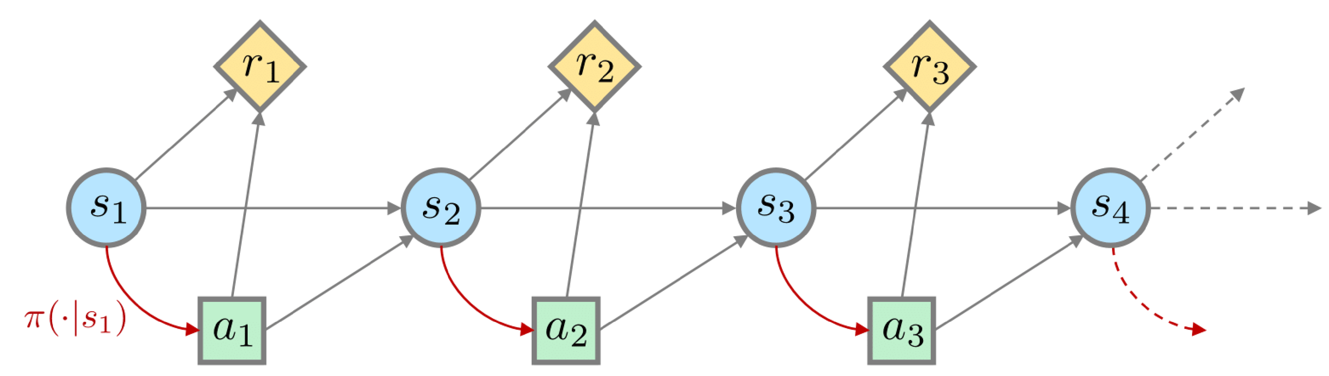
Model-based Reinforcement Learning with Multinomial Function Approximation

Taehyun Hwang[1] & **Min-hwan Oh**[2]

[1]th.hwang@snu.ac.kr, [2]minoh@snu.ac.kr

SEOUL NATIONAL UNIVERSITY

Finite-horizon Markov Decision Processes (MDPs)



- Policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ determines which action the agent takes in state s_h
  - Value function of policy π

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

  - Optimal value function & optimal policy

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

- Goal: Minimize the cumulative regret of π over K episodes

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

Existing RL Algorithms with Function Approx.

- Low-rank MDPs: $P(\cdot \mid s,a) = \langle \phi(s,a), \boldsymbol{\mu}^* \rangle$, $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$: d-dim feature map
  - Optimism: `LSVI-UCB` $\widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}T^{1/2}\right)$ [3]
  - Randomization: `OPT-RLSVI` $\widetilde{\mathcal{O}}\left(d^2H^2T^{1/2}\right)$ [5], `LSVI-PHE` $\widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}T^{1/2}\right)$ [1]
- Bilinear transition model: $P(s' \mid s,a) = \phi(s,a)^\top M^* \psi(s')$, $\psi : \mathcal{S} \to \mathbb{R}^{d'}$, $M^* \in \mathbb{R}^{d \times d'}$
  - Optimism: `UCMatrixRL` $\widetilde{\mathcal{O}}\left(dH^2T^{1/2}\right)$ [4]
- Linear mixture models: $P(ds' \mid s,a) = \sum_{j=1}^{d} \theta_j P_j(ds' \mid s,a)$, P_j: basis transition
  - Optimism: `UCRL-VTR` $\widetilde{\mathcal{O}}\left(dH^{3/2}T^{1/2}\right)$ [2]

Limitation of Linear Transition Model

- For an arbitrary set of features about an MDP, there exist no linear transition model that can induce a proper probability distribution over next states.
  - e.g., Difficult to ensure $\sum_{s'} \hat{P}(s' \mid s,a) = 1$
- `UCMatrixRL` [4] based on the linear model has the regret of $\widetilde{\mathcal{O}}\left(|\mathcal{S}|dH^2T^{1/2}\right)$
  - Leading to serious deterioration in performance for large state space

Multinomial Logistic (MNL) Transition Model

- **MNL transition model:**

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

  - $\varphi(s,a,s') \in \mathbb{R}^d$: 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
- **Can we design a provably efficient algorithm for MNL transition model?**

Upper Confidence model-based RL for MNL (`UCRL-MNL`)

- Ridge penalized maximum likelihood estimation for MNL transition model

$$\hat{\theta}_k = \operatorname*{argmax}_\theta \sum_{\substack{k'<k \\ h \leq H}} \sum_{s' \in \mathcal{S}_{k',h}} y_{k',h}^{s'} \log P_\theta(s' \mid s_{k',h}, a_{k',h}) - \frac{\lambda}{2}\|\theta\|_2^2$$

  - $y_{k,h} = (y_{k,h}^{s'})_{s' \in \mathcal{S}_{k,h}}$ where $y_{k,h}^{s'} = \mathbb{1}(s_{k,h+1} = s')$: Transition response variable
  - $\lambda > 0$: Regularization parameter
- UCB-based optimistic value function

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

  - $\beta_k = \mathcal{O}(\sqrt{d})$: Confidence radius
  - $A_k = \lambda I_d + \sum_{\substack{k'<k \\ h \leq H}} \sum_{s' \in \mathcal{S}_{k',h}} \varphi_{k',h,s'} \varphi_{k',h,s'}^\top$: Gram matrix
- For each episode $k = 1, \ldots, K$:
  1. Construct the optimistic value function $\hat{Q}_{k,h}(s,a)$ for $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$
  2. for $h = 1, \ldots, H$, select $a_{k,h} = \operatorname{argmax}_a \hat{Q}_{k,h}(s_{k,h}, a)$ and observe $s_{k,h+1}$
  3. Update A_{k+1} and compute $\hat{\theta}_{k+1}$

Regret Analysis for `UCRL-MNL`

- Regularity assumptions (standard in previous literature)
  1. (Bounded feature & parameter) $\|\varphi(s,a,s')\|_2 \leq L_\varphi$, $\|\theta_*\|_2 \leq L_\theta$
  2. (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 $\beta_k = \mathcal{O}(\sqrt{d})$, $\theta^* \in \mathcal{C}_k = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_k\|_{A_k} \leq \beta_k\right\}$ and $Q_h^*(s,a) \leq \hat{Q}_{k,h}(s,a)$ with high probability.

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 UCB policy can be controlled.

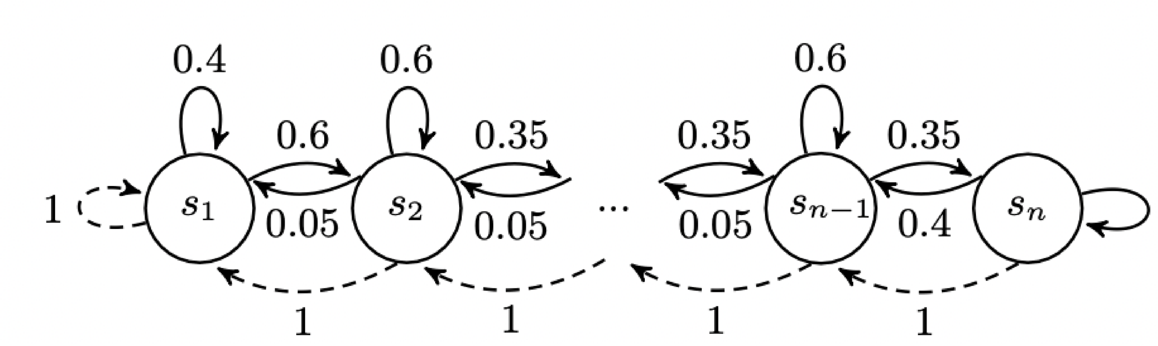
Theorem (Regret of `UCRL-MNL`)

The regret of `UCRL-MNL` is bounded by $\mathbf{Regret}_\pi(K) = \widetilde{\mathcal{O}}(d\sqrt{H^3T})$

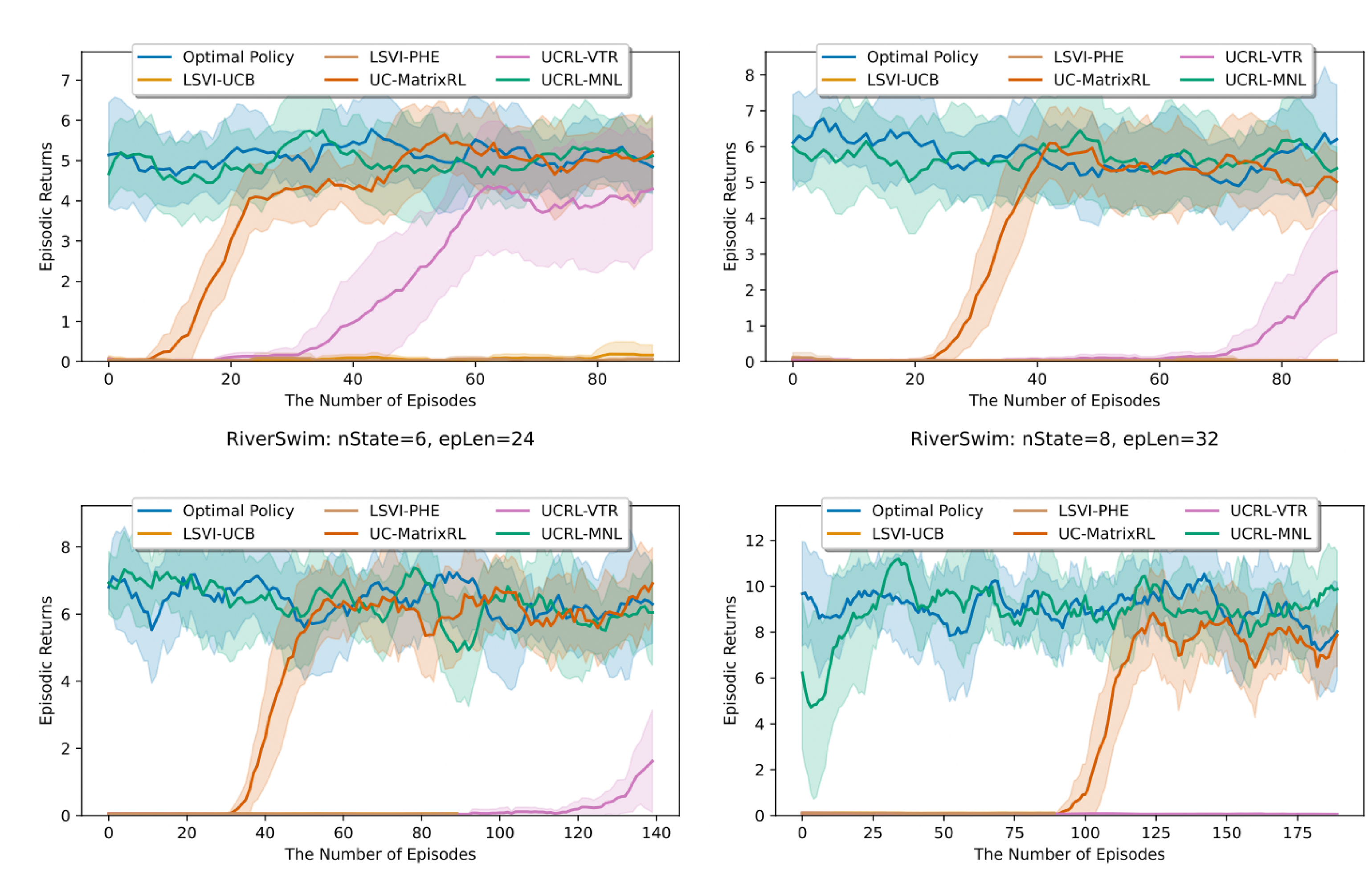
- Applies to any feature representation of state-action and parameter
- Sublinear regret in total timesteps $T = KH \to$ converges to optimality
- First theoretical guarantee for RL with MNL function approximation

Numerical Experiments

- RiverSwim environment with n states ($n = 6, 8, 10, 20$)
  - The environment requires deeper exploration to reach optimality.



- Comparison with provable RL algorithms with function approximation
  - `UCRL-MNL` outperforms the existing algorithms by significant margins.



References

[1] H. Ishfaq, Q. Cui, V. Nguyen, A. Ayoub, Z. Yang, Z. Wang, D. Precup, and L. Yang, *Randomized exploration in reinforcement learning with general value function approximation*, in International Conference on Machine Learning, vol. 139, PMLR, 2021, pp. 4607–4616.

[2] Z. Jia, L. Yang, C. Szepesvari, and M. Wang, *Model-based reinforcement learning with value-targeted regression*, in Learning for Dynamics and Control, PMLR, 2020, pp. 666–686.

[3] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, *Provably efficient reinforcement learning with linear function approximation*, in Conference on Learning Theory, PMLR, 2020, pp. 2137–2143.

[4] L. Yang and M. Wang, *Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound*, in International Conference on Machine Learning, PMLR, 2020, pp. 10746–10756.

[5] A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric, *Frequentist regret bounds for randomized least-squares value iteration*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1954–1964.