Lecture 12: recent progress on reinforcement learning with function approximation

Dabeen Lee

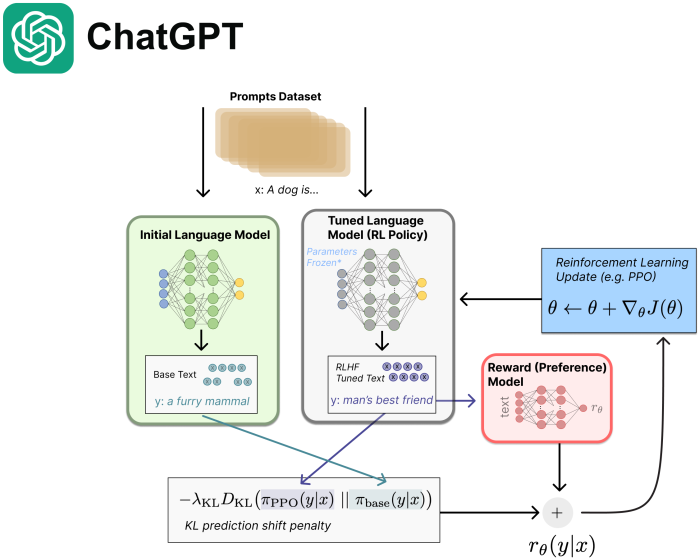
Industrial and Systems Engineering, KAIST

2025 Winter Lecture Series on Combinatorial Optimization

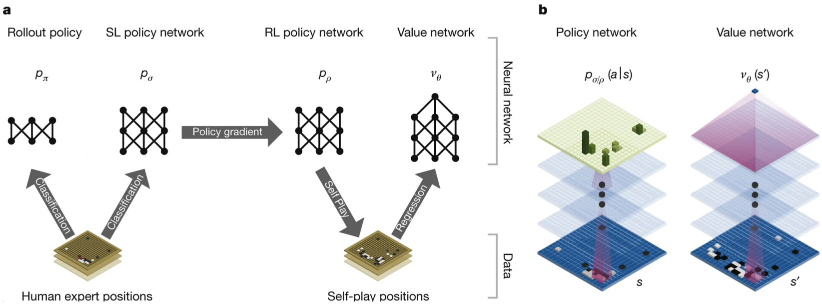
Based on Joint Works with Woojin Chae, Junyeop Kwon, Jaehyun Park (KAIST) &
Kihyuk Hong, Yufan Zhang, Ambuj Tewari (The University of Michigan)

January 17, 2025

Reinforcement Learning for LLM



Reinforcement Learning for AlphaGo



Function Approximation for Reinforcement Learning

- Model the **reward** function, the **transition** kernel, or the **value** function with a function class, e.g., neural networks.
- Applications (of mostly neural function approximation):
  - Atari games [Mnih et al., 2015]
  - Go [Silver et al., 2017]
  - Robotics [Kober et al., 2013]
  - Autonomous driving [Yurtsever et al., 2020].
- Despite this empirical success, **we lack theoretical understanding of function approximation frameworks.**

Today's Theme

Design and analyze function approximation frameworks and **algorithms** for reinforcement learning with **provable guarantees**.

Outline

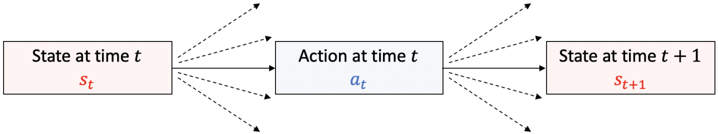
- Markov Decision Process (MDP) (Background)
- Linear Function Approximation for Reinforcement Learning (RL)
- Multinomial Logistic (MNL) Function Approximation for RL

Outline

- Markov Decision Process (MDP) (Background)
- **Linear Function Approximation for Reinforcement Learning (RL)**
- Multinomial Logistic (MNL) Function Approximation for RL

Markov Decision Process (MDP)

Formulation



- $\pi(a \mid s)$: policy, given by the probability of taking action a at state s
- $r(s, a)$: reward from choosing action a at state s
- $\mathbb{P}(s' \mid s, a)$: probability of transitioning to state s' from state s when the chosen action is a.

Markov Decision Process (MDP)

Settings

- Finite-Horizon MDP
- **Infinite-Horizon Average-Reward MDP**
- Infinite-Horizon Discounted-Reward MDP

Markov Decision Process (MDP)

Finite-Horizon MDP

- Fixed initial state (or a fixed distribution of the initial state).
- H: the finite length of the horizon.
- For example, arcade games.

- Basically, run an episode and **reset**.

Markov Decision Process (MDP)

Infinite-Horizon Average-Reward MDP

- **Continue the process without resetting**.
- Start with the initial state s_1.
- Given state s_t in time t, take action a_t and observe the next state s_{t+1}.
- For example, inventory management and financial planning.

- **Average reward** (under policy π):

$$J^\pi(s_1) = \liminf_{T\to\infty} \quad \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t)\right].$$

- **Optimal policy**:

$$\pi^* \in \mathrm{argmax}_\pi \{J^\pi(s_1)\}.$$

Markov Decision Process (MDP)

Infinite-Horizon Discounted-Reward MDP

- Similar to the infinite-horizon average-reward setting.
- **Discounted reward** (under policy π):

$$V^{\pi}(s_1) = \liminf_{T \to \infty} \quad \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t)\right]$$

for some discount factor $\gamma \in (0, 1)$.

Computing Optimal Policies for MDPs

- **If the reward and transition functions are known**, we can **efficiently compute an optimal policy** for both finite- and infinite- horizon MDPs.
- One may use the following frameworks to compute an optimal policy.
  1. Linear programming-based methods.
  2. Value iteration.
  3. Policy iteration.
  4. Policy gradient.

Reinforcement Learning for MDPs

Reinforcement Learning for Infinite-Horizon Average-Reward MDP

- At state s_t for time step t, take action a_t from policy π^t
- Observe $r(s_t, a_t) + \epsilon_t$ (noisy reward) and the next state s_{t+1}.
- Learn the reward function r and the transition function $\mathbb{P}$.
- Update π^t to obtain policy π^{t+1} for time step $t+1$.
- **Total cumulative reward over T steps**:

$$\sum_{t=1}^{T} r\left(s_t, a_t\right).$$

- **Regret**:

$$T \cdot \underbrace{\max_{\pi} \left\{ \liminf_{T \to \infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} r\left(s_t^{\pi}, a_t^{\pi}\right) \right] \right\}}_{\text{optimal average reward}} - \sum_{t=1}^{T} r\left(s_t, a_t\right)$$

Infinite-Horizon Average-Reward MDP

- **Not all MDPs are learnable**!
- Not learnable means that no algorithm can guarantee a sublinear regret.

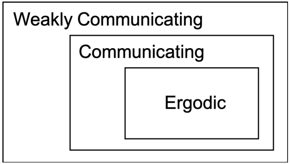
$$\text{Regret}(T) = T \cdot \max_{\pi} \left\{ \liminf_{T \to \infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} r\left(s_t^{\pi}, a_t^{\pi}\right) \right] \right\} - \sum_{t=1}^{T} r\left(s_t, a_t\right) = \underbrace{o(T)}_{\text{sublinear in } T}$$

(sublinear in T: $\text{Regret}(T)/T \to 0$ as $T \to \infty$).

Regret Bounds

Infinite-Horizon Average-Reward MDP

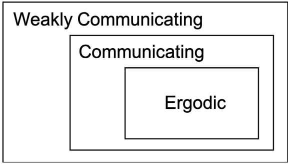
- Recovery from a bad state to a good state should be possible!



- Ergodic MDP: every policy induces a single recurrent class.
- Communicating MDP: one can travel from one state to any other state by a policy.
- Weakly Communicating MDP: state space $\mathcal{S}$ has a set of communicating states, and the others are transient states.

Regret Bounds

Infinite-Horizon Average-Reward Tabular MDP



- Communicating MDP: MDPs with **bounded diameter**, where

$$\underbrace{D}_{\text{diameter of an MDP } M} = \max_{s \neq s' \in \mathcal{S}} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}\left[\underbrace{T(s' \mid M, \pi, s)}_{\text{travel time from } s \text{ to } s'} \right].$$

- Weakly Communicating MDP: MDPs with **bounded span**, where

$$\mathrm{sp}(v^*) = \max_{s \in \mathcal{S}} v^*(s) - \min_{s \in \mathcal{S}} v^*(s)$$

  and v^* is the optimal associated bias function.

- For communicating MDPs, $\mathrm{sp}(v^*) \leq D$.

Regret Bounds

- **Regret** (S: # of states, A: # of actions):

Regret Lower Bound [Jaksch et al., 2010]	$\Omega(\sqrt{\mathrm{sp}(v^*)SAT})$
UCRL2 [Jaksch et al., 2010]	$\tilde{O}(DS\sqrt{AT})$
Thompson Sampling [Agrawal and Jia, 2017]	$\tilde{O}(D\sqrt{SAT})$
REGAL.D [Bartlett and Tewari, 2009]	$\tilde{O}(\mathrm{sp}(v^*)S\sqrt{AT})$
EBF [Zhang and Ji, 2019]	$\tilde{O}(\sqrt{\mathrm{sp}(v^*)SAT})$

General Goal

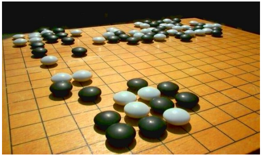
1. Prove a strong lower bound
2. Develop an algorithm whose regret upper bound is close to the lower bound.

RL with Function Approximation

- For infinite-horizon average-reward MDPs, the regret lower bound is

Infinite-horizon [Jaksch et al., 2010]	$\Omega(\sqrt{\mathrm{sp}(v^*)SAT})$

- When S or A is large, the regret is large.
- Atari: 10^{100} states, Go: 10^{170} states.

RL with Function Approximation

- There can be some underlying structures for a given MDP.
- Hence, we may **approximate** the reward function or the transition kernel by a function class, e.g., neural networks.
- Applications (of mostly neural function approximation):
  - Atari games [Mnih et al., 2015]
  - Go [Silver et al., 2017]
  - Robotics [Kober et al., 2013]
  - Autonomous driving [Yurtsever et al., 2020].

Question

Assuming that the reward and transition functions come from a function class, can we guarantee a **smaller regret bound?**

Linear Function Approximation

Linear MDP

- Assume that the transition probability is given by

$$\mathbb{P}(s' \mid s, a) = \varphi(s, a)^\top \mu(s').$$

- $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a **known** feature mapping.
- $\mu : \mathcal{S} \to \mathbb{R}^d$ is an **unknown** parameter function.
- We are interested in the regime where the dimension d is small.
- The task is to learn the unknown parameter function μ.

Linear Function Approximation

Linear Mixture MDP

- Assume that the transition probability is given by

$$\mathbb{P}(s' \mid s, a) = \varphi(s, a, s')^\top \theta.$$

- $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is a **known** feature mapping.
- $\theta \in \mathbb{R}^d$ is an **unknown** parameter.
- We are interested in the regime where the dimension d is small.
- The task is to learn the unknown parameter θ.

Linear Function Approximation

Regret for Infinite-Horizon Linear MDP

Lower Bound [Wu et al., 2022]	$\Omega(d\sqrt{\mathrm{sp}(v^*)T})$
FOPO [Wei et al., 2021] (inefficient)	$\tilde{O}(d^{1.5}\mathrm{sp}(v^*)\sqrt{T})$
OLSVI.FH [Wei et al., 2021]	$\tilde{O}(d^{0.75}\mathrm{sp}(v^*)^{0.5}T^{0.75})$
LOOP [He et al., 2024] (inefficient)	$\tilde{O}(d^{1.5}\mathrm{sp}(v^*)^{1.5}\sqrt{T})$
MDP-EXP2 [Wei et al., 2021] (ergodic)	$\tilde{O}(d\tau_{\mathrm{mix}}^{1.5}\sqrt{T})$

Theorem (Hong, Chae, Zhang, Lee, and Tewari, 2024+)

An efficient *value iteration-based algorithm guarantees that for* weakly communicating linear MDPs *with span* $\mathrm{sp}(v^*)$,

$$Regret = \tilde{O}\left(d^{1.5}\mathrm{sp}(v^*)\sqrt{T}\right).$$

- We achieve the best regret upper bound with an efficient algorithm.

Linear Function Approximation

Corollary (Hong, Chae, Zhang, Lee, and Tewari, 2024+)

There is an efficient model-free *algorithm that guarantees that*

$$\mathit{Regret} = \tilde{O}\left(\mathrm{sp}(v^*) S^{1.5} A^{1.5} \sqrt{T}\right)$$

for weakly communicating MDPs with span $\mathrm{sp}(v^*)$ *where* S *and* A *are the numbers of states and actions.*

- This improves upon the regret upper bound of

$$\text{Regret} = \tilde{O}\left(\mathrm{sp}(v^*) S^5 A^2 \sqrt{T}\right)$$

due to [Zhang and Xie, 2023].

Linear Function Approximation

Regret for Infinite-Horizon Linear Mixture MDP

Lower Bound [Wu et al., 2022]	$\Omega(d\sqrt{\mathrm{sp}(v^*)T})$
UCRL2-VTR [Wu et al., 2022] (communicating)	$\tilde{O}(d\sqrt{DT})$

Theorem (Chae, Hong, Zhang, Tewari and Lee, 2024+)

An efficient value iteration-based algorithm guarantees that for weakly communicating linear mixture MDPs with span $\mathrm{sp}(v^*)$,

$$Regret = \tilde{O}\left(d\sqrt{\mathrm{sp}(v^*)T}\right).$$

- Our algorithm is **minimax optimal**!

Parameter Estimation

1. There exists $\theta^* \in \mathbb{R}^d$ such that

$$\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[V(s') \right] = \varphi(s,a)^\top \theta^*$$

for any value function V and state-action pair (s,a).

2. To obtain θ, we apply ridge regression:

$$\widehat{\theta} = \mathrm{argmin}_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{\tau} \left(\underbrace{\varphi(s_\tau, a_\tau)^\top \theta}_{\text{expected value}} - \underbrace{V(s_{\tau+1})}_{\text{realized value}} \right)^2 .$$

Algorithm for Linear MDPs and Analysis

Value Iteration with Clipping

1. Approximate the average-reward MDP by a discounted-reward MDP.
2. Run value iteration on the disounted-reward MDP:

$$Q_{n+1}(s,a) = \Big[\underbrace{r(s,a) + \gamma \cdot \varphi(s,a)^\top \bar{\theta}_n}_{\text{discounted value iteration}} + \underbrace{\beta \left\|\varphi(s,a)\right\|_{\Sigma^{-1}}}_{\text{bonus term for optimism}} \Big]_{[0,(1-\gamma)^{-1}]} .$$

3. Apply the following clipping operation to **control span**:

$$V_{n+1}(s) = \min \left\{ \max_a Q_{n+1}(s,a), \quad \underbrace{\min_{s'} \max_a Q_{n+1}(s',a) + 2 \cdot \mathrm{sp}(v^*)}_{\text{threshold}} \right\} .$$

Algorithm for Linear MDPs and Analysis

Input: Discounting factor $\gamma \in (0,1)$, regularization $\lambda > 0$, span H, bonus factor β.
Initialize: $t \leftarrow 1$, $k \leftarrow 1$, $t_k \leftarrow 1$, $\Lambda_1 \leftarrow \lambda I$, $\bar{\Lambda}_0 \leftarrow \lambda I$, $Q_t^1(\cdot,\cdot) \leftarrow \frac{1}{1-\gamma}$ for $t \in [T]$.

Receive state s_1.
for *time step* $t = 1, \ldots, T$ **do**
- Take action $a_t = \mathrm{argmax}_a\, Q_t^k(s_t, a)$. Receive reward $r(s_t, a_t)$. Receive next state s_{t+1}.
- $\bar{\Lambda}_t \leftarrow \bar{\Lambda}_{t-1} + \boldsymbol{\varphi}(s_t, a_t)\boldsymbol{\varphi}(s_t, a_t)^T$.
- **if** $2\det(\Lambda_k) < \det(\bar{\Lambda}_t)$ **then**
  - $k \leftarrow k+1$, $t_k \leftarrow t+1$, $\Lambda_k \leftarrow \bar{\Lambda}_t$.
  - `// Run value iteration to plan for remaining` $T - t_k + 1$ `time steps in the new episode.`
  - $\widetilde{V}_{T+1}^k(\cdot) \leftarrow \frac{1}{1-\gamma}$, $V_{T+1}^k(\cdot) \leftarrow \frac{1}{1-\gamma}$.
  - **for** $u = T, T-1, \ldots, t_k$ **do**
    - $\boldsymbol{w}_{u+1}^k \leftarrow \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)(V_{u+1}^k(s_{\tau+1}) - \min_{s'} \widetilde{V}_{u+1}^k(s')))$.
    - $Q_u^k(\cdot,\cdot) \leftarrow \Big(r(\cdot,\cdot) + \gamma(\langle \boldsymbol{\varphi}(\cdot,\cdot), \boldsymbol{w}_{u+1}^k\rangle + \min_{s'} \widetilde{V}_{u+1}^k(s') + \beta \|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda_k^{-1}})\Big) \wedge \frac{1}{1-\gamma}$.
    - $\widetilde{V}_u^k(\cdot) \leftarrow \max_a Q_u^k(\cdot, a)$.
    - $V_u^k(\cdot) \leftarrow \widetilde{V}_u^k(\cdot) \wedge (\min_{s'} \widetilde{V}_u^k(s') + H)$.

Proof for Regret Upper Bound

- The regret function can be decomposed as follows.

Lemma

Regret

$$\leq \underbrace{\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma) V_{t+1}^k(s_{t+1}))}_{(a)} + \gamma \underbrace{\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} (V_{t+1}^k(s_{t+1}) - Q_t^k(s_t, a_t))}_{(b)}$$

$$+ \gamma \underbrace{\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t, a_t)} \left[V_{t+1}^k(s') \right] - V_{t+1}^k(s_{t+1}) \right)}_{(c)}$$

$$+ 4\beta \underbrace{\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t)\|_{\Lambda_t^{-1}}}_{(d)}$$

Proof for Regret Upper Bound

- Term (a), given by

$$\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma) V_{t+1}^k(s_{t+1})),$$

is due to approximation by the discounted-reward MDP.

Lemma

Let J^* and v^* be the optimal average reward and the optimal bias function, and let V^* be the optimal discounted value function with discount factor $\gamma \in [0,1)$. Then it holds that

$$\max_{s \in \mathcal{S}} |J^* - (1-\gamma) V^*(s)| \leq (1-\gamma)\mathrm{sp}(v^*),$$
$$\mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*).$$

Proof for Regret Upper Bound

- Term (b), given by

$$\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(V_{t+1}^k(s_{t+1}) - Q_t^k(s_t, a_t)\right),$$

can be upper bounded based on

$$V_{t+1}^k(s_{t+1}) \leq \max_a Q_{t+1}^k(s_{t+1}, a) = Q_{t+1}^k(s_{t+1}, a_{t+1}).$$

- This leads to a telescoping sum.

Proof for Regret Upper Bound

- Term (c), given by

$$\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \left(\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \left[V_{t+1}^k(s') \right] - V_{t+1}^k(s_{t+1}) \right),$$

  is bounded based on the covering argument due to [Jin et al., 2020] along with the Azuma-Hoeffding inequality for martingales.

- Term (d), given by

$$\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \|\varphi(s_t, a_t)\|_{\Lambda_t^{-1}},$$

  is bounded based on the self-normalization inequality due to [Abbasi-yadkori et al., 2011].

Algorithm for Linear MDPs and Analysis

Regret for Infinite-Horizon Linear MDP

Lower Bound [Wu et al., 2022]	$\Omega(d\sqrt{\mathrm{sp}(v^*)T})$
FOPO [Wei et al., 2021] (inefficient)	$\tilde{O}(d^{1.5}\mathrm{sp}(v^*)\sqrt{T})$
OLSVI.FH [Wei et al., 2021]	$\tilde{O}(d^{0.75}\mathrm{sp}(v^*)^{0.5}T^{0.75})$
LOOP [He et al., 2024] (inefficient)	$\tilde{O}(d^{1.5}\mathrm{sp}(v^*)^{1.5}\sqrt{T})$
MDP-EXP2 [Wei et al., 2021] (ergodic)	$\tilde{O}(d\tau_{\mathrm{mix}}^{1.5}\sqrt{T})$

Theorem (Hong, Chae, Zhang, Lee, and Tewari, 2024+)

An efficient value iteration-based algorithm guarantees that for weakly communicating linear MDPs with span $\mathrm{sp}(v^*)$,

$$\mathit{Regret} = \tilde{O}\left(d^{1.5}\mathrm{sp}(v^*)\sqrt{T}\right).$$

Algorithm for Linear Mixture MDPs and Analysis

Regret for Infinite-Horizon Linear Mixture MDP

Lower Bound [Wu et al., 2022]	$\Omega(d\sqrt{\mathrm{sp}(v^*)T})$
UCRL2-VTR [Wu et al., 2022] (communicating)	$\tilde{O}(d\sqrt{DT})$

Theorem (Chae, Hong, Zhang, Tewari and Lee, 2024+)

An efficient value iteration-based algorithm guarantees that for weakly communicating linear mixture MDPs with span $\mathrm{sp}(v^*)$,

$$\mathit{Regret} = \tilde{O}\left(d\sqrt{\mathrm{sp}(v^*)T}\right).$$

Key Components for Improvement

- For linear mixture MDPs, the **clipped value iteration** procedure converges!
- We apply variance-aware weighted linear regression for estimating θ.

RL with Non-Linear Function Approximation

- Perhaps, the linearity assumption is too restrictive.
- It is not always clear how to impose $0 \leq \mathbb{P}(s' \mid s, a) \leq 1$ for the linear case.
- The underlying model function can be non-linear.

RL with Multinomial Logistic Function Approximation

- [Hwang and Oh, 2023] proposed the multinomial logistic (MNL) function approximation framework.
- Assume that the transition probability is given by

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

- Advantage: the MNL framework is natural for modeling transition probabilities.
- As the linear mixture MDP, $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ is a **known** feature mapping.
- Moreover, $\theta^* \in \mathbb{R}^d$ is an **unknown** parameter.
- Again, we are interested in the regime where the dimension d is small.

RL with Multinomial Logistic Function Approximation

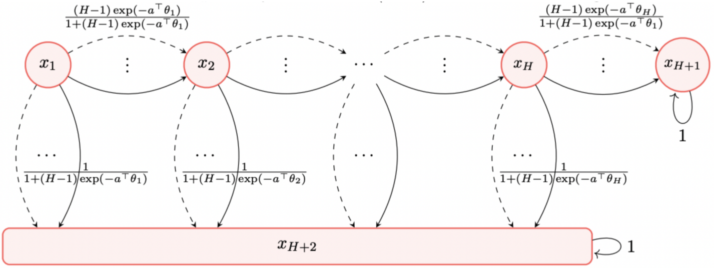
Regret Bounds for MNL transitions (Finite-Horizon)

UCRL-MNL-LL+ [Cho et al., 2024]	$\tilde{O}(dH^2\sqrt{K})$
Lower Bound [**Our Result: Park, Kwon, and Lee, 2024+**]	$\Omega(dH^{1.5}\sqrt{K})$

Regret Bounds for MNL transitions (Infinite-Horizon)

UCMNLK [**Our Result: Park, Kwon, and Lee, 2024+**]	$\tilde{O}(d\,\mathrm{sp}(v^*)\sqrt{T})$
Lower Bound [**Our Result: Park, Kwon, and Lee, 2024+**]	$\Omega(d\sqrt{\mathrm{sp}(v^*)T})$

Finite-Horizon Lower Bound



Theorem (Park, Kwon, and Lee, 2024+)

There is an MDP M with $K \geq \{(d-1)^2 H/2, H^3(d-1)^2/32\}$, $d \geq 2$, and $H \geq 3$ for which any algorithm $\mathfrak{A}$ incurs

$$\text{Regret} \geq \frac{(d-1)H^{1.5}\sqrt{K}}{480\sqrt{2}} = \boldsymbol{\Omega(dH^{1.5}\sqrt{K})}.$$

Theorem (Park, Kwon, and Lee, 2024+)

An efficient value iteration-based algorithm guarantees that for weakly communicating MNL MDPs with span $\mathrm{sp}(v^*)$,

$$\text{Regret} = \tilde{O}\left(d\,\mathrm{sp}(v^*)\sqrt{T}\right).$$

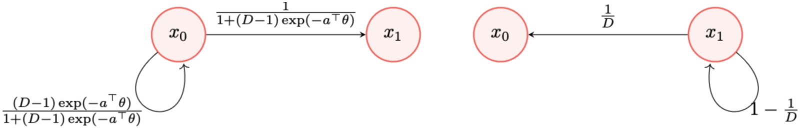
- **Log-likelihood function**:

$$\ell_t(\theta) = \sum_{i=1}^{t-1} \sum_{s' \in \mathcal{S}_{s_i,a_i}} y_{i,s'} \log p_i(s', \theta).$$

- Apply the **online Newton method** of [Zhang and Sugiyama, 2023] to estimate the transition parameter θ^*:

$$\widehat{\theta}_{t+1} = \mathrm{argmin}_{\theta \in \Theta} \left\{ \nabla_\theta(\ell_t(\widehat{\theta}_t))^\top (\theta - \widehat{\theta}_t) + \frac{1}{2\eta} \|\theta - \widehat{\theta}_t\|^2_{\widehat{\Sigma}_t} \right\}.$$

Infinite-Horizon Lower Bound



Theorem (Park, Kwon, and Lee, 2024+)

There is an MDP instance M with $d \geq 2$, $\mathrm{sp}(v^*) \geq 101$, and $T \geq 45(d-1)^2\mathrm{sp}(v^*)$ for which any algorithm $\mathfrak{A}$ incurs

$$\text{Regret} \geq \frac{1}{4050} d\sqrt{DT} = \boldsymbol{\Omega}\left(\boldsymbol{d}\sqrt{\mathbf{sp}(\boldsymbol{v}^*)\boldsymbol{T}}\right).$$

Thank you!

Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3621f1454cacf995530ea53652ddf8fb-Paper.pdf.

P. L. Bartlett and A. Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 35–42, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

W. Cho, T. Hwang, J. Lee, and M. hwan Oh. Randomized exploration for reinforcement learning with multinomial logistic function approximation, 2024.

J. He, H. Zhong, and Z. Yang. Sample-efficient learning of infinite-horizon average-reward MDPs with general function approximation. In *The Twelfth*

International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=fq1wNrC2ai.

T. Hwang and M.-h. Oh. Model-based reinforcement learning with multinomial logistic function approximation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):7971–7979, Jun. 2023. doi: 10.1609/aaai.v37i7.25964. URL https://ojs.aaai.org/index.php/AAAI/article/view/25964.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, aug 2010. ISSN 1532-4435.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013. doi: 10.1177/0278364913495721. URL https://doi.org/10.1177/0278364913495721.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen,

C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236. URL https://doi.org/10.1038/nature14236.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017. doi: 10.1038/nature24270. URL https://doi.org/10.1038/nature24270.

C.-Y. Wei, M. Jafarnia Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3007–3015. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/wei21d.html.

Y. Wu, D. Zhou, and Q. Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3883–3913. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/wu22a.html.

E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8: 58443–58469, 2020. doi: 10.1109/ACCESS.2020.2983149.

Y.-J. Zhang and M. Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29741–29782. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5ef04392708bb2340cb9b7da41225660-Paper-Conference.pdf.

Z. Zhang and X. Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/9e984c108157cea74c894b5cf34efc44-Paper.pdf.

Z. Zhang and Q. Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5476–5477. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/zhang23b.html.