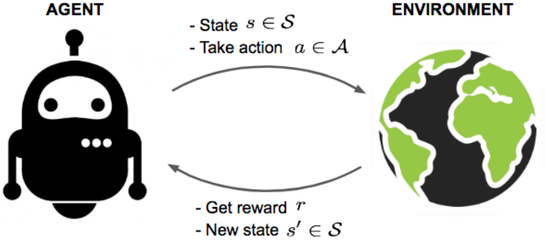
Model-free Reinforcement Learning in Infinite-horizon Average-reward MDPs

Chen-Yu Wei Mehdi Jafarnia-Jahromi Haipeng Luo
Hiteshi Sharma Rahul Jain

University of Southern California

Problem Formulation



Consider a Markov Decision Process (MDP) with

- A finite set of states $\mathcal{S}$
- A finite set of actions $\mathcal{A}$
- known reward function $r(s, a)$
- unknown transition kernel $p(s'|s, a)$

Goal

Maximize the sum of reward $\sum_{t=1}^{T} r(s_t, a_t)$.
(average-reward setting)

Goal

Maximize the sum of reward $\sum_{t=1}^{T} r(s_t, a_t)$.
(average-reward setting)

To **evaluate the performance**, define

$$J^* = \max_{\pi:\mathcal{S}\to\mathcal{A}} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r(s_t, \pi(s_t))\right],$$

and

$$\text{Regret}_T = TJ^* - \sum_{t=1}^{T} r(s_t, a_t).$$

Goal

Maximize the sum of reward $\sum_{t=1}^{T} r(s_t, a_t)$.
(average-reward setting)

To **evaluate the performance**, define

$$J^* = \max_{\pi:S\to\mathcal{A}} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r(s_t, \pi(s_t))\right],$$

and

$$\text{Regret}_T = TJ^* - \sum_{t=1}^{T} r(s_t, a_t).$$

A regret sublinear in T implies that the learner's performance is asymptotically same as the best policy.

Model-based vs. Model-free Methods

- Model-based methods: learns the **underlying rules of the world** and performs planning based on them.
  e.g., modeling the transition probability $p(s'|s, a)$
- Model-free methods: directly learns **how to act**
  e.g., modeling the state-action value $Q^*(s, a)$ or the optimal policy $\pi^*(a|s)$

Model-based vs. Model-free Methods

- Pros and Cons: Empirically, model-based methods are more **sample efficient**; however, model-free methods are more **memory efficient** and **robust against model error**.

Model-based vs. Model-free Methods

- Pros and Cons: Empirically, model-based methods are more **sample efficient**; however, model-free methods are more **memory efficient** and **robust against model error**.

- In many applications, model-free methods achieve state-of-the-art performance (e.g., many Atari games).

Model-based vs. Model-free Methods

- Pros and Cons: Empirically, model-based methods are more **sample efficient**; however, model-free methods are more **memory efficient** and **robust against model error**.
- In many applications, model-free methods achieve state-of-the-art performance (e.g., many Atari games).
- Theoretical analysis on model-free methods is relatively scarce despite its empirical success (there is a resurge since the recent work of [Jin et al.'18]).

Contribution

We provide the state-of-the-art regret bound for the average-reward setting under model-free methods.

Contribution

We provide the state-of-the-art regret bound for the average-reward setting under model-free methods.

Comparing with previous work:

Sub-class of MDP	Ours	Best MF	Best MB
Ergodic	$\boldsymbol{O(\sqrt{T})}$	$O(T^{\frac{3}{4}})$ (Politex)	$O(\sqrt{T})$
Weakly-comm.	$\boldsymbol{O(T^{\frac{2}{3}})}$	No previous bound	$O(\sqrt{T})$

Contribution

We provide the state-of-the-art regret bound for the average-reward setting under model-free methods.

Comparing with previous work:

Sub-class of MDP	Ours	Best MF	Best MB
Ergodic	$\boldsymbol{O(\sqrt{T})}$	$O(T^{\frac{3}{4}})$ (Politex)	$O(\sqrt{T})$
Weakly-comm.	$\boldsymbol{O(T^{\frac{2}{3}})}$	No previous bound	$O(\sqrt{T})$

- Ergodic MDPs $\subset$ weakly communicating MDPs

Contribution

We provide the state-of-the-art regret bound for the average-reward setting under model-free methods.

Comparing with previous work:

Sub-class of MDP	Ours	Best MF	Best MB
Ergodic	$\boldsymbol{O(\sqrt{T})}$	$O(T^{\frac{3}{4}})$ (Politex)	$O(\sqrt{T})$
Weakly-comm.	$\boldsymbol{O(T^{\frac{2}{3}})}$	No previous bound	$O(\sqrt{T})$

- Ergodic MDPs $\subset$ weakly communicating MDPs
- Weakly communicating is the minimal assumption required for sublinear regret to be possible.

Contribution

We provide the state-of-the-art regret bound for the average-reward setting under model-free methods.

Comparing with previous work:

Sub-class of MDP	Ours	Best MF	Best MB
Ergodic	$\boldsymbol{O(\sqrt{T})}$	$O(T^{\frac{3}{4}})$ (Politex)	$O(\sqrt{T})$
Weakly-comm.	$\boldsymbol{O(T^{\frac{2}{3}})}$	No previous bound	$O(\sqrt{T})$

- Ergodic MDPs $\subset$ weakly communicating MDPs
- Weakly communicating is the minimal assumption required for sublinear regret to be possible.
- **open problem:** Can the bound of model-free methods match that of model-based methods?

Two cases that we study

- **Ergodic MDPs**:
  $\Rightarrow$ all policies are explorative in the state space
  $\Rightarrow$ no need to worry about exploring the state space
  $\Rightarrow$ $O(\sqrt{T})$ regret

- **Weakly communicating MDPs**:
  $\Rightarrow$ there exist strategies that explores the state space
  $\Rightarrow$ adding exploration bonus to guide exploration
  $\Rightarrow$ $O(T^{\frac{2}{3}})$ regret

Case 1. Ergodic MDP

MDP-OOMD (Optimistic Online Mirror Descent)

Ergodic MDPs

1. **Uniformly mixing**: Under any policy π, any initial state s_1,

$$\left| \Pr\{s_t = s\} - \underbrace{\mu^\pi(s)}_{\text{stationary distribution}} \right| \le O(e^{-t/t_{\text{mix}}}).$$

2. **Lower bounded stationary probability**: for any policy π,

$$\mu^\pi(s) \ge \epsilon > 0.$$

Ergodic MDPs – MDP-OOMD Algorithm

- Running a multi-armed bandit (MAB) algorithm on each state.
- Feed cumulative reward in the trajectory of length $N \approx t_{\text{mix}}$ to the MAB algorithms.

Ergodic MDPs – MDP-OOMD Algorithm

- Running a multi-armed bandit (MAB) algorithm on each state.
- Feed cumulative reward in the trajectory of length $N \approx t_{\text{mix}}$ to the MAB algorithms.

Parameters: B (epoch length), N (trajectory length)
For $k = 1, 2, \ldots$

Ergodic MDPs – MDP-OOMD Algorithm

- Running a multi-armed bandit (MAB) algorithm on each state.
- Feed cumulative reward in the trajectory of length $N \approx t_{\text{mix}}$ to the MAB algorithms.

Parameters: B (epoch length), N (trajectory length)
For $k = 1, 2, \ldots$

- Execute π_k for B steps and get the sequence

$$\mathcal{T} = (s_1, a_1, s_2, a_2, \ldots, s_B, a_B).$$

Ergodic MDPs – MDP-OOMD Algorithm

- Running a multi-armed bandit (MAB) algorithm on each state.
- Feed cumulative reward in the trajectory of length $N \approx t_{\text{mix}}$ to the MAB algorithms.

Parameters: B (epoch length), N (trajectory length)
For $k = 1, 2, \ldots$

- Execute π_k for B steps and get the sequence
$$\mathcal{T} = (s_1, a_1, s_2, a_2, \ldots, s_B, a_B).$$
- For each state-action pair (s, a),
  - Find several length-N sub-trajectories of $\mathcal{T}$ that starts from (s, a). Let $R(s, a)$ be their reward average.

Ergodic MDPs – MDP-OOMD Algorithm

- Running a multi-armed bandit (MAB) algorithm on each state.
- Feed cumulative reward in the trajectory of length $N \approx t_{\text{mix}}$ to the MAB algorithms.

Parameters: B (epoch length), N (trajectory length)
For $k = 1, 2, \ldots$

- Execute π_k for B steps and get the sequence

$$\mathcal{T} = (s_1, a_1, s_2, a_2, \ldots, s_B, a_B).$$

- For each state-action pair (s, a),
  - Find several length-N sub-trajectories of $\mathcal{T}$ that starts from (s, a). Let $R(s, a)$ be their reward average.
- For each state s,
  - Update the MAB on state s with rewards $R(s, a)$ $\forall a$.

Ergodic MDPs – Regret Bound

The MAB algorithm we use is **Optimistic Online Mirror Descent** with log-barrier as the regularizer.

Ergodic MDPs – Regret Bound

The MAB algorithm we use is **Optimistic Online Mirror Descent** with log-barrier as the regularizer.

MDP-OOMD achieves

$$\mathbb{E}\left[TJ^* - \sum_{t=1}^{T} r(s_t, a_t)\right] \leq O\left(\sqrt{t_{\text{mix}}^3 \rho |\mathcal{A}| T}\right)$$

where

$$\rho = \max_{\pi} \sum_{s} \frac{\mu^{\pi^*}(s)}{\mu^{\pi}(s)} \qquad \text{(distribution mismatch coefficient)}$$

Ergodic MDPs – Regret Bound

The MAB algorithm we use is **Optimistic Online Mirror Descent** with log-barrier as the regularizer.

MDP-OOMD achieves

$$\mathbb{E}\left[TJ^* - \sum_{t=1}^{T} r(s_t, a_t)\right] \leq O\left(\sqrt{t_{\text{mix}}^3 \rho |\mathcal{A}| T}\right)$$

where

$$\rho = \max_{\pi} \sum_{s} \frac{\mu^{\pi^*}(s)}{\mu^{\pi}(s)} \qquad \text{(distribution mismatch coefficient)}$$

Remark. MDP-OOMD is essentially a policy-gradient algorithm with some new **variance reduction** scheme.

Case 2. Weakly Communicating MDP

Discounted optimistic Q-learning

Weakly Communicating MDPs

Bounded bias span: for any pair of states s, s', under the best policy π^*, the advantage of starting from s over starting from s' is bounded.

$$\mathbb{E}\left[\sum_{t=1}^{\tau} r(s_t, \pi^*(s_t)) \;\middle|\; s_1 = s\right] - \mathbb{E}\left[\sum_{t=1}^{\tau} r(s_t, \pi^*(s_t)) \;\middle|\; s_1 = s'\right] \leq D$$

for any τ.

Weakly Communicating MDPs – Optimistic Q-learning

- Discounted Q-learning with adaptive discount factor γ_t
- Exploration bonus b_τ
- Carefully tuned learning rates α_τ

Weakly Communicating MDPs – Optimistic Q-learning

- Discounted Q-learning with adaptive discount factor γ_t
- Exploration bonus b_τ
- Carefully tuned learning rates α_τ

Parameters: C (bonus parameter)
For $t = 1, 2, \ldots$

Weakly Communicating MDPs – Optimistic Q-learning

- Discounted Q-learning with adaptive discount factor γ_t
- Exploration bonus b_τ
- Carefully tuned learning rates α_τ

Parameters: C (bonus parameter)
For $t = 1, 2, \ldots$

- Choose $a_t = \operatorname{argmax}_a Q_t(s_t, a)$

Weakly Communicating MDPs – Optimistic Q-learning

- Discounted Q-learning with adaptive discount factor γ_t
- Exploration bonus b_τ
- Carefully tuned learning rates α_τ

Parameters: C (bonus parameter)
For $t = 1, 2, \dots$

- Choose $a_t = \mathrm{argmax}_a\, Q_t(s_t, a)$
- Update $Q_{t+1}(s_t, a_t) =$

$$Q_t(s_t, a_t) + \underbrace{\alpha_\tau}_{\text{learning rate}} \left(r(s_t, a_t) + \gamma_t V_t(s_{t+1}) - Q_t(s_t, a_t) + \underbrace{b_\tau}_{\text{bonus}} \right)$$

Weakly Communicating MDPs – Optimistic Q-learning

- Discounted Q-learning with adaptive discount factor γ_t
- Exploration bonus b_τ
- Carefully tuned learning rates α_τ

Parameters: C (bonus parameter)

For $t = 1, 2, \ldots$

- Choose $a_t = \text{argmax}_a Q_t(s_t, a)$
- Update $Q_{t+1}(s_t, a_t) =$

$$Q_t(s_t, a_t) + \underbrace{\alpha_\tau}_{\text{learning rate}} \left(r(s_t, a_t) + \gamma_t V_t(s_{t+1}) - Q_t(s_t, a_t) + \underbrace{b_\tau}_{\text{bonus}} \right)$$

where $\gamma_t = 1 - \left(\frac{t}{|\mathcal{S}||\mathcal{A}|}\right)^{-\frac{1}{3}}$, $\tau \triangleq \#\text{visit}(s_t, a_t)$

$$\alpha_\tau = \frac{1}{1+\tau(1-\gamma_t)}, \quad b_\tau = C\sqrt{\frac{1}{\tau(1-\gamma_t)}}$$

Weakly Communicating MDPs – Regret Bound

Optimistic Q-learning achieves

$$\mathbb{E}\left[TJ^* - \sum_{t=1}^{T} r(s_t, a_t)\right] \leq O\left(D\sqrt[3]{SAT^2}\right)$$

where D is the bias span.

Weakly Communicating MDPs – Regret Bound

Optimistic Q-learning achieves

$$\mathbb{E}\left[TJ^* - \sum_{t=1}^{T} r(s_t, a_t)\right] \le O\left(D\sqrt[3]{SAT^2}\right)$$

where D is the bias span.

Technical contribution: how to use a discount algorithm to solve an average-reward problem.

Summary

- We propose two **model-free** online reinforcement learning algorithms for MDPs with finite states and actions in the **average-reward setting**

Summary

- We propose two **model-free** online reinforcement learning algorithms for MDPs with finite states and actions in the **average-reward setting**
- We formalize the regret bounds of our algorithms, which are either new or improve over previous results.

Summary

- We propose two **model-free** online reinforcement learning algorithms for MDPs with finite states and actions in the **average-reward setting**
- We formalize the regret bounds of our algorithms, which are either new or improve over previous results.
- **MDP-OOMD** gets $O(\sqrt{T})$ regret under the ergodic assumption. **Optimistic Q-learning** gets $O(T^{\frac{2}{3}})$ regret under weakly communicating assumption.