Tiered Reinforcement Learning: Pessimistic in the Face of Uncertainty and Constant Regret

Jiawei Huang, Li Zhao, Tao Qin, Wei Chen, Nan Jiang, Tie-Yan Liu

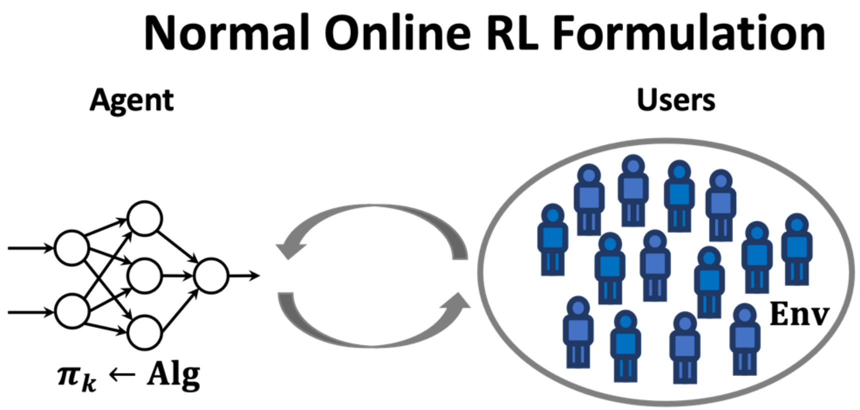
ETH zürich

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Microsoft

Introduction

- RL has been applied in many applications with user interaction:
  - Medical Treatment
  - Recommendation System
  - Other Online Application Services



- The normal learning protocol (Fig. RHS)
  - Repeatedly:
    - **Policy Improvement:** Learn a policy from collected data
    - **Collect New Data from Env:** User comes; generate trajectories during interaction.

Introduction

However, users may have ' **Tiered Structure**':
The users can be divided into 2 (or more) groups by their different preference on exploration risk.

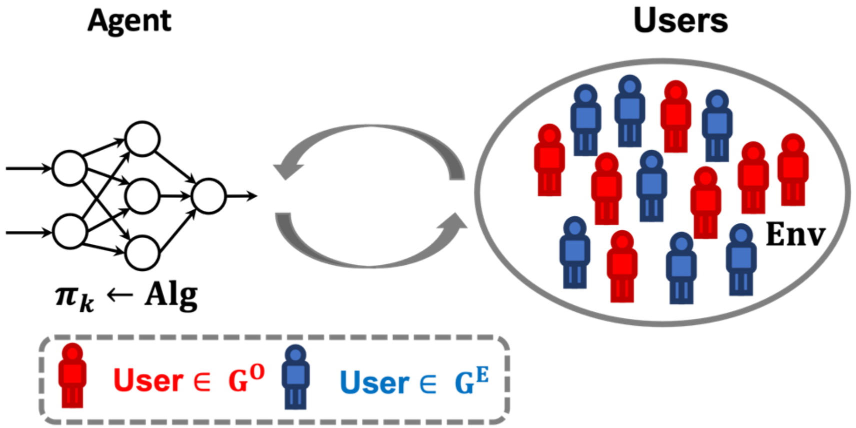
- RL has been applied in many applications with user interaction:
  - Medical Treatment

    Paid Volunteers v.s. Conservative Patients
  - Recommendation System

    Paid Tester v.s. General Customers
  - Other Online Application Services

    Customers Using Free Services v.s. Paid VIP Customers

Normal Online RL Formulation



G^O (short for $\text{Group}^{\text{Online}}$): risk-tolerant user group
G^E (short for $\text{Group}^{\text{Exploit}}$): risk-averse user group

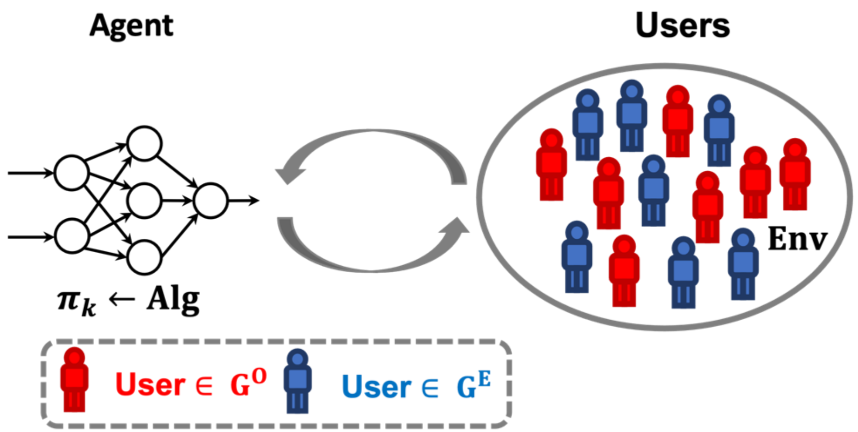
- The normal learning protocol (Fig. RHS)
  - Repeatedly:
    - **Policy Improvement:** Learn a policy from collected data
    - **Collect New Data from Env:** User comes; generate trajectories during interaction.

Introduction

However, users may have ‘ **Tiered Structure**’:
The users can be divided into 2 (or more) groups by their different preference on exploration risk.

- RL has been applied in many applications with user interaction:
  - Medical Treatment
    Paid Volunteers v.s. Conservative Patients
  - Recommendation System
    Paid Tester v.s. General Customers
  - Other Online Application Services
    Customers Using Free Services v.s. Paid VIP Customers

Normal Online RL Formulation



G^O (short for $\text{Group}^{\text{Online}}$): risk-tolerant user group
G^E (short for $\text{Group}^{\text{Exploit}}$): risk-averse user group

- The normal learning protocol (Fig. RHS)
  - Repeatedly:
    - **Policy Improvement:** Learn a policy from collected data
    - **Collect New Data from Env:** User comes; generate trajectories during interaction.
  - Users in different groups will be treated equivalently and suffer similar loss…

Tiered RL Framework

Initialize $D_1 \leftarrow \{\}$.

for $k = 1, 2, ..., K$ **do**

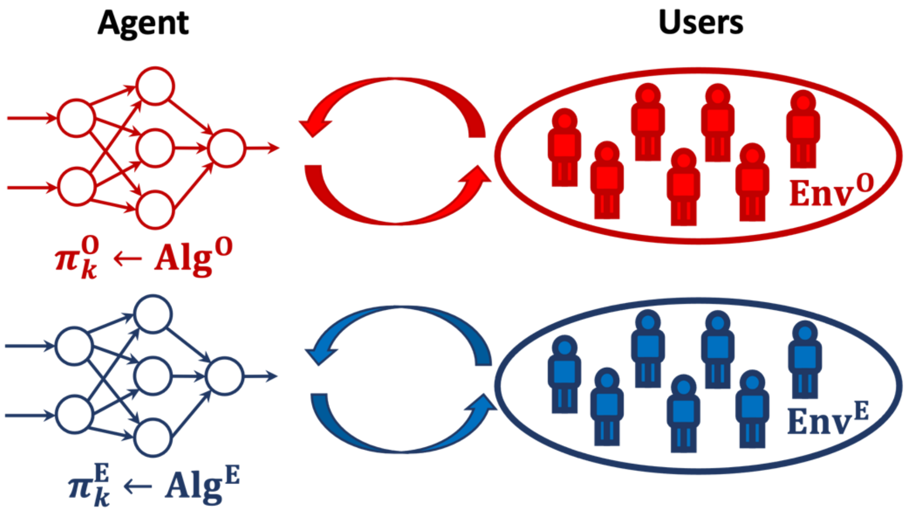
  $\pi_{\mathrm{O},k} \leftarrow \mathrm{Alg}^{\mathrm{O}}(D_k)$; $\pi_{\mathrm{E},k} \leftarrow \mathrm{Alg}^{\mathrm{E}}(D_k)$.

  $\pi_{\mathrm{O},k}/\pi_{\mathrm{E},k}$ interacts with users in exploration/exploitation tier, and collect data $\tau_{\mathrm{O},k}/\tau_{\mathrm{E},k}$.

  $D_{k+1} = D_k \cup \{\tau_{\mathrm{O},k}\} \cap \{\tau_{\mathrm{E},k}\}$.

end

Tiered Online RL Formulation (Ours)



Assume $\mathrm{Env}^{\mathrm{O}} = \mathrm{Env}^{\mathrm{E}}$;
Relaxtion of this assumption leave to future work.

Tiered RL Framework

Initialize $D_1 \leftarrow \{\}$.
for $k = 1, 2, ..., K$ **do**
  $\pi_{\mathrm{O},k} \leftarrow \mathrm{Alg}^{\mathrm{O}}(D_k)$; $\pi_{\mathrm{E},k} \leftarrow \mathrm{Alg}^{\mathrm{E}}(D_k)$.
  $\pi_{\mathrm{O},k}/\pi_{\mathrm{E},k}$ interacts with users in exploration/exploitation tier, and collect data $\tau_{\mathrm{O},k}/\tau_{\mathrm{E},k}$.
  $D_{k+1} = D_k \cup \{\tau_{\mathrm{O},k}\} \cap \{\tau_{\mathrm{E},k}\}$.
end

- Objective
  - Consider the pseudo-regret $\mathbf{Regret_K}(\cdot)$:
    - $\mathbf{Regret_K}(\mathbf{Alg^E}) \coloneqq \mathbf{E}\left[\sum_{k=1}^{K} V^*(s_1) - V^{\pi_k^{\mathrm{E}}}(s_1)\right]$; $\mathbf{Regret_K}(\mathbf{Alg^O}) \coloneqq \mathbf{E}\left[\sum_{k=1}^{K} V^*(s_1) - V^{\pi_k^{\mathrm{O}}}(s_1)\right]$.
  - Is it possible for **$\mathbf{Regret_K}(\mathbf{Alg^E})$ to be strictly lower than any online learning algorithms** in certain scenarios, while **keeping $\mathbf{Regret_K}(\mathbf{Alg^O})$ near-optimal**?

Benefits for G^{E} under our framework.

Not too much additional cost for G^{O}.

Highlight of Main Results

Normal Tabular RL Setting

- No benefits by comparing with standard online RL (from minimax optimality perspective)
  - $\min_{\mathrm{Alg}^{\mathrm{O}},\mathrm{Alg}^{\mathrm{E}}} \max_{\mathrm{MDP}} \mathrm{Regret}(\mathrm{Alg}^{\mathrm{E}}) \geq O(\sqrt{H^3SAK})$

Minimax lower bound of normal online RL setting

Tabular RL with Strictly Positive Gap

$\forall h,s,a$: $\Delta_h(s,a) = 0$ or $\Delta_h(s,a) \geq \Delta_{min} > 0$

where $\Delta_h(s,a) \coloneqq V_h^*(s) - Q_h^*(s,a)$

- By choosing:
  - **P**essimistic **V**alue **I**teration (PVI) as $\mathrm{Alg}^{\mathrm{E}}$,
  - Arbitrary online algorithm with near-optimal regret as $\mathrm{Alg}^{\mathrm{O}}$
- Guarantee
  - $\mathrm{Regret}_K(\mathrm{Alg}^{\mathrm{O}})$ keeps near-optimal.
  - $\mathrm{Regret}_K(\mathrm{Alg}^{\mathrm{E}})$ is constant/independent of K.

In contrast with $O(\log K)$ lower bound in online setting

Why Pessimistic Value Iteration?

- Why Pessimism?
  - Key property [Jin et. al., 2021]:
    - The more optimal trajectories occurs in D_k, the smaller $V^* - V^{\pi_k^{\mathrm{PVI}}}$ would be

Why Pessimistic Value Iteration?

- Why Pessimism?
  - Key property [Jin et. al., 2021]:
    - The more optimal trajectories occurs in D_k, the smaller $V^* - V^{\pi_k^{\mathrm{PVI}}}$ would be
  - Coincide with the optimality constraint of **$\mathbf{Alg^O}$**
    - Low regret of **$\mathbf{Alg^O}$** ⇔ Faster accumulation of optimal trajectories in D_k

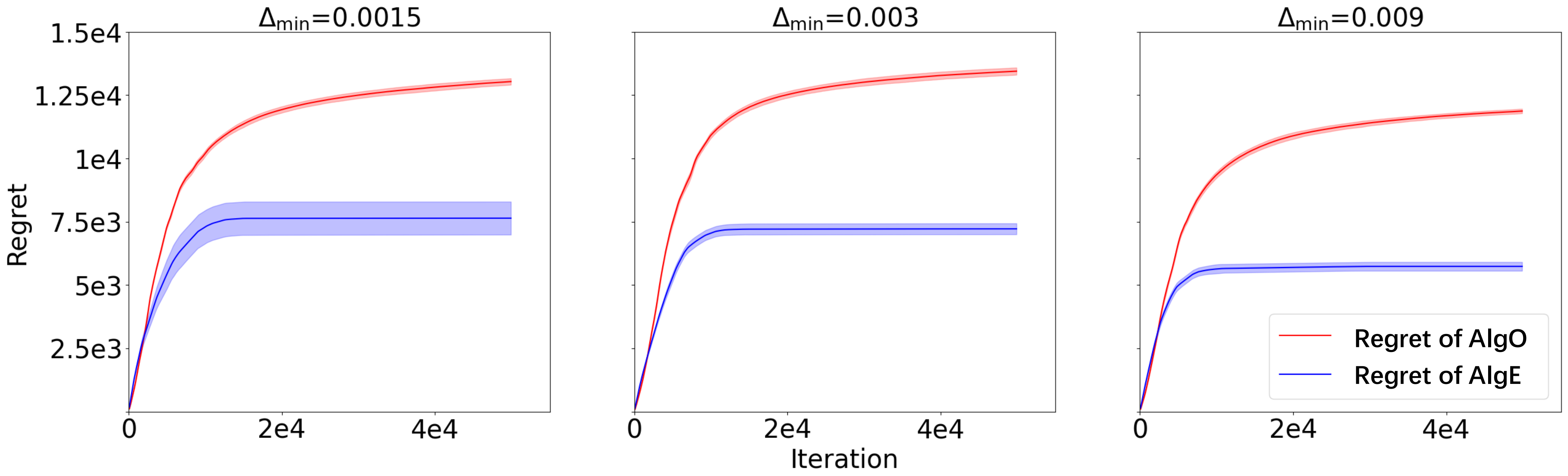
Why Pessimistic Value Iteration?

- Why Pessimism?
  - Key property [Jin et. al., 2021]:
    - The more optimal trajectories occurs in D_k, the smaller $V^* - V^{\pi_k^{\mathrm{PVI}}}$ would be
  - Coincide with the optimality constraint of **Alg$^{\mathbf{O}}$**
    - Low regret of **Alg$^{\mathbf{O}}$** $\Leftrightarrow$ Faster accumulation of optimal trajectories in D_k

Thanks to strictly positive gap, $V^* - V^{\pi_k^{\mathrm{PVI}}}$ will be zero after certain steps, which implies constant regret.

Verification Experiments in Tabular MDP

- S=A=H=5. Random generated transition/reward functions.
- $\mathrm{Alg}^{\mathrm{O}}$: StrongEulder [2]; $\mathrm{Alg}^{\mathrm{E}}$: PVI with Adaptive Bonus Term in [2]



[2] NeurIPS 2019, Max Simchowitz and Kevin Jamieson, Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs.

Thanks