Model-Based Reinforcement Learning with Multinomial Logistic Function Approximation

Taehyun Hwang & Min-hwan Oh

Seoul National University

AAAI 2023

More Recent Results in RL

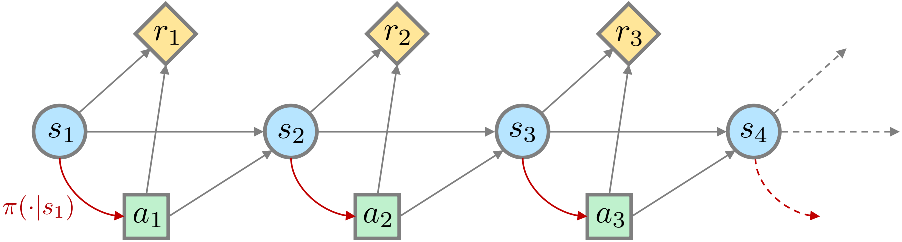
RL with function approximation has made significant advances in empirical studies. However,

More Recent Results in RL

RL with function approximation has made significant advances in empirical studies. However,

- Theoretical understanding of these methods is still limited
- Most existing theoretical works in RL with function approximation consider linear function approximation
- Trying to close the gap between theory and empirical findings

Markov Decision Processes (MDPs)



A finite-horizon Markov Decision Processes (MDPs), $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, r)$

- $\mathcal{S}$: State space
- $\mathcal{A}$: Set of actions
- H: Length of horizon
- $P = \{\mathbb{P}(\cdot \mid s, a) \mid (s, a) \in \mathcal{S} \times \mathcal{A}\}$: Collection of transition probability
- r: Reward function

Value Functions & Performance Measure

Value function of a policy π

$$Q_h^\pi(s,a) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r\left(s_{h'}, \pi(s_{h'}, h')\right) \mid s_h = s, a_h = a\right]$$

$$V_h^\pi(s) = Q^\pi(s, \pi_h(s))$$

Optimal value function & policy

$$Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$$

$$\pi_h^*(s) := \operatorname*{argmax}_{a\in\mathcal{A}} Q_h^*(s,a)$$

Performance measure

$$\textbf{Regret}_\pi(K) := \sum_{k=1}^{K}(V_1^* - V_1^{\pi_k})(s_{k,1})$$

- K: total number of episodes
- $T = KH$: total number of timesteps

Existing Works: Tabular Methods

A large number of works both on model-based and model-free methods

- Model-based: (Jaksch et al., 2010; Osband and Roy, 2014; Azar et al., 2017; Dann et al., 2017; Agrawal and Jia, 2017; Ouyang et al., 2017)
- Model-free: (Jin et al., 2018; Osband et al., 2019; Russo, 2019; Zhang et al., 2020, 2021)

Model-based and model-free methods can achieve $\widetilde{\mathcal{O}}(H\sqrt{SAT})$ regret.

- *optimal up to logarithmic factors* (Jin et al., 2018; Zhang et al., 2020).
- $S = |\mathcal{S}|$: the total number of states
- $A = |\mathcal{A}|$: the total number of actions

Existing Works: Tabular Methods

A large number of works both on model-based and model-free methods

- Model-based: (Jaksch et al., 2010; Osband and Roy, 2014; Azar et al., 2017; Dann et al., 2017; Agrawal and Jia, 2017; Ouyang et al., 2017)
- Model-free: (Jin et al., 2018; Osband et al., 2019; Russo, 2019; Zhang et al., 2020, 2021)

Model-based and model-free methods can achieve $\widetilde{\mathcal{O}}(H\sqrt{SAT})$ regret.

- *optimal up to logarithmic factors* (Jin et al., 2018; Zhang et al., 2020).
- $S = |\mathcal{S}|$: the total number of states
- $A = |\mathcal{A}|$: the total number of actions

But these methods do not perform well with large $\mathcal{S}$ & $\mathcal{A}$.

- No generalization across states (or actions)

Existing Works: Linear Function Approximation

Low-rank linear MDPs (Model-free): $\mathbb{P}(s' \mid s, a) = \langle \phi(s, a), \boldsymbol{\mu}^*(s') \rangle$

- Optimism: `LSVI-UCB` $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ (Jin et al., 2020)
- Randomization: `OPT-RLSVI` $\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$ (Zanette et al., 2020)

Existing Works: Linear Function Approximation

Low-rank linear MDPs (Model-free): $\mathbb{P}(s' \mid s, a) = \langle \phi(s, a), \boldsymbol{\mu}^*(s') \rangle$

- Optimism: `LSVI-UCB` $\widetilde{\mathcal{O}}(d^{3/2} H^{3/2} \sqrt{T})$ (Jin et al., 2020)
- Randomization: `OPT-RLSVI` $\widetilde{\mathcal{O}}(d^2 H^2 \sqrt{T})$ (Zanette et al., 2020)

Bilinear transition model (Model-based): $\mathbb{P}(s' \mid s, a) = \phi(s, a)^\top M^* \psi(s')$

- Optimism: `UC-MatrixRL` $\widetilde{\mathcal{O}}(dH^2 \sqrt{T})$ (Yang and Wang, 2020)

Existing Works: Linear Function Approximation

Low-rank linear MDPs (Model-free): $\mathbb{P}(s' \mid s, a) = \langle \phi(s, a), \boldsymbol{\mu}^*(s') \rangle$

- Optimism: `LSVI-UCB` $\widetilde{\mathcal{O}}(d^{3/2} H^{3/2} \sqrt{T})$ (Jin et al., 2020)
- Randomization: `OPT-RLSVI` $\widetilde{\mathcal{O}}(d^2 H^2 \sqrt{T})$ (Zanette et al., 2020)

Bilinear transition model (Model-based): $\mathbb{P}(s' \mid s, a) = \phi(s, a)^\top M^* \psi(s')$

- Optimism: `UC-MatrixRL` $\widetilde{\mathcal{O}}(dH^2 \sqrt{T})$ (Yang and Wang, 2020)

Linear mixture models (Model-based) :$\mathbb{P}(s' \mid s, a) = \sum_{j=1}^{d} \theta_j^* \mathbb{P}_j(s' \mid s, a)$

- Optimism: `UCRL-VTR` $\widetilde{\mathcal{O}}(dH^{3/2} \sqrt{T})$ (Ayoub et al., 2020)

Existing Works: Linear Function Approximation

Low-rank linear MDPs (Model-free): $\mathbb{P}(s' \mid s, a) = \langle \phi(s, a), \boldsymbol{\mu}^*(s') \rangle$

- Optimism: `LSVI-UCB` $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ (Jin et al., 2020)
- Randomization: `OPT-RLSVI` $\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$ (Zanette et al., 2020)

Bilinear transition model (Model-based): $\mathbb{P}(s' \mid s, a) = \phi(s, a)^\top M^* \psi(s')$

- Optimism: `UC-MatrixRL` $\widetilde{\mathcal{O}}(dH^2\sqrt{T})$ (Yang and Wang, 2020)

Linear mixture models (Model-based) : $\mathbb{P}(s' \mid s, a) = \sum_{j=1}^{d} \theta_j^* \mathbb{P}_j(s' \mid s, a)$

- Optimism: `UCRL-VTR` $\widetilde{\mathcal{O}}(dH^{3/2}\sqrt{T})$ (Ayoub et al., 2020)

Generalized linear function approximation (Model-free)

- Optimism: `LSVI-UCB with GLM` $\widetilde{\mathcal{O}}(d^{3/2}H\sqrt{T})$ Wang et al. (2021)

  ↳ But, this is not GLM approximation of the transition model

Limitation of Linear Transition Model

Proposition (Limited admissible features)

For an arbitrary set of features, a linear transition model cannot induce a proper probability distribution over next states.

- Difficult to ensure $\sum_{s'} \hat{P}(s' \mid s, a) = 1$

Limitation of Linear Transition Model

Proposition (Limited admissible features)

For an arbitrary set of features, a linear transition model cannot induce a proper probability distribution over next states.

- Difficult to ensure $\sum_{s'} \hat{P}(s' \mid s, a) = 1$

Proposition (Dependence on state space)

`UC-MatrixRL` (Yang and Wang, 2020) based on the linear model has the regret of $\widetilde{\mathcal{O}}(|\mathcal{S}| dH^2 \sqrt{T})$.

- Potentially leading to serious deterioration of the performances

Multinomial Logistic Transition model

Motivation

- State transition in MDP is essentially categorical distribution.
- Multinomial Logistic (MNL) model is a natural way of modeling a categorical distribution. Works for any set of features.

Multinomial Logistic Transition model

Motivation

- State transition in MDP is essentially categorical distribution.
- Multinomial Logistic (MNL) model is a natural way of modeling a categorical distribution. Works for any set of features.

MNL Transition model

$$P(s' \mid s, a) = \frac{\exp(\varphi(s, a, s')^\top \theta^*)}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\varphi(s, a, \widetilde{s})^\top \theta^*)}$$

- $\varphi(s, a, s') \in \mathbb{R}^d$: given feature vector
- $\theta^* \in \mathbb{R}^d$: Unknown transition core parameter
- $\mathcal{S}_{s,a} := \{s' \in \mathcal{S} : P(s' \mid s, a) \neq 0\}$: set of reachable states from (s, a)

Multinomial Logistic Transition model

Motivation

- State transition in MDP is essentially categorical distribution.
- Multinomial Logistic (MNL) model is a natural way of modeling a categorical distribution. Works for any set of features.

MNL Transition model

$$P(s' \mid s, a) = \frac{\exp(\varphi(s, a, s')^\top \theta^*)}{\sum_{\widetilde{s} \in \mathcal{S}_{s,a}} \exp(\varphi(s, a, \widetilde{s})^\top \theta^*)}$$

- $\varphi(s, a, s') \in \mathbb{R}^d$: given feature vector
- $\theta^* \in \mathbb{R}^d$: Unknown transition core parameter
- $\mathcal{S}_{s,a} := \{s' \in \mathcal{S} : P(s' \mid s, a) \neq 0\}$: set of reachable states from (s, a)

Can we design a provably efficient RL algorithm for the multinomial logistic transition model?

Upper Confidence Model-based RL for MNL

Algorithm Upper Confidence Model-based RL for MNL (UCRL-MNL)

Initialize $A_1 = \lambda I_d$, $\hat{\theta}_1 = \mathbf{0} \in \mathbb{R}^d$
for episode $k = 1, ..., K$ **do**
  Construct optimistic value functions for $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$

$$\hat{Q}_{k,H+1}(s,a) = 0 \quad \text{and} \quad \hat{V}_{k,h}(s) = \min\{\max_a \hat{Q}_{k,h}(s,a), H\}$$

$$\hat{Q}_{k,h}(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} \frac{\exp(\varphi(s,a,s')^\top \hat{\theta}_k)\hat{V}_{k,h+1}(s')}{\sum_{\tilde{s} \in \mathcal{S}_{s,a}} \exp(\varphi(s,a,\tilde{s})^\top \hat{\theta}_k)} + 2H\beta_k \max_{s' \in \mathcal{S}_{s,a}} \|\varphi(s,a,s')\|_{A_k^{-1}}$$

  for horizon $h = 1, \cdots, H$ **do**
    Select $a_{k,h} = \operatorname{argmax}_{a \in \mathcal{A}} \hat{Q}_{k,h}(s_{k,h}, a)$ and observe $s_{k,h+1}$
  end for
  Update $A_{k+1} = A_k + \sum_{h \leq H} \sum_{s' \in \mathcal{S}_{k,h}} \varphi_{k,h,s'} \varphi_{k,h,s'}^\top$
  Compute $\hat{\theta}_{k+1}$ using the ridge penalized MLE
end for

Regret Analysis

Regularity assumptions (standard in previous literature)

- (Bounded feature & parameter) $\|\varphi(s, a, s')\|_2 \leq L_\varphi$, $\|\theta^*\|_2 \leq L_\theta$
- (Non-singular Fisher info. matrix) $\inf_{\theta \in \mathbb{R}^d} p_{k,h}(s', \theta) p_{k,h}(s'', \theta) > 0$

Regret Analysis

Regularity assumptions (standard in previous literature)

- (Bounded feature & parameter) $\|\varphi(s, a, s')\|_2 \leq L_\varphi$, $\|\theta^*\|_2 \leq L_\theta$
- (Non-singular Fisher info. matrix) $\inf_{\theta \in \mathbb{R}^d} p_{k,h}(s', \theta) p_{k,h}(s'', \theta) > 0$

Lemma (Concentration of $\hat{\theta}_k$ and Optimsim)

For $\beta_k = \widetilde{\mathcal{O}}(\sqrt{d})$, $\theta^* \in \mathcal{C}_k = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_k\|_{A_k} \leq \beta_k\right\}$ and $\hat{Q}_{k,h}(s, a) \geq Q^*_h(s, a)$ with high probability.

- Allows us to work with the *estimated* value function in stead of *unknown optimal* value function

Regret Analysis

Lemma (Value iteration error per step)

$$\hat{Q}_{k,h}(s_{k,h}, a_{k,h}) - \left[r(s_{k,h}, a_{k,h}) + P_h \hat{V}_{k,h+1}(s_{k,h}, a_{k,h}) \right] \leq 2H\beta_k \max_{s' \in \mathcal{S}_{k,h}} \|\varphi_{k,h,s'}\|_{A_k^{-1}}$$

- Hence, the regret under the optimistic policy can be controlled.

Regret Analysis

Lemma (Value iteration error per step)

$$\hat{Q}_{k,h}(s_{k,h}, a_{k,h}) - \left[r(s_{k,h}, a_{k,h}) + P_h \hat{V}_{k,h+1}(s_{k,h}, a_{k,h})\right] \leq 2H\beta_k \max_{s' \in \mathcal{S}_{k,h}} \|\varphi_{k,h,s'}\|_{A_k^{-1}}$$

- Hence, the regret under the optimistic policy can be controlled.

Theorem (Regret of `UCRL-MNL`)

For $\beta_k = \mathcal{O}(\sqrt{d})$, with high probability, the cumulative regret of the `UCRL-MNL` policy π is upper-bounded by

$$\textbf{Regret}_\pi(K) = \widetilde{\mathcal{O}}(d\sqrt{H^3 T} + H\sqrt{T})$$

- Applied to any feature representation of state-action and parameter
- Cumulative regret, sublinear in $T \Rightarrow$ converges to optimality
- First theoretical guarantee for RL with MNL function approximation

Numerical Experiments: RiverSwim Envrionment

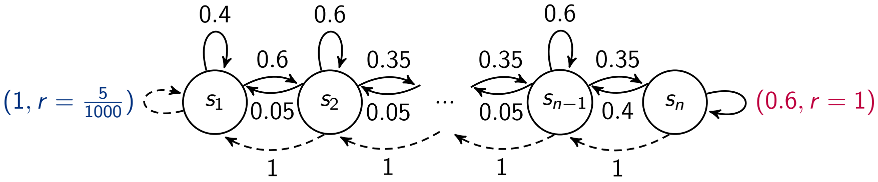
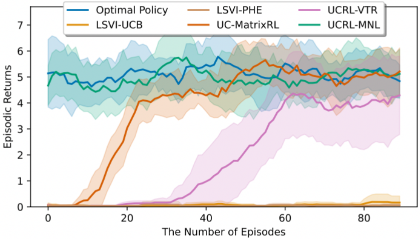


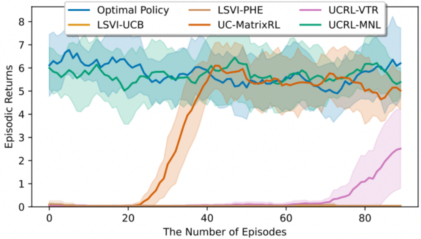
Figure: RiverSwim environment with n states

- Small reward on the left-most state
- Large reward on the right-most state
- Challenging for myopic policies. The environment requires *deeper exploration* to solve.

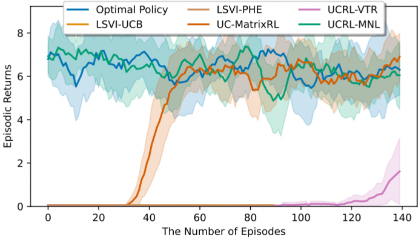
Numerical Experiments: Results



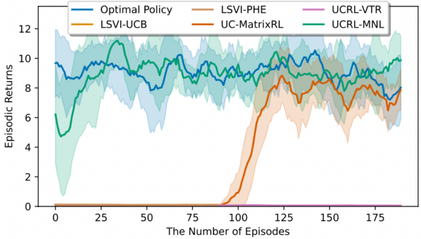
RiverSwim: nState=6, epLen=24



RiverSwim: nState=8, epLen=32



RiverSwim: nState=10, epLen=40



RiverSwim: nState=20, epLen=80

Summary

- MNL function approximation: a new model for provable RL
  - Natural function approximation for transition probabilities
- Propose a RL algorithm, UCRL-MNL, under this new model
  - Achieves the provable guarantees on regret performance
- Superior numerical performances compared to existing methods
- Attains both theoretical and practical efficiency

References I

Agrawal, S. and Jia, R. (2017). Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 5713–5723.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4).

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, pages 4868–4878.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Osband, I. and Roy, B. V. (2014). Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474.

Osband, I., Van Roy, B., Russo, D. J., Wen, Z., et al. (2019). Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017). Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. (2021). Optimism in reinforcement learning with generalized linear function approximation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

References II

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. (2020). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learningvia reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 15198–15207.

Zhang, Z., Zhou, Y., and Ji, X. (2021). Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pages 12653–12662. PMLR.