ETH zürich

Foundations of Reinforcement Learning - Spring 2024
Project nr 9

Model-Based RL as Stackelberg Games: Don't Be Fooled by Your Own Model

Nils Cremer
Kacper Ozieblowski
Yanick Zengaffinen

Supervisor: Vinzenz Thoma

1 Background

Model-based RL enables efficient learning and can be framed as a Stackelberg game [1]. In [2] they propose an algorithm to learn Stackelberg equilibria (in simple iterated matrix games).

Model Based RL as Stackelberg Game [1]

Environment M, Model $\hat{M}$, Policy π

Policy as Leader: (PAL) $\max_{\pi} \left\{ J(\pi, \widehat{M}^{\pi}) \ s.t.\ \widehat{M}^{\pi} \in \arg\min_{\widehat{M}} \ell(\widehat{M}, \mu_M^{\pi}) \right\}$

Model as Leader: (MAL) $\min_{\widehat{M}} \left\{ \ell(\widehat{M}, \mu_M^{\pi_{\widehat{M}}}) \ s.t.\ \pi_{\widehat{M}} \in \arg\max_{\pi} J(\pi, \widehat{M}) \right\}$

where

$\mu_M^{\pi} = \frac{1}{T} \sum_{t=0}^{T} P(s_t = s, a_t = a)$ and $\ell(\hat{M}, \mu) = \mathbb{E}_{(s,a)\sim\mu} \left[D_{KL}(P(\cdot|s,a), \hat{P}(\cdot|s,a)) \right]$

Theorem 1: Given policy π and model $\hat{M}$, such that

$\ell(\hat{M}, \mu_M^{\pi}) \leq \epsilon_M$ and $J(\pi, \hat{M}) \geq J(\pi', \hat{M}) - \epsilon_\pi \ \forall \pi'$

then for any optimal policy π^*

$$J(\pi^*, M) - J(\pi, M) \leq O(\epsilon_\pi + \frac{\sqrt{\epsilon_M}}{(1-\gamma)^2} + \frac{1}{1-\gamma} D_{TV}(\mu_M^{\pi^*}, \mu_{\hat{M}}^{\pi^*}))$$

Learning a Stackelberg Equilibrium [2]

Contextualized Follower

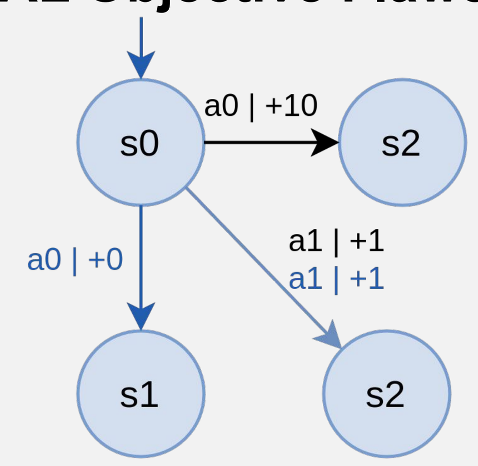
for each pre-training iteration **do** **// Follower Pretraining**
  sample random leader
  query leader -> leader description ω
  train follower given leader description ω (e.g. PPO)
for each training iteration **do** **// Leader Training**
  query leader -> leader description ω
  get best-response follower using ω
  train leader

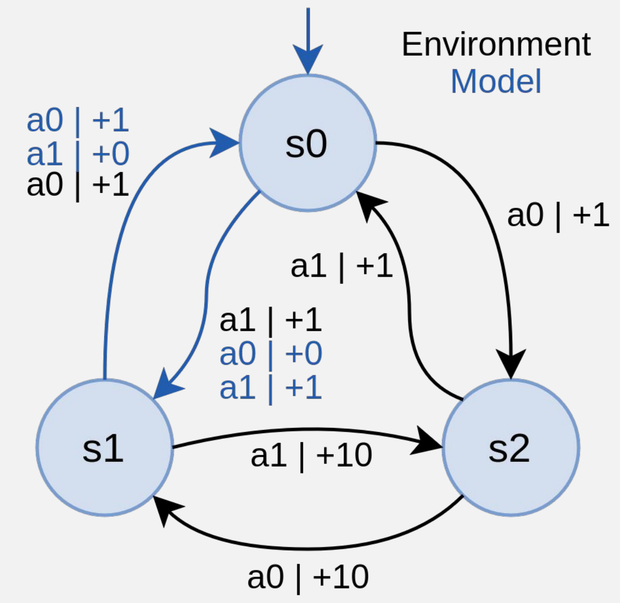
Inner-Outer Loop

for each leader iteration **do** **// Outer Loop**
  for each follower iteration **do** **// Inner Loop**
    train follower
  train leader

2 Theory

MAL Objective Flawed





MDP can be designed such that the final policy performs arbitrarily bad. We can prove that PAL does not suffer from this.

Theorem (ours): Given an approximate solution π to **PAL** with approximate best-responding models $\hat{M}(\pi)$, such that

$\ell(\hat{M}(\pi), \mu_M^{\pi}) \leq \epsilon_M, \forall \pi$ and $J(\pi, \hat{M}(\pi)) \geq \sup_{\pi'} J(\pi', \hat{M}(\pi')) - \epsilon_\pi$

then $J(\pi^*, M) - J(\pi, M) \leq \epsilon_\pi + \frac{4\gamma\sqrt{\epsilon_M} R_{max}}{(1-\gamma)^2}$

where π^* is an optimal policy and R_{max} is a bound on the absolute values of all rewards.

Fixing MAL

(MAL + Agent Reward)

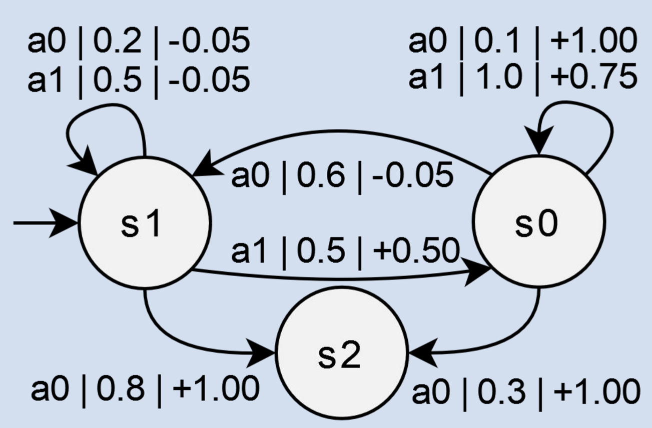
$$\min_{\hat{M}} \left\{ \alpha \cdot \ell(\hat{M}, \mu_M^{\pi_{\hat{M}}}) - (1-\alpha) \cdot J(\pi_{\hat{M}}, \hat{M}) \ s.t.\ \pi_{\hat{M}} \in \arg\max_{\pi} J(\pi, \hat{M}) \right\}$$

(MAL + Random Noise)

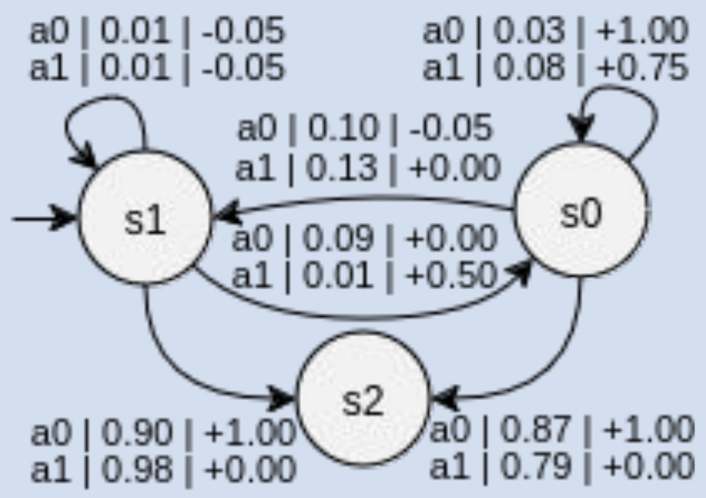
$$\min_{\hat{M}} \left\{ \ell\left(\hat{M}, \alpha\mu_M^{\pi_{\hat{M}}} + (1-\alpha)\mathbb{E}_{\pi' \in \Pi}\left[\mu_M^{\pi'}\right]\right) \ s.t.\ \pi_{\hat{M}} \in \arg\max_{\pi} J(\pi, \hat{M}) \right\}$$

3 Experiments

Environment



MAL - Final Model



4 Results

Problem Formulation	Avg. Ep. Reward
MAL	00.97 ± 00.05
MAL + Agent Reward	36.93 ± 00.57
MAL + Random Noise	36.74 ± 00.67
PAL	36.77 ± 00.64

The average reward is the mean of 5 experiments.

5 Discussion

Model as Leader

- Can fail because the model is not incentivized to maximize the reward, thus it can hide parts of the environment
- Ergodicity and determinism are not strong enough
- Good mixing $D_{TV}(\mu_M^{\pi^*}, \mu_{\hat{M}}^{\pi^*}) \to 0$ can solve the issue [1]
- MAL with changed leader objective can work
- Learns model only for best responding policies

Policy as Leader

- Provably aligned with the actual goal
- Reduces to normal model based RL in MDPs

6 Future Work

- Sample efficiency of PAL (e.g. on POMDPs)
- Theoretical guarantees for our MAL formulations
- Complex MDPs
- Thoroughly investigate why [1] worked
- Continuous environments

References

[1] Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A game theoretic framework for model based reinforcement learning. CoRR, abs/2004.07804, 2020. URL https://arxiv.org/abs/2004.07804.

[2] Matthias Gerstgrasser and David C. Parkes. Oracles followers: Stackelberg equilibria in deep multi-agent reinforcement learning, 2023. URL https://proceedings.mlr.press/v202/gerstgrasser23a/gerstgrasser23a.pdf