_\xi(t \mid s)$
  - Reward function：$\hat{r}_\psi(s, a, s')$
- 模型学习：

$$\mathcal{L}_{\xi,\psi} = \mathbb{E}_{(s,a,r,s')}\left[||\hat{T}_\xi(s,a) - s'||^2 + \mathbb{H}\left(d(t \mid s'), \hat{d}_\xi(t \mid \hat{T}_\xi(s,a))\right) + (\hat{r}_\psi(s,a,s') - r)^2\right]$$

  - 其中，$d(t|s') = \begin{cases} 1, & s' \; is \; terminal \; state \\ 0, & otherwise \end{cases}$

MVE

- MVE Target的计算
- 给定模型 $\hat{T}_\xi(s,a)$ $\hat{d}_\xi(t \mid s)$ ，$\hat{r}_\psi(s,a,s')$ ，roll out固定长度H步

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

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

- Loss function： $\mathcal{L}_\theta = \mathbb{E}_{(s,a,r,s')}[(\hat{Q}^\pi_\theta(s,a) - \mathcal{T}_H^{MVE}(r,s'))^2]$

MVE与STEVE

- MVE实际存在的问题：需要调整H来平衡“模型的误差”和“Q值函数的估计误差”，且H的确定依赖于具体任务（task-specific）。
  - 模型准确时，H越大，Target的估算越准确，误差越小，学习效果越好。模型不准确时，H越大，模型误差累积越大。
- 针对MVE存在的问题，Stochastic Ensemble Value Expansion（STEVE）采用了一种特殊的加权方法。

STEVE

- 介系泥门煤油见过的船新版本：
- 设定三组参数：$\boldsymbol{\theta} = \{\theta_1, ..., \theta_L\}, \boldsymbol{\psi} = \{\psi_1, ..., \psi_N\}, \boldsymbol{\xi} = \{\xi_1, ..., \xi_M\}$
- M个模型各自roll out，得到M条长度H的轨迹，$\tau^{\xi_1}, ..., \tau^{\xi_M}$
- M条轨迹分别与L个Q函数、N个奖励函数组合计算$\mathcal{T}_i^{\text{MVE}}$,$(0 \le i \le H)$
- 计算均值$\mathcal{T}_i^{\mu}$ 和方差$\mathcal{T}_i^{\sigma^2}$
- 计算：

$$\mathcal{T}_H^{\text{STEVE}}(r, s') = \sum_{i=0}^{H} \frac{\tilde{w}_i}{\sum_j \tilde{w}_j} \mathcal{T}_i^{\mu}, \qquad \tilde{w}_i^{-1} = \mathcal{T}_i^{\sigma^2}$$

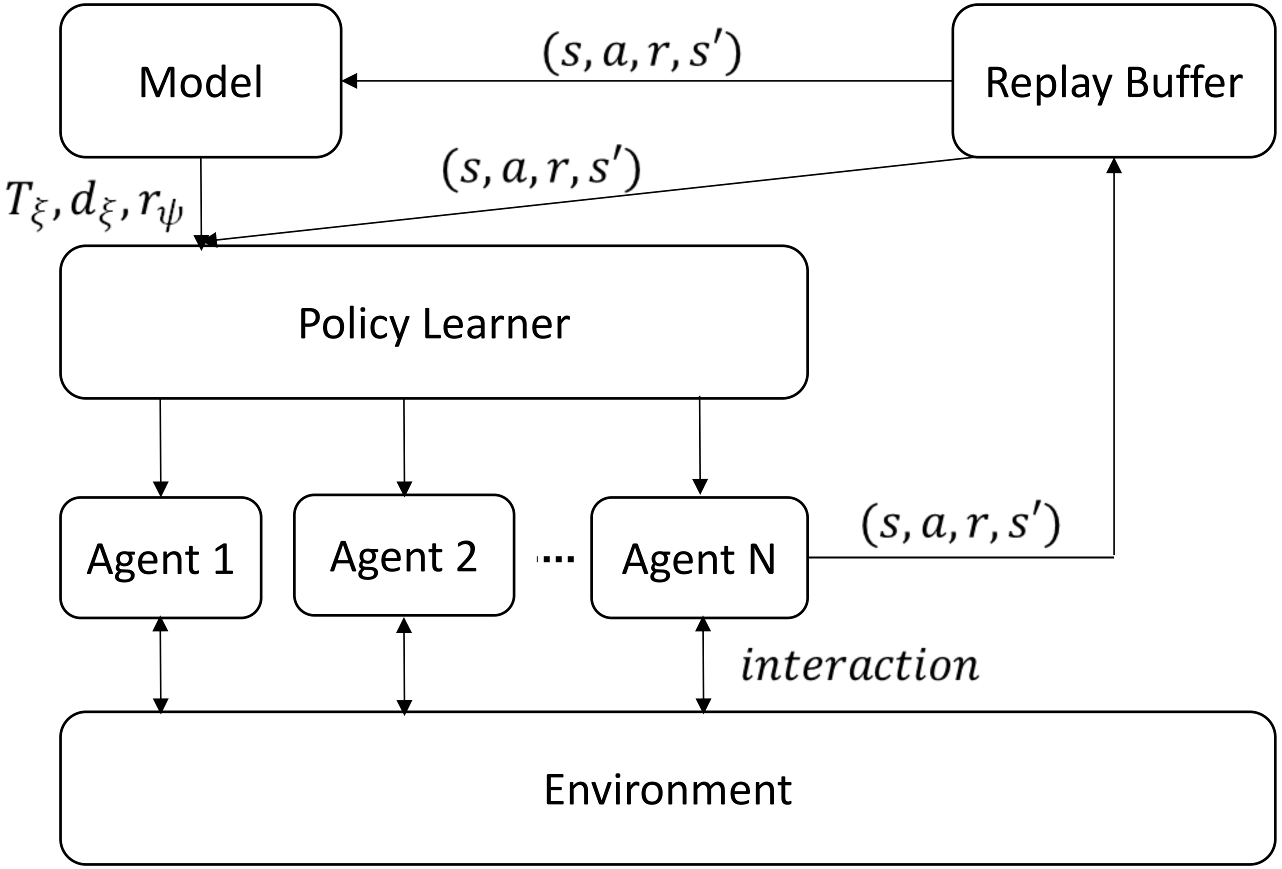
STEVE

- 目标：最小化加权值与真实Q值的均方误差（玄学推导）

$$\mathbb{E}\left[\left(\sum_{i=0}^{H} w_i \mathcal{T}_i^{\text{MVE}} - Q^{\pi}(s,a)\right)^2\right] = \text{Bias}\left(\sum_i w_i \mathcal{T}_i^{\text{MVE}}\right)^2 + \text{Var}\left(\sum_i w_i \mathcal{T}_i^{\text{MVE}}\right)$$
$$\approx \text{Bias}\left(\sum_i w_i \mathcal{T}_i^{\text{MVE}}\right)^2 + \sum_i w_i^2 \, \text{Var}(\mathcal{T}_i^{\text{MVE}}),$$

- 对于 $\sum_i w_i^2 \, \text{Var}(\mathcal{T}_i^{\text{MVE}})$ ，设定$w_i = \frac{Var(\mathcal{T}_i^{MVE})^{-1}}{\sum_j Var(\mathcal{T}_j^{MVE})^{-1}}$，可取得最小值。

STEVE实现



Model
(s, a, r, s')
Replay Buffer
(s, a, r, s')
T_ξ, d_ξ, r_ψ
Policy Learner
Agent 1
Agent 2
···
Agent N
(s, a, r, s')
$interaction$
Environment