Reinforcement Learning for Infinite-Horizon Average-Reward MDPs with Multinomial Logistic Function Approximation

Dabeen Lee

Industrial and Systems Engineering, KAIST

1st Korean AI Theory Community Workshop: Bandits

Joint work with Jaehyun Park (KAIST)

Outline

- Tabular MDP
- Linear Function Approximation
- General Function Approximation
- Multinomial Logistic Function Approximation
- Our Results

Markov Decision Process (MDP)

Setting

Markov Decision Process (MDP)

Setting

- $\mathcal{S}$: finite state space with $|\mathcal{S}| = S$.

Markov Decision Process (MDP)

Setting

- $\mathcal{S}$: finite state space with $|\mathcal{S}| = S$.
- $\mathcal{A}$: finite action space with $|\mathcal{A}| = A$.

Markov Decision Process (MDP)

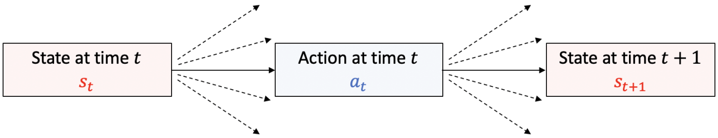
Setting

- $\mathcal{S}$: finite state space with $|\mathcal{S}| = S$.
- $\mathcal{A}$: finite action space with $|\mathcal{A}| = A$.
- $\mathbb{P}(s' \mid s, a)$: probability of transitioning to state s' from state s when the chosen action is a.

Markov Decision Process (MDP)

Setting

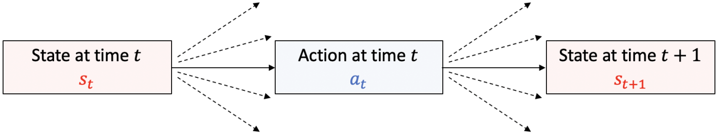
- $\mathcal{S}$: finite state space with $|\mathcal{S}| = S$.
- $\mathcal{A}$: finite action space with $|\mathcal{A}| = A$.
- $\mathbb{P}(s' \mid s, a)$: probability of transitioning to state s' from state s when the chosen action is a.



Markov Decision Process (MDP)

Setting

- $\mathcal{S}$: finite state space with $|\mathcal{S}| = S$.
- $\mathcal{A}$: finite action space with $|\mathcal{A}| = A$.
- $\mathbb{P}(s' \mid s, a)$: probability of transitioning to state s' from state s when the chosen action is a.

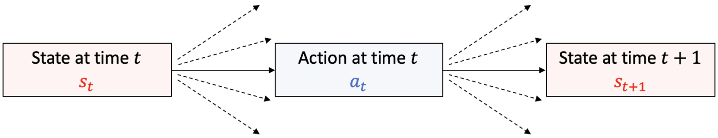


- $r(s, a)$: reward from choosing action a at state s

Markov Decision Process (MDP)

Setting

- $\mathcal{S}$: finite state space with $|\mathcal{S}| = S$.
- $\mathcal{A}$: finite action space with $|\mathcal{A}| = A$.
- $\mathbb{P}(s' \mid s, a)$: probability of transitioning to state s' from state s when the chosen action is a.



- $r(s, a)$: reward from choosing action a at state s
- $\pi(a \mid s)$: policy, given by the probability of taking action a at state s

Markov Decision Process (MDP)

Finite-Horizon MDP

Markov Decision Process (MDP)

Finite-Horizon MDP

- Fixed initial state (or a fixed distribution of the initial state).

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.

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.
- Starting from the initial state s_1, given state s_h in step h, take action a_h and observe the next state s_{h+1}.

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.
- Starting from the initial state s_1, given state s_h in step h, take action a_h and observe the next state s_{h+1}.
- **Cumulative reward**:

$$\sum_{h=1}^{H} r(s_h, a_h).$$

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.
- Starting from the initial state s_1, given state s_h in step h, take action a_h and observe the next state s_{h+1}.
- **Cumulative reward**:

$$\sum_{h=1}^{H} r(s_h, a_h).$$

- **Optimal policy**:

$$\pi^* \in \mathrm{argmax}_{\pi} \left\{ \mathbb{E}\left[\sum_{h=1}^{H} r\left(s_h^{\pi}, a_h^{\pi}\right) \right] \right\}.$$

Markov Decision Process (MDP)

Infinite-Horizon Average-Reward MDP

Markov Decision Process (MDP)

Infinite-Horizon Average-Reward MDP

- Starting from the initial state s_1, given state s_t in time t, take action a_t and observe the next state s_{t+1}.

Markov Decision Process (MDP)

Infinite-Horizon Average-Reward MDP

- Starting from the initial state s_1, given state s_t in time t, take action a_t and observe the next state s_{t+1}.
- **Average reward**:

$$\lim_{T\to\infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right].$$

Markov Decision Process (MDP)

Infinite-Horizon Average-Reward MDP

- Starting from the initial state s_1, given state s_t in time t, take action a_t and observe the next state s_{t+1}.
- **Average reward**:

$$\lim_{T\to\infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right].$$

- **Optimal policy**:

$$\pi^* \in \operatorname{argmax}_{\pi} \left\{ \lim_{T\to\infty} \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} r\left(s_t^{\pi}, a_t^{\pi}\right)\right]\right\}.$$

Reinforcement Learning for MDP

- We assume that the reward function $r(s, a)$ is known.

Reinforcement Learning for MDP

- We assume that the reward function $r(s, a)$ is known.
- If the transition probability $\mathbb{P}(s' \mid s, a)$ is known, we can efficiently compute an optimal policy for both finite- and infinite- horizon MDPs.

Reinforcement Learning for MDP

- We assume that the reward function $r(s, a)$ is known.
- If the transition probability $\mathbb{P}(s' \mid s, a)$ is known, we can efficiently compute an optimal policy for both finite- and infinite- horizon MDPs.
- If not, we apply reinforcement learning to learn near-optimal policies.

Reinforcement Learning for MDP

- We assume that the reward function $r(s, a)$ is known.
- If the transition probability $\mathbb{P}(s' \mid s, a)$ is known, we can efficiently compute an optimal policy for both finite- and infinite- horizon MDPs.
- If not, we apply reinforcement learning to learn near-optimal policies.
- **Basic idea**:

$$\text{trajectory } \{s_1, a_1, \ldots, s_t, a_t\} \text{ up to step } t$$
$$\rightarrow \text{policy } \pi^{t+1} \text{ for step } t+1$$

Reinforcement Learning for MDP

Finite-Horizon Episodic Reinforcement Learning

Reinforcement Learning for MDP

Finite-Horizon Episodic Reinforcement Learning

- Run the finite-horizon MDP multiple times: episodes.

Reinforcement Learning for MDP

Finite-Horizon Episodic Reinforcement Learning

- Run the finite-horizon MDP multiple times: episodes.
- For episode k, we prepare $\pi^k = \{\pi_h^k\}_{h=1}^H$, a collection of policies over the H-horizon.

Reinforcement Learning for MDP

Finite-Horizon Episodic Reinforcement Learning

- Run the finite-horizon MDP multiple times: episodes.
- For episode k, we prepare $\pi^k = \{\pi_h^k\}_{h=1}^H$, a collection of policies over the H-horizon.
- After episode k, we observe trajectory $\left\{s_1, a_1^{\pi_1^k}, s_2^{\pi_1^k}, a_2^{\pi_2^k}, \ldots, s_H^{\pi_{H-1}^k}, a_H^{\pi_H^k}\right\}$.

Reinforcement Learning for MDP

Finite-Horizon Episodic Reinforcement Learning

- Run the finite-horizon MDP multiple times: episodes.
- For episode k, we prepare $\pi^k = \{\pi_h^k\}_{h=1}^H$, a collection of policies over the H-horizon.
- After episode k, we observe trajectory $\left\{s_1, a_1^{\pi_1^k}, s_2^{\pi_1^k}, a_2^{\pi_2^k}, \ldots, s_H^{\pi_{H-1}^k}, a_H^{\pi_H^k}\right\}$.
- Based on the trajectories over the first k episodes, we construct π^{k+1}.

Reinforcement Learning for MDP

Finite-Horizon Episodic Reinforcement Learning

- Run the finite-horizon MDP multiple times: episodes.
- For episode k, we prepare $\pi^k = \{\pi_h^k\}_{h=1}^H$, a collection of policies over the H-horizon.
- After episode k, we observe trajectory $\left\{s_1, a_1^{\pi_1^k}, s_2^{\pi_1^k}, a_2^{\pi_2^k}, \ldots, s_H^{\pi_{H-1}^k}, a_H^{\pi_H^k}\right\}$.
- Based on the trajectories over the first k episodes, we construct π^{k+1}.
- **Total cumulative reward**:

$$\sum_{k=1}^{K} \sum_{h=1}^{H} r\left(s_h^{\pi^k}, a_h^{\pi^k}\right).$$

Reinforcement Learning for MDP

Finite-Horizon Episodic Reinforcement Learning

- Run the finite-horizon MDP multiple times: episodes.
- For episode k, we prepare $\pi^k = \{\pi_h^k\}_{h=1}^H$, a collection of policies over the H-horizon.
- After episode k, we observe trajectory $\left\{ s_1, a_1^{\pi_1^k}, s_2^{\pi_1^k}, a_2^{\pi_2^k}, \ldots, s_H^{\pi_{H-1}^k}, a_H^{\pi_H^k} \right\}$.
- Based on the trajectories over the first k episodes, we construct π^{k+1}.
- **Total cumulative reward**:

$$\sum_{k=1}^K \sum_{h=1}^H r\left(s_h^{\pi^k}, a_h^{\pi^k}\right).$$

- **Regret**:

$$\underbrace{K \sum_{h=1}^H r\left(s_h^*, a_h^*\right)}_{\text{total cumulative reward under an optimal policy}} - \sum_{k=1}^K \sum_{h=1}^H r\left(s_h^{\pi^k}, a_h^{\pi^k}\right).$$

Reinforcement Learning for Infinite-Horizon Average-Reward MDP

Reinforcement Learning for Infinite-Horizon Average-Reward MDP

- At state s_t for time step t, prepare a policy π^t.

Reinforcement Learning for MDP

Reinforcement Learning for Infinite-Horizon Average-Reward MDP

- At state s_t for time step t, prepare a policy π^t.
- Take action a_t from policy π^t & Observe the next state s_{t+1}.

Reinforcement Learning for MDP

Reinforcement Learning for Infinite-Horizon Average-Reward MDP

- At state s_t for time step t, prepare a policy π^t.
- Take action a_t from policy π^t & Observe the next state s_{t+1}.
- **Total cumulative reward over T steps**:

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

Reinforcement Learning for MDP

Reinforcement Learning for Infinite-Horizon Average-Reward MDP

- At state s_t for time step t, prepare a policy π^t.
- Take action a_t from policy π^t & Observe the next state s_{t+1}.
- **Total cumulative reward over T steps**:

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

- **Regret**:

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

Model-Based RL

- We may compute a policy based on an estimation of $\mathbb{P}(s' \mid s, a)$.

Model-Based RL

- We may compute a policy based on an estimation of $\mathbb{P}(s' \mid s, a)$.
- Tabular RL: Learn the probability $\mathbb{P}(s' \mid s, a)$ for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

Model-Based RL

- We may compute a policy based on an estimation of $\mathbb{P}(s' \mid s, a)$.
- Tabular RL: Learn the probability $\mathbb{P}(s' \mid s, a)$ for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.
- Finite-horizon episodic case:

Model-Based RL

- We may compute a policy based on an estimation of $\mathbb{P}(s' \mid s, a)$.
- Tabular RL: Learn the probability $\mathbb{P}(s' \mid s, a)$ for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.
- Finite-horizon episodic case: assuming non-stationary transitions, i.e., $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_{H-1}$,

Model-Based RL

- We may compute a policy based on an estimation of $\mathbb{P}(s' \mid s, a)$.
- Tabular RL: Learn the probability $\mathbb{P}(s' \mid s, a)$ for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.
- Finite-horizon episodic case: assuming non-stationary transitions, i.e., $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_{H-1}$,

UCRL [Jaksch et al., 2010]	$\tilde{O}(H^{3/2} S \sqrt{AK})$
UCBVI [Azar et al., 2017]	$\tilde{O}(H^{3/2} \sqrt{SAK})$
Regret Lower Bound [Jin et al., 2018]	$\Omega(H^{3/2} \sqrt{SAK})$

Infinite-Horizon Tabular MDP

Infinite-Horizon Tabular MDP

- Not all MDPs are learnable!

Model-Based RL

Infinite-Horizon Tabular MDP

- Not all MDPs are learnable!
- Communicating MDPs: MDPs with bounded diameter

Infinite-Horizon Tabular MDP

- Not all MDPs are learnable!
- Communicating MDPs: MDPs with bounded diameter

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

Model-Based RL

Infinite-Horizon Tabular MDP

- Not all MDPs are learnable!
- Communicating MDPs: MDPs with bounded diameter

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

- Weakly communicating MDPs: state space $\mathcal{S}$ has a set of communicating states, and the others are transient states.

Model-Based RL

Infinite-Horizon Tabular MDP

- Not all MDPs are learnable!
- Communicating MDPs: MDPs with bounded diameter

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

- Weakly communicating MDPs: state space $\mathcal{S}$ has a set of communicating states, and the others are transient states.
- A weakly communicating MDP satisfies that $\mathrm{sp}(v^*)$ is bounded where $\mathrm{sp}(v^*)$ is the span of the optimal associated bias function.

Model-Based RL

Infinite-Horizon Tabular MDP

- Not all MDPs are learnable!
- Communicating MDPs: MDPs with bounded diameter

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

- Weakly communicating MDPs: state space $\mathcal{S}$ has a set of communicating states, and the others are transient states.
- A weakly communicating MDP satisfies that $\mathrm{sp}(v^*)$ is bounded where $\mathrm{sp}(v^*)$ is the span of the optimal associated bias function.
- For communicating MDPs, $\mathrm{sp}(v^*) \leq D$.

Model-Based RL

Infinite-Horizon Tabular MDP

- **Regret**:

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]	$\tilde{O}(\sqrt{\mathrm{sp}(v^*)SAT})$
Regret Lower Bound [Jaksch et al., 2010]	$\Omega(\sqrt{DSAT})$

RL with Function Approximation

- For tabular MDPs, regret lower bounds are

Finite-horizon [Jin et al., 2018]	$\Omega(H^{3/2}\sqrt{SAK})$
Infinite-horizon [Jaksch et al., 2010]	$\Omega(\sqrt{DSAT})$

RL with Function Approximation

- For tabular MDPs, regret lower bounds are

Finite-horizon [Jin et al., 2018]	$\Omega(H^{3/2}\sqrt{SAK})$
Infinite-horizon [Jaksch et al., 2010]	$\Omega(\sqrt{DSAT})$

- When the state space $\mathcal{S}$ or the action space $\mathcal{A}$ is large, the regret is large.

RL with Function Approximation

- For tabular MDPs, regret lower bounds are

Finite-horizon [Jin et al., 2018]	$\Omega(H^{3/2}\sqrt{SAK})$
Infinite-horizon [Jaksch et al., 2010]	$\Omega(\sqrt{DSAT})$

- When the state space $\mathcal{S}$ or the action space $\mathcal{A}$ is large, the regret is large.
- A resolution is to approximate the transition model $\mathbb{P}(s' \mid s, a)$ by a function class, e.g., neural networks.

RL with Function Approximation

- For tabular MDPs, regret lower bounds are

Finite-horizon [Jin et al., 2018]	$\Omega(H^{3/2}\sqrt{SAK})$
Infinite-horizon [Jaksch et al., 2010]	$\Omega(\sqrt{DSAT})$

- When the state space $\mathcal{S}$ or the action space $\mathcal{A}$ is large, the regret is large.
- A resolution is to approximate the transition model $\mathbb{P}(s' \mid s, a)$ 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], and autonomous driving [Yurtsever et al., 2020].

RL with Function Approximation

- For tabular MDPs, regret lower bounds are

Finite-horizon [Jin et al., 2018]	$\Omega(H^{3/2}\sqrt{SAK})$
Infinite-horizon [Jaksch et al., 2010]	$\Omega(\sqrt{DSAT})$

- When the state space $\mathcal{S}$ or the action space $\mathcal{A}$ is large, the regret is large.
- A resolution is to approximate the transition model $\mathbb{P}(s' \mid s, a)$ 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], and autonomous driving [Yurtsever et al., 2020].
- Question: does some function structure lead to a smaller regret bound?

Linear MDP

RL with Linear Function Approximation

Linear MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$.

RL with Linear Function Approximation

Linear MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$.
- There is an **unknown** parameter function $\theta : \mathcal{S} \to \mathbb{R}^d$.

RL with Linear Function Approximation

Linear MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$.
- There is an **unknown** parameter function $\theta : \mathcal{S} \to \mathbb{R}^d$.
- Assume that the transition probability is given by

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

RL with Linear Function Approximation

Linear MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$.
- There is an **unknown** parameter function $\theta : \mathcal{S} \to \mathbb{R}^d$.
- Assume that the transition probability is given by

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

- We are interested in the regime where the dimension d is small.

RL with Linear Function Approximation

Linear MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$.
- There is an **unknown** parameter function $\theta : \mathcal{S} \to \mathbb{R}^d$.
- Assume that the transition probability is given by

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

- We are interested in the regime where the dimension d is small.
- Model-based RL boils down to learning the unknown parameter function $\theta(s')$.

Linear Mixture MDP

RL with Linear Function Approximation

Linear Mixture MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.
- There is an **unknown** parameter $\theta \in \mathbb{R}^d$.

RL with Linear Function Approximation

Linear Mixture MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.
- There is an **unknown** parameter $\theta \in \mathbb{R}^d$.
- Assume that the transition probability is given by

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

Linear Mixture MDP

- There is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.
- There is an **unknown** parameter $\theta \in \mathbb{R}^d$.
- Assume that the transition probability is given by

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

- We are interested in the regime where the dimension d is small.
- Model-based RL boils down to learning the unknown parameter θ.

RL with Linear Function Approximation

Regret for Linear MDP

Finite-Horizon Upper Bound [Agarwal et al., 2023, He et al., 2023, Hu et al., 2022]	$\tilde{O}(dH^{3/2}\sqrt{K})$
Finite-Horizon Lower Bound [Zhou et al., 2021]	$\Omega(dH^{3/2}\sqrt{K})$
Infinite-Horizon Upper Bound [Hong et al., 2024]	$\tilde{O}(d^{3/2}\mathrm{sp}(v^*)\sqrt{T})$
Infinite-Horizon Lower Bound [Wu et al., 2022]	$\Omega(d\sqrt{DT})$

RL with Linear Function Approximation

Regret for Linear Mixture MDP

Finite-Horizon Upper Bound [Zhou et al., 2021]	$\tilde{O}(dH^{3/2}\sqrt{K})$
Finite-Horizon Lower Bound [Zhou et al., 2021]	$\Omega(dH^{3/2}\sqrt{K})$
Infinite-Horizon Upper Bound [Wu et al., 2022]	$\tilde{O}(d\sqrt{DT})$
Infinite-Horizon Lower Bound [Wu et al., 2022]	$\Omega(d\sqrt{DT})$

RL with Non-Linear Function Approximation

- Perhaps, the linearity assumption is too restrictive.

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.

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 Non-Linear Function Approximation

General Function Approximation

RL with Non-Linear Function Approximation

General Function Approximation

- One way to consider non-linear functions that are still not too complex is to define a **structural complexity measure**.

General Function Approximation

- One way to consider non-linear functions that are still not too complex is to define a **structural complexity measure**.
- Then we may focus on functions that have **a small value** with respect to a given measure.

RL with Non-Linear Function Approximation

General Function Approximation

- One way to consider non-linear functions that are still not too complex is to define a **structural complexity measure**.
- Then we may focus on functions that have **a small value** with respect to a given measure.
- Eluder dimension [Wang et al., 2020].
- Bellman eluder dimension [Jin et al., 2021].
- Bilinear class [Du et al., 2021].
- Decision-estimation coefficient [Foster et al., 2023].
- Generalized eluder coefficient [Zhong et al., 2023].

RL with Non-Linear Function Approximation

General Function Approximation

- One way to consider non-linear functions that are still not too complex is to define a **structural complexity measure**.
- Then we may focus on functions that have **a small value** with respect to a given measure.
- Eluder dimension [Wang et al., 2020].
- Bellman eluder dimension [Jin et al., 2021].
- Bilinear class [Du et al., 2021].
- Decision-estimation coefficient [Foster et al., 2023].
- Generalized eluder coefficient [Zhong et al., 2023].
- Issue 1: requires solving an abstract optimization / regression problem.

RL with Non-Linear Function Approximation

General Function Approximation

- One way to consider non-linear functions that are still not too complex is to define a **structural complexity measure**.
- Then we may focus on functions that have **a small value** with respect to a given measure.
- Eluder dimension [Wang et al., 2020].
- Bellman eluder dimension [Jin et al., 2021].
- Bilinear class [Du et al., 2021].
- Decision-estimation coefficient [Foster et al., 2023].
- Generalized eluder coefficient [Zhong et al., 2023].
- Issue 1: requires solving an abstract optimization / regression problem.
- Issue 2: no lower bound.

RL with Multinomial Logistic Function Approximation

- [Hwang and Oh, 2023] proposed the multinomial logistic (MNL) function approximation framework.

RL with Multinomial Logistic Function Approximation

- [Hwang and Oh, 2023] proposed the multinomial logistic (MNL) function approximation framework.
- As the linear mixture MDP, there is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.

RL with Multinomial Logistic Function Approximation

- [Hwang and Oh, 2023] proposed the multinomial logistic (MNL) function approximation framework.
- As the linear mixture MDP, there is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.
- Moreover, there is an **unknown** parameter $\theta^* \in \mathbb{R}^d$.

RL with Multinomial Logistic Function Approximation

- [Hwang and Oh, 2023] proposed the multinomial logistic (MNL) function approximation framework.
- As the linear mixture MDP, there is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.
- Moreover, there is an **unknown** parameter $\theta^* \in \mathbb{R}^d$.
- Assume that the transition probability is given by

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

RL with Multinomial Logistic Function Approximation

- [Hwang and Oh, 2023] proposed the multinomial logistic (MNL) function approximation framework.
- As the linear mixture MDP, there is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.
- Moreover, there is an **unknown** parameter $\theta^* \in \mathbb{R}^d$.
- Assume that the transition probability is given by

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

- Again, we are interested in the regime where the dimension d is small.

RL with Multinomial Logistic Function Approximation

- [Hwang and Oh, 2023] proposed the multinomial logistic (MNL) function approximation framework.
- As the linear mixture MDP, there is a **known** feature mapping $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$.
- Moreover, there is an **unknown** parameter $\theta^* \in \mathbb{R}^d$.
- Assume that the transition probability is given by

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

- Again, we are interested in the regime where the dimension d is small.
- Advantage: the MNL framework is natural for modeling transition probabilities.

RL with Multinomial Logistic Function Approximation

Regret Bounds for RL with MNL transitions

RL with Multinomial Logistic Function Approximation

Regret Bounds for RL with MNL transitions

- **Assumption**: there exists $0 < \kappa < 1$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s', s'' \in \mathcal{S}$, we have

$$\inf_{\theta \in \mathbb{R}^d} \mathbb{P}(s' \mid s, a, \theta)\mathbb{P}(s'' \mid s, a, \theta) \geq \kappa.$$

Regret Bounds for RL with MNL transitions

- **Assumption**: there exists $0 < \kappa < 1$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s', s'' \in \mathcal{S}$, we have

$$\inf_{\theta \in \mathbb{R}^d} \mathbb{P}(s' \mid s, a, \theta)\mathbb{P}(s'' \mid s, a, \theta) \geq \kappa.$$

- Let $0 < \kappa^* < 1$ satisfy that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s', s'' \in \mathcal{S}$, we have

$$\mathbb{P}(s' \mid s, a, \theta^*)\mathbb{P}(s'' \mid s, a, \theta^*) \geq \kappa^*.$$

RL with Multinomial Logistic Function Approximation

Regret Bounds for RL with MNL transitions

- **Assumption**: there exists $0 < \kappa < 1$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s', s'' \in \mathcal{S}$, we have

$$\inf_{\theta \in \mathbb{R}^d} \mathbb{P}(s' \mid s, a, \theta)\mathbb{P}(s'' \mid s, a, \theta) \geq \kappa.$$

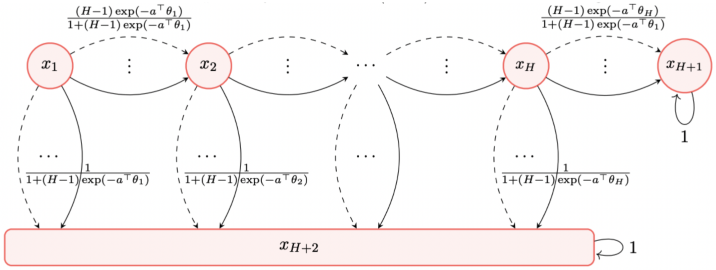
- Let $0 < \kappa^* < 1$ satisfy that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s', s'' \in \mathcal{S}$, we have

$$\mathbb{P}(s' \mid s, a, \theta^*)\mathbb{P}(s'' \mid s, a, \theta^*) \geq \kappa^*.$$

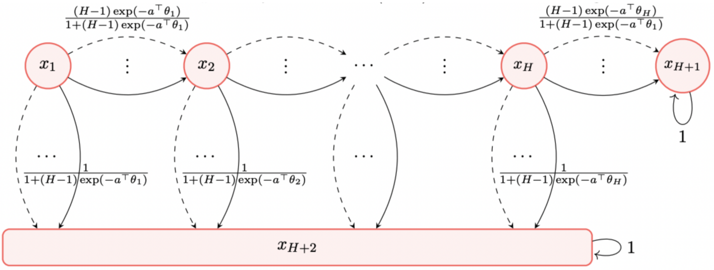
- **Regret**:

UCRL-MNL [Hwang and Oh, 2023]	$\tilde{O}(\kappa^{-1} dH^2\sqrt{K})$
UCRL-MNL-LL+ [Li et al., 2024]	$\tilde{O}(dH^2\sqrt{K} + \kappa^{-1} d^2 H^2)$
UCRL-MNL+ [Cho et al., 2024]	$\tilde{O}(dH^2\sqrt{K} + \kappa^{-1} d^2 H^2)$
Regret Lower Bound [Li et al., 2024]	$\Omega(dH\sqrt{K\kappa^*})$

Our Result 1: Tighter Lower Bound



Our Result 1: Tighter Lower Bound



Theorem

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 a regret at least

$$\mathbb{E}\left[\mathrm{regret}(M, \mathfrak{A}, K)\right] \geq \frac{(d-1)H^{3/2}\sqrt{K}}{480\sqrt{2}}$$

where the expectation is taken over the randomness generated by M and $\mathfrak{A}$.

Our Result 2: Algorithms for Infinite-Horizon Average-Reward Setting

Theorem

Let M be a communicating MDP governed by the MNL transition model, and let D denote the diameter of M. There is an algorithm, called `UCRL2-MNL`, *that guarantees that for any initial state s_1,*

$$\mathrm{Regret}(M, \texttt{UCRL2-MNL}, s_1, T) = \tilde{O}\left(\kappa^{-1} D d \sqrt{T}\right)$$

with probability at least $1 - 2\delta$.

Our Result 2: Algorithms for Infinite-Horizon Average-Reward Setting

Theorem

Let M be a communicating MDP governed by the MNL transition model, and let D denote the diameter of M. There is an algorithm, called `UCRL2-MNL`*, that guarantees that for any initial state s_1,*

$$\mathrm{Regret}(M, \texttt{UCRL2-MNL}, s_1, T) = \tilde{O}\left(\kappa^{-1} D d \sqrt{T}\right)$$

with probability at least $1 - 2\delta$.

- `UCRL2-MNL` is an adaptation of `UCRL2` due to [Jaksch et al., 2010].

Our Result 2: Algorithms for Infinite-Horizon Average-Reward Setting

Theorem

Let M be a communicating MDP governed by the MNL transition model, and let D denote the diameter of M. There is an algorithm, called `UCRL2-MNL`, that guarantees that for any initial state s_1,

$$\mathrm{Regret}(M, \texttt{UCRL2-MNL}, s_1, T) = \tilde{O}\left(\kappa^{-1} D d \sqrt{T}\right)$$

with probability at least $1 - 2\delta$.

- `UCRL2-MNL` is an adaptation of `UCRL2` due to [Jaksch et al., 2010].
- The main component is running extended value iteration.

Our Result 2: Algorithms for Infinite-Horizon Average-Reward Setting

Theorem

Let M be a weakly communicating MDP governed by the MNL transition model, and let $\mathrm{sp}(v^*)$ denote the span of the associated optimal bias function. There is an algorithm, called `OVIFH-MNL`, that guarantees that for any initial state s_1,

$$\mathrm{Regret}(M, \texttt{OVIFH-MNL}, s_1, T) = \tilde{O}\left(\kappa^{-2/5}\mathrm{sp}(v^*)d^{2/5}T^{4/5}\right)$$

with probability at least $1 - 2\delta$.

Our Result 2: Algorithms for Infinite-Horizon Average-Reward Setting

Theorem

Let M be a weakly communicating MDP governed by the MNL transition model, and let $\mathrm{sp}(v^*)$ denote the span of the associated optimal bias function. There is an algorithm, called `OVIFH-MNL`, that guarantees that for any initial state s_1,

$$\mathrm{Regret}(M, \mathtt{OVIFH\text{-}MNL}, s_1, T) = \tilde{O}\left(\kappa^{-2/5}\mathrm{sp}(v^*)d^{2/5}T^{4/5}\right)$$

with probability at least $1 - 2\delta$.

- `OVIFH-MNL` decomposes the T-horizon to T/H episodes of length H.

Our Result 2: Algorithms for Infinite-Horizon Average-Reward Setting

Theorem

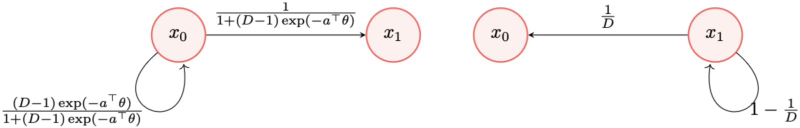
Let M be a weakly communicating MDP governed by the MNL transition model, and let $\mathrm{sp}(v^*)$ denote the span of the associated optimal bias function. There is an algorithm, called `OVIFH-MNL`, that guarantees that for any initial state s_1,

$$\mathrm{Regret}(M, \texttt{OVIFH-MNL}, s_1, T) = \tilde{O}\left(\kappa^{-2/5}\mathrm{sp}(v^*)d^{2/5}T^{4/5}\right)$$

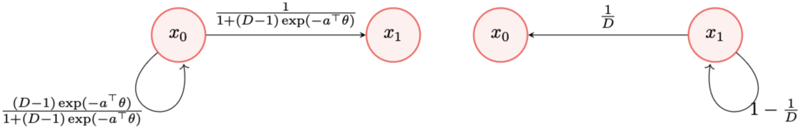
with probability at least $1 - 2\delta$.

- `OVIFH-MNL` decomposes the T-horizon to T/H episodes of length H.
- For each episode, we apply `UCRL-MNL` [Hwang and Oh, 2023].

Our Result 3: Strong Lower Bound for Learning Communicating MDPs



Our Result 3: Strong Lower Bound for Learning Communicating MDPs



Theorem

There is an MDP instance M with $d \geq 2$, $D \geq 101$, and $T \geq 45(d-1)^2 D$ for which any algorithm $\mathfrak{A}$ incurs a regret at least

$$\mathbb{E}\left[\mathrm{regret}(M, \mathfrak{A}, x_0, T)\right] \geq \frac{1}{4050} d\sqrt{DT}$$

where the expectation is taken over the randomness generated by M and $\mathfrak{A}$.

Summary

- Tighter lower bound for the finite-horizon setting:

$$\Omega(dH^{3/2}\sqrt{K})$$

improves upon the lower bound $\Omega(dH\sqrt{K\kappa^*})$ due to [Li et al., 2024].

Summary

- Tighter lower bound for the finite-horizon setting:

$$\Omega(dH^{3/2}\sqrt{K})$$

  improves upon the lower bound $\Omega(dH\sqrt{K\kappa^*})$ due to [Li et al., 2024].
- UCRL2-MNL for the infinite-horizon average-reward setting with regret

$$\tilde{O}(dD\sqrt{T})$$

  for communicating MDPs with diameter at most D.

Summary

- Tighter lower bound for the finite-horizon setting:

$$\Omega(dH^{3/2}\sqrt{K})$$

  improves upon the lower bound $\Omega(dH\sqrt{K\kappa^*})$ due to [Li et al., 2024].
- UCRL2-MNL for the infinite-horizon average-reward setting with regret

$$\tilde{O}(dD\sqrt{T})$$

  for communicating MDPs with diameter at most D.
- OVIFH-MNL for the infinite-horizon average-reward setting with regret

$$\tilde{O}\left(\kappa^{-2/5}\mathrm{sp}(v^*)d^{2/5}T^{4/5}\right)$$

  for weakly communicating MDPs.

Summary

- Tighter lower bound for the finite-horizon setting:

$$\Omega(dH^{3/2}\sqrt{K})$$

  improves upon the lower bound $\Omega(dH\sqrt{K\kappa^*})$ due to [Li et al., 2024].

- UCRL2-MNL for the infinite-horizon average-reward setting with regret

$$\tilde{O}(dD\sqrt{T})$$

  for communicating MDPs with diameter at most D.

- OVIFH-MNL for the infinite-horizon average-reward setting with regret

$$\tilde{O}\left(\kappa^{-2/5}\mathrm{sp}(v^*)d^{2/5}T^{4/5}\right)$$

  for weakly communicating MDPs.

- Lower bound for the finite-horizon setting:

$$\Omega(d\sqrt{DT}).$$

Confidence Sets for the Transition Parameter

- **Log-likelihood function**:

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

Confidence Sets for the Transition Parameter

- **Log-likelihood function**:

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

- **Ridge-penalized MLE**:

$$\widehat{\theta}_t = \mathrm{argmax}_\theta \left\{ \ell_t(\theta) - \frac{\lambda}{2} \|\theta\|_2^2 \right\}.$$

Confidence Sets for the Transition Parameter

- **Log-likelihood function**:

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

- **Ridge-penalized MLE**:

$$\widehat{\theta}_t = \mathrm{argmax}_\theta \left\{ \ell_t(\theta) - \frac{\lambda}{2} \|\theta\|_2^2 \right\}.$$

- **Gram matrix**:

$$A_{t+1} := \lambda I_d + \sum_{i=1}^{t} \sum_{s' \in \mathcal{S}_{s_i, a_i}} \varphi_{i,s'} \varphi_{i,s'}^\top = A_t + \sum_{s' \in \mathcal{S}_{s_t, a_t}} \varphi_{t,s'} \varphi_{t,s'}^\top$$

Confidence Sets for the Transition Parameter

- **Confidence sets**:

$$\mathcal{C}_t := \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \widehat{\theta}_t \right\|_{A_t} \leq \beta_t \right\}$$

where

$$\beta_t = \frac{1}{\kappa} \sqrt{d \log \left(1 + \frac{t \mathcal{U} L_\varphi^2}{d \lambda} \right) + 2 \log \frac{1}{\delta}} + \frac{\sqrt{\lambda}}{\kappa} L_\theta$$

and $\mathcal{U} = \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\mathcal{S}_{s,a}|$.

Confidence Sets for the Transition Parameter

- **Confidence sets**:

$$\mathcal{C}_t := \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \widehat{\theta}_t \right\|_{A_t} \leq \beta_t \right\}$$

where

$$\beta_t = \frac{1}{\kappa} \sqrt{d \log \left(1 + \frac{t \mathcal{U} L_\varphi^2}{d \lambda} \right) + 2 \log \frac{1}{\delta}} + \frac{\sqrt{\lambda}}{\kappa} L_\theta$$

and $\mathcal{U} = \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\mathcal{S}_{s,a}|$.

Lemma

With probability at least $1 - \delta$, it holds that $\theta^* \in \mathcal{C}_t$ for all $t \in [T]$.

UCRL2-MNL: Extended Value Iteration

Algorithm 0 Extended Value Iteration ($\texttt{EVI}(\mathcal{C},\epsilon)$)

Inputs: confidence set $\mathcal{C}$, a desired accuracy level ϵ
Initialize: $u^{(0)}(s) = 0$ for every $s \in \mathcal{S}$ and $i = 0$.
while $\max_{s\in\mathcal{S}} \left\{u^{(i+1)}(s) - u^{(i)}(s)\right\} - \min_{s\in\mathcal{S}} \left\{u^{(i+1)}(s) - u^{(i)}(s)\right\} > \epsilon$ **do**

Set

$$u^{(i+1)}(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \max_{\theta\in\mathcal{C}} \left\{ \sum_{s'\in\mathcal{S}_{s,a}} p(s' \mid s,a,\theta) u^{(i)}(s') \right\} \right\}$$

Set $i = i + 1$
end while
Return $u^{(i)}(s)$ for $s \in \mathcal{S}$

UCRL2-MNL: Greedy Policy

- For $s \in \mathcal{S}$,

$$\pi(s) = \text{argmax}_{a \in \mathcal{A}} \left\{ r(s,a) + \max_{\theta \in \mathcal{C}} \left\{ \sum_{s' \in \mathcal{S}_{s,a}} p(s' \mid s, a, \theta) u(s') \right\} \right\}.$$

UCRL2-MNL

Algorithm 1 `UCRL2-MNL`

Input: feature map $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, confidence level $\delta \in (0, 1)$, and parameters $\lambda, L_\varphi, L_\theta, \kappa, \mathcal{U}$
Initialize: $t = 1$, $\widehat{\theta}_1 = 0$, $A_1 = \lambda I_d$, and observe the initial state $s_1 \in \mathcal{S}$
for episodes $k = 1, 2, \ldots,$ **do**
  Set $t_k = t$
  Set $u_k(s)$ as the output of $\texttt{EVI}(\mathcal{C}_{t_k}, \epsilon)$ for $s \in \mathcal{S}$ where $\mathcal{C}_{t_k}$
  Set $w_k(s) = u_k(s) - (\max_{s \in \mathcal{S}} u_k(s) + \min_{s \in \mathcal{S}} u_k(s))/2$ for $s \in \mathcal{S}$
  Take policy π_k by setting $\pi_k(s)$ with $u = w_k$ and $\mathcal{C} = \mathcal{C}_{t_k}$ for $s \in \mathcal{S}$
  while $\det(A_t) \leq 2 \det(A_{t_k})$ **do**
    Take action $a_t = \pi_k(s_t)$ and observe s_{t+1} sampled from $p(\cdot \mid s_t, a_t)$
    Set $A_{t+1} = A_t + \sum_{s' \in \mathcal{S}_t} \varphi_{t,s'} \varphi_{t,s'}^\top$
    Update $t = t + 1$
  end while
end for

OVIFH-MNL

- UCRL2-MNL requires solving

$$\max_{\theta \in \mathcal{C}} \left\{ \sum_{s' \in \mathcal{S}_{s,a}} p(s' \mid s, a, \theta) u(s') \right\}$$

which is a non-convex optimization problem.

OVIFH-MNL

- UCRL2-MNL requires solving

$$\max_{\theta \in \mathcal{C}} \left\{ \sum_{s' \in \mathcal{S}_{s,a}} p(s' \mid s, a, \theta) u(s') \right\}$$

  which is a non-convex optimization problem.
- UCRL2-MNL does not apply to general weakly communicating MDPs.

OVIFH-MNL

- UCRL2-MNL requires solving

$$\max_{\theta \in \mathcal{C}} \left\{ \sum_{s' \in \mathcal{S}_{s,a}} p(s' \mid s, a, \theta) u(s') \right\}$$

  which is a non-convex optimization problem.
- UCRL2-MNL does not apply to general weakly communicating MDPs.
- Question: can we find a more computationally efficient algorithm that applies to weakly communicating MDPs?

OVIFH-MNL

- **Idea**: decompose the T-horizon to T/H episodes of fixed length H.

OVIFH-MNL

- **Idea**: decompose the T-horizon to T/H episodes of fixed length H.
- Then apply optimistic value iteration such as UCRL-MNL by [Hwang and Oh, 2023].

OVIFH-MNL

- **Idea**: decompose the T-horizon to T/H episodes of fixed length H.
- Then apply optimistic value iteration such as UCRL-MNL by [Hwang and Oh, 2023].
- **Optimistic value function**:

$$\widehat{Q}_{k,h}(s,a) := r(s,a) + \sum_{s' \in \mathcal{S}_{s,a}} p\left(s' \mid s, a, \widehat{\theta}_{t_k}\right) \widehat{V}_{k,h+1}(s') + 2H\beta_{t_k} \max_{s' \in \mathcal{S}_{s,a}} \|\phi(s,a,s')\|_{A_{t_k}^{-1}}$$

Algorithm 2 `OVIFH-MNL`

Input: feature map $\varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, confidence level $\delta \in (0, 1)$, and parameters $\lambda, L_\varphi, L_\theta, \kappa, \mathcal{U}$

Initialize: $\widehat{\theta}_1 = 0$, $A_1 = \lambda I_d$, and observe the initial state $s_1 \in \mathcal{S}$

for episodes $k = 1, 2, \ldots, T/H$ **do**

  Set $\widehat{Q}_{k,h}(s, a)$ for $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$

  for steps $h = 1, \ldots, H$ **do**

    Set $t = (k - 1)H + h$

    Take action $a_t = \mathrm{argmax}_{a \in \mathcal{A}} \widehat{Q}_{k,h}(s_t, a)$ and observe s_{t+1} sampled from $p(\cdot \mid s_t, a_t)$

  end for

end for

Thank you!

A draft is now available online: `https://dabeenl.github.io`

A. Agarwal, Y. Jin, and T. Zhang. Voql: Towards optimal regret in model-free rl with nonlinear function approximation. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 987–1063. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/agarwal23a.html.

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.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/azar17a.html.

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.

S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in rl. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/du21a.html.

D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin. The statistical complexity of interactive decision making, 2023.

J. He, H. Zhao, D. Zhou, and Q. Gu. Nearly minimax optimal reinforcement learning for linear Markov decision processes. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12790–12822. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/he23d.html.

K. Hong, Y. Zhang, and A. Tewari. Provably efficient reinforcement learning for infinite-horizon average-reward linear mdps, 2024.

P. Hu, Y. Chen, and L. Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th*

International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 8971–9019. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hu22a.html.

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. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=b8Kl8mcK6tb.

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.

L.-F. Li, Y.-J. Zhang, P. Zhao, and Z.-H. Zhou. Provably efficient reinforcement learning with multinomial logit function approximation, 2024.

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.

R. Wang, R. R. Salakhutdinov, and L. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6123–6135. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/440924c5948e05070663f88e69e8242b-Paper.pdf.

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.

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.

H. Zhong, W. Xiong, S. Zheng, L. Wang, Z. Wang, Z. Yang, and T. Zhang. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond, 2023.

D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4532–4576. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/zhou21a.html.