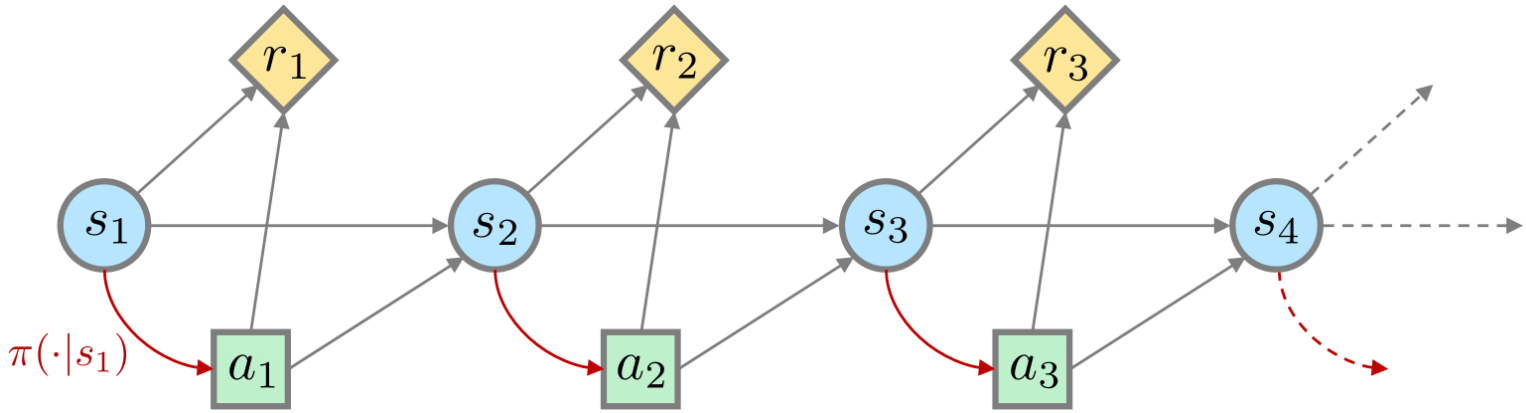
Randomized Exploration for Reinforcement Learning with Multinomial Logistic Function Approximation

Wooseong Cho Taehyun Hwang Joongkyu Lee Min-hwan Oh

Seoul National University

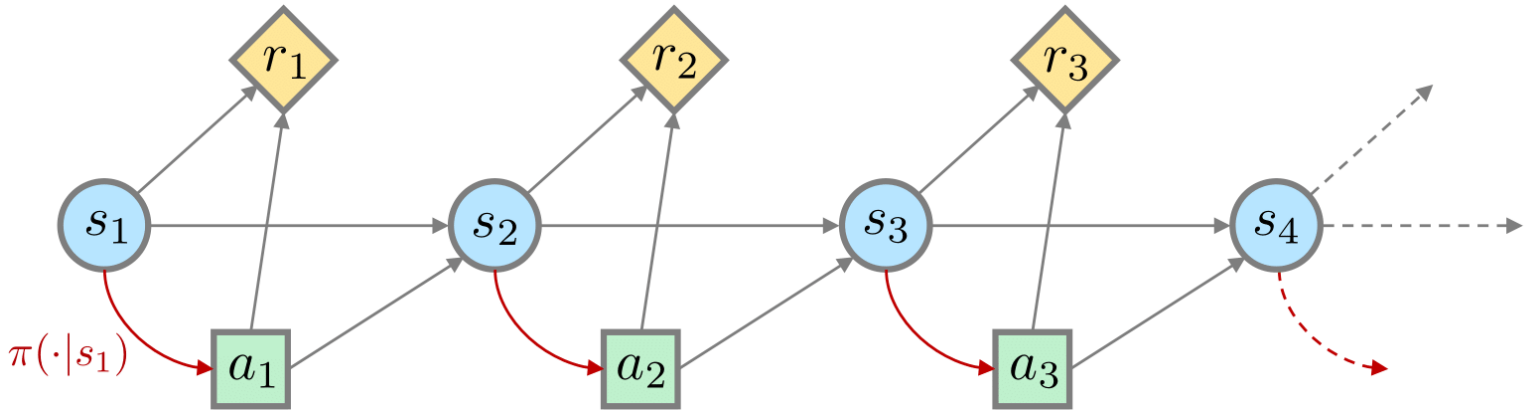
Markov Decision Processes



Episodic *Markov decision processes* (MDPs), $\mathcal{M}\left(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, r\right)$

- $\mathcal{S}$: state space, $\mathcal{A}$: action space
- H: length of each episode
- $\{P_h\}_{h=1}^H$: collection of transition probabilities
- r: reward function

Markov Decision Processes



Episodic *Markov decision processes* (MDPs), $\mathcal{M}\left(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, r\right)$

- $\mathcal{S}$: state space, $\mathcal{A}$: action space
- H: length of each episode
- $\{P_h\}_{h=1}^H$: collection of transition probabilities
- r: reward function

Goal: Maximizing sum of rewards $\equiv$ Minimizing cumulative regret

$$\text{Regret}_\pi(K) := \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)$$

- K: total number of episode

Markov Decision Processes (MDPs)

Low-rank linear MDPs (Jin et al., 2020)

$$P_h(s' \mid s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\mu}_h(s') \rangle$$

- $\boldsymbol{\phi}(s, a) \in \mathbb{R}^d$: feature vector of (s, a)
- $\boldsymbol{\mu}_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$: *unknown* (signed) measure

Markov Decision Processes (MDPs)

Low-rank linear MDPs (Jin et al., 2020)

$$P_h(s' \mid s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\mu}_h(s') \rangle$$

- $\boldsymbol{\phi}(s, a) \in \mathbb{R}^d$: feature vector of (s, a)
- $\boldsymbol{\mu}_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$: *unknown* (signed) measure

Multinomial Logistic (MNL) MDPs (Hwang and Oh, 2023)

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

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

Main Contributions

Algorithm	Model-based	Transition model	Exploration	Regret
LSVI-UCB (Jin et al., 2020)	✗	Linear	UCB	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
OPT-RLSVI (Zanette et al., 2020)	✗	Linear	Randomized	$\widetilde{\mathcal{O}}(d^{2}H^{2}\sqrt{T})$
LSVI-PHE (Ishfaq et al., 2021)	✗	Linear	Randomized	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
UC-MatrixRL (Yang and Wang, 2020)	✓	Linear	UCB	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{2}\sqrt{T})$
UCRL-VTR (Ayoub et al., 2020)	✓	Linear mixture	UCB	$\widetilde{\mathcal{O}}(dH^{\frac{3}{2}}\sqrt{T})$
UCRL-MNL (Hwang and Oh, 2023)	✓	MNL	UCB	$\widetilde{\mathcal{O}}(\kappa^{-1}dH^{\frac{3}{2}}\sqrt{T})$

Can we design a provably efficient and tractable randomized algorithm for MNL-MDPs?

Main Contributions

Algorithm	Model-based	Transition model	Exploration	Regret
LSVI-UCB (Jin et al., 2020)	✗	Linear	UCB	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
OPT-RLSVI (Zanette et al., 2020)	✗	Linear	Randomized	$\widetilde{\mathcal{O}}(d^2H^2\sqrt{T})$
LSVI-PHE (Ishfaq et al., 2021)	✗	Linear	Randomized	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
UC-MatrixRL (Yang and Wang, 2020)	✓	Linear	UCB	$\widetilde{\mathcal{O}}(d^{\frac{3}{2}}H^2\sqrt{T})$
UCRL-VTR (Ayoub et al., 2020)	✓	Linear mixture	UCB	$\widetilde{\mathcal{O}}(dH^{\frac{3}{2}}\sqrt{T})$
UCRL-MNL (Hwang and Oh, 2023)	✓	MNL	UCB	$\widetilde{\mathcal{O}}(\kappa^{-1}dH^{\frac{3}{2}}\sqrt{T})$
RRL-MNL **(this work)**	✓	MNL	Randomized	$\widetilde{\mathcal{O}}(\kappa^{-1}d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T})$
ORRL-MNL **(this work)**	✓	MNL	Randomized	$\widetilde{\mathcal{O}}\left(d^{\frac{3}{2}}H^{\frac{3}{2}}\sqrt{T}+\kappa^{-1}d^2H^2\right)$
UCRL-MNL+ **(this work)**	✓	MNL	UCB	$\widetilde{\mathcal{O}}\left(dH^{\frac{3}{2}}\sqrt{T}+\kappa^{-1}d^2H^2\right)$

- First randomized algorithms for MNL-MDPs
- Frequentist regret analysis *without assuming stochastic optimism*
- Statistically improved algorithms for MNL-MDPs

Algorithm: RRL-MNL

Algorithm Randomized RL for MNL-MDPs (RRL-MNL)

Initialize: $\mathbf{A}_{1,h} = \lambda \mathbf{I}_d, \boldsymbol{\theta}_h^1 = \mathbf{0}_d$ for all $h \in [H]$
for each episode $k = 1, \ldots, K$ **do**
  Observe s_1^k and sample $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{A}_{k,h}^{-1})$ for $m \in [M]$
  Compute

$$Q_h^k(s,a) = r(s,a) + \underbrace{\sum_{s' \in \mathcal{S}_{s,a}} P_{\boldsymbol{\theta}_h^k}(s' \mid s,a) V_{h+1}^k(s')}_{\text{Value induced by estimated model}} + \underbrace{\max_{m \in [M]} \hat{\boldsymbol{\varphi}}_{k,h}(s,a)^\top \boldsymbol{\xi}_{k,h}^{(m)}}_{\text{Randomized bonus}}$$

  for each horizon $h = 1, \ldots, H$ **do**
    Select $a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} Q_h^k(s_h^k, a)$ and observe s_{h+1}^k
    Update $\mathbf{A}_{k+1,h}$ and $\boldsymbol{\theta}_h^{k+1}$
  end for
end for

$\hat{\boldsymbol{\varphi}}_{k,h}(s,a) := \boldsymbol{\varphi}(s,a,\hat{s})$ for $\hat{s} = \operatorname{argmax}_{s'} \|\boldsymbol{\varphi}(s,a,s')\|_{\mathbf{A}_{k,h}^{-1}}$

Regret Bound (RRL-MNL)

Assumptions

- [Boundedness] $\|\boldsymbol{\varphi}(s, a, s')\|_2 \leq L_{\boldsymbol{\varphi}}$, $\|\boldsymbol{\theta}_h^*\|_2 \leq L_{\boldsymbol{\theta}}$
- [Non-singular Fisher info. matrix] $\kappa := \inf_{\|\boldsymbol{\theta}\|_2 \leq L_{\boldsymbol{\theta}}} P_{\boldsymbol{\theta}}(s' \mid s, a) P_{\boldsymbol{\theta}}(\widetilde{s} \mid s, a) > 0$

Regret Bound (RRL-MNL)

Assumptions

- [Boundedness] $\|\boldsymbol{\varphi}(s,a,s')\|_2 \leq L_{\boldsymbol{\varphi}}$, $\|\boldsymbol{\theta}_h^*\|_2 \leq L_{\boldsymbol{\theta}}$
- [Non-singular Fisher info. matrix] $\kappa := \inf_{\|\boldsymbol{\theta}\|_2 \leq L_{\boldsymbol{\theta}}} P_{\boldsymbol{\theta}}(s' \mid s,a) P_{\boldsymbol{\theta}}(\widetilde{s} \mid s,a) > 0$

Theorem (Regret bound of RRL-MNL)

Let $T = KH$. The cumulative regret over K episodes is bounded by

$$\textbf{\textit{Regret}}_{\pi}(K) = \widetilde{\mathcal{O}}\left(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T}\right).$$

Regret Bound (RRL-MNL)

Assumptions

- [Boundedness] $\|\boldsymbol{\varphi}(s,a,s')\|_2 \le L_{\boldsymbol{\varphi}}$, $\|\boldsymbol{\theta}_h^*\|_2 \le L_{\boldsymbol{\theta}}$
- [Non-singular Fisher info. matrix] $\kappa := \inf_{\|\boldsymbol{\theta}\|_2 \le L_{\boldsymbol{\theta}}} P_{\boldsymbol{\theta}}(s' \mid s,a) P_{\boldsymbol{\theta}}(\widetilde{s} \mid s,a) > 0$

Theorem (Regret bound of RRL-MNL)

Let $T = KH$. The cumulative regret over K episodes is bounded by

$$\textbf{\textit{Regret}}_{\pi}(K) = \widetilde{\mathcal{O}}\left(\kappa^{-1} d^{\frac{3}{2}} H^{\frac{3}{2}} \sqrt{T}\right).$$

- We do not assume stochastic optimism, unlike Ishfaq et al. (2021).

Lemma (Stochastic Optimism)

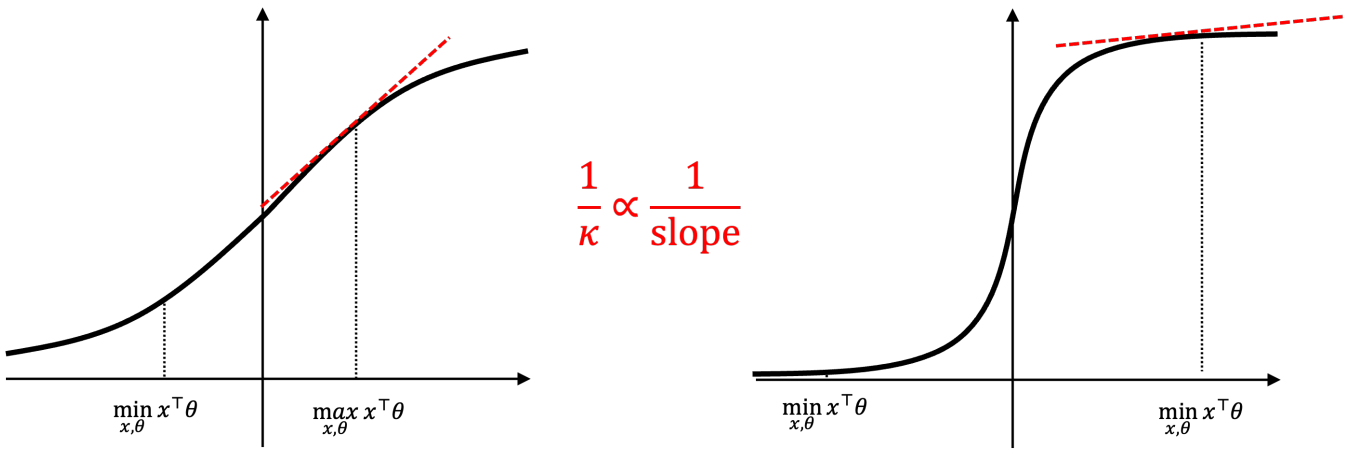
Let Φ be a normal CDF. For $M = \mathcal{O}(\log H)$,

$$\mathbb{P}\left((V_1^k - V_1^*)(s_1^k) \ge 0\right) \ge \Phi(-1)/2.$$

Problem-dependent Constant κ

$$\kappa := \inf_{\|\boldsymbol{\theta}\|_2 \leq L_{\boldsymbol{\theta}}} P_{\boldsymbol{\theta}}(s' \mid s, a) P_{\boldsymbol{\theta}}(\widetilde{s} \mid s, a) > 0$$

- Characterizes the degree of non-linearity of the MNL function



- In the worst case, κ^{-1} can be exponentially large in $|\mathcal{S}_{s,a}|$.

Algorithm: ORRL-MNL

Algorithm Optimistic Randomized RL for MNL-MDPs (ORRL-MNL)

Initialize: $\mathbf{B}_{1,h} = \lambda \mathbf{I}_d, \widetilde{\boldsymbol{\theta}}_h^1 = \mathbf{0}_d$ for all $h \in [H]$
for each episode $k = 1, \ldots, K$ **do**
  Observe s_1^k and sample $\boldsymbol{\xi}_{k,h}^{(m)} \sim \mathcal{N}(\mathbf{0}_d, \sigma_k^2 \mathbf{B}_{k,h}^{-1})$.
  Compute

$$\widetilde{Q}_h^k(s,a) = r(s,a) + \underbrace{\sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) V_{h+1}^k(s')}_{\text{Value induced by estimated model}} + \underbrace{\nu_{k,h}^{\text{rand}}(s,a)}_{\text{Optimistic randomized bonus}}$$

  for each horizon $h = 1, \ldots, H$ **do**
    Select $a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} \widetilde{Q}_h^k(s_h^k, a)$ and observe s_{h+1}^k
    Update $\widetilde{\boldsymbol{\theta}}_h^{k+1}$ and $\mathbf{B}_{k+1,h}$.
  end for
end for

$$\nu_{k,h}^{\text{rand}}(s,a) = \sum_{s' \in \mathcal{S}_{s,a}} P_{\widetilde{\boldsymbol{\theta}}_h^k}(s' \mid s,a) \bar{\boldsymbol{\varphi}}_{s,a,s'} (\widetilde{\boldsymbol{\theta}}_h^k)^\top \boldsymbol{\xi}_{k,h}^{s'} + 3H\beta_k^2 \max_{s' \in \mathcal{S}_{s,a}} \|\boldsymbol{\varphi}_{s,a,s'}\|_{\mathbf{B}_{k,h}^{-1}}^2$$

Regret Bound (ORRL-MNL)

Theorem (Regret bound of ORRL-MNL)

Let $T = KH$. The cumulative regret over K episodes is bounded by

$$\textbf{\textit{Regret}}_{\pi}(K) = \widetilde{\mathcal{O}}\left(d^{3/2}H^{3/2}\sqrt{T} + \kappa^{-1}d^{2}H^{2}\right).$$

- Compared to RRL-MNL, leading term does not suffer from κ^{-1}
- Improved dependence on κ^{-1} for randomized exploration RL

Numerical Experiments

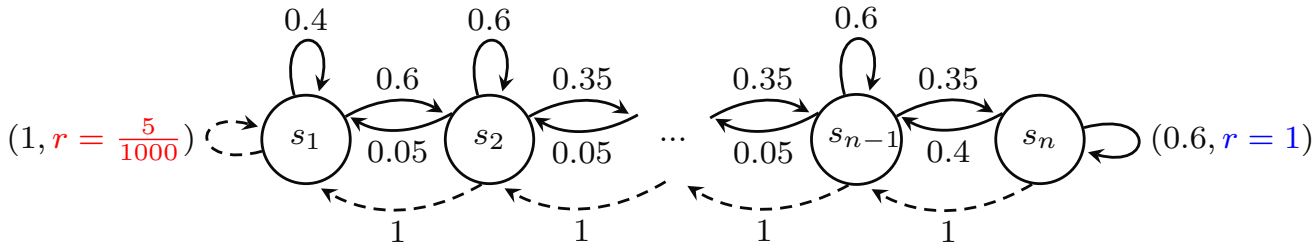
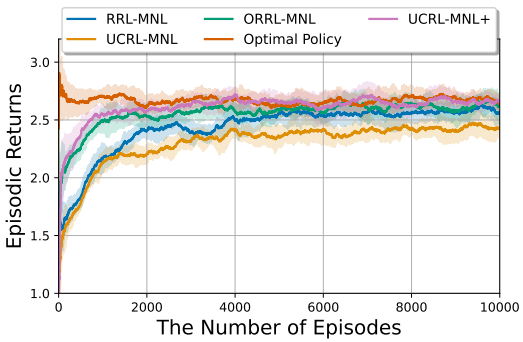
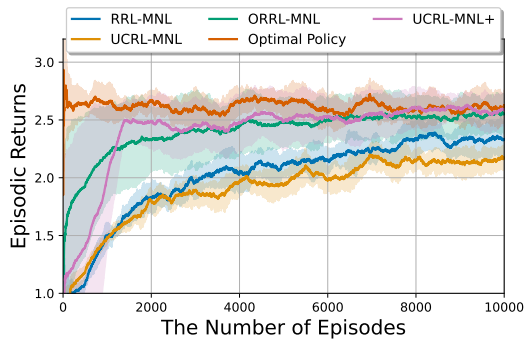


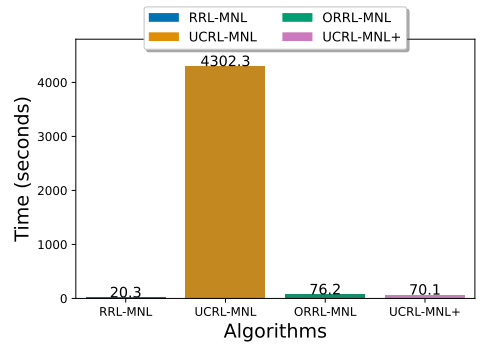
Figure: "RiverSwim" with n states



(a) $S = 4, H = 12$



(b) $S = 8, H = 24$



(c) Runtime (1,000 episodes)

References

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.

Hwang, T. and Oh, M.-h. (2023). Model-based reinforcement learning with multinomial logistic function approximation. In Proceedings of the AAAI conference on artificial intelligence, pages 7971–7979.

Ishfaq, H., Cui, Q., Nguyen, V., Ayoub, A., Yang, Z., Wang, Z., Precup, D., and Yang, L. (2021). Randomized exploration in reinforcement learning with general value function approximation. In International Conference on Machine Learning, volume 139, pages 4607–4616. PMLR, PMLR.

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.

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.