Concurrent Reinforcement Learning with Aggregated States via Randomized Least Squares Value Iteration

Yan Chen

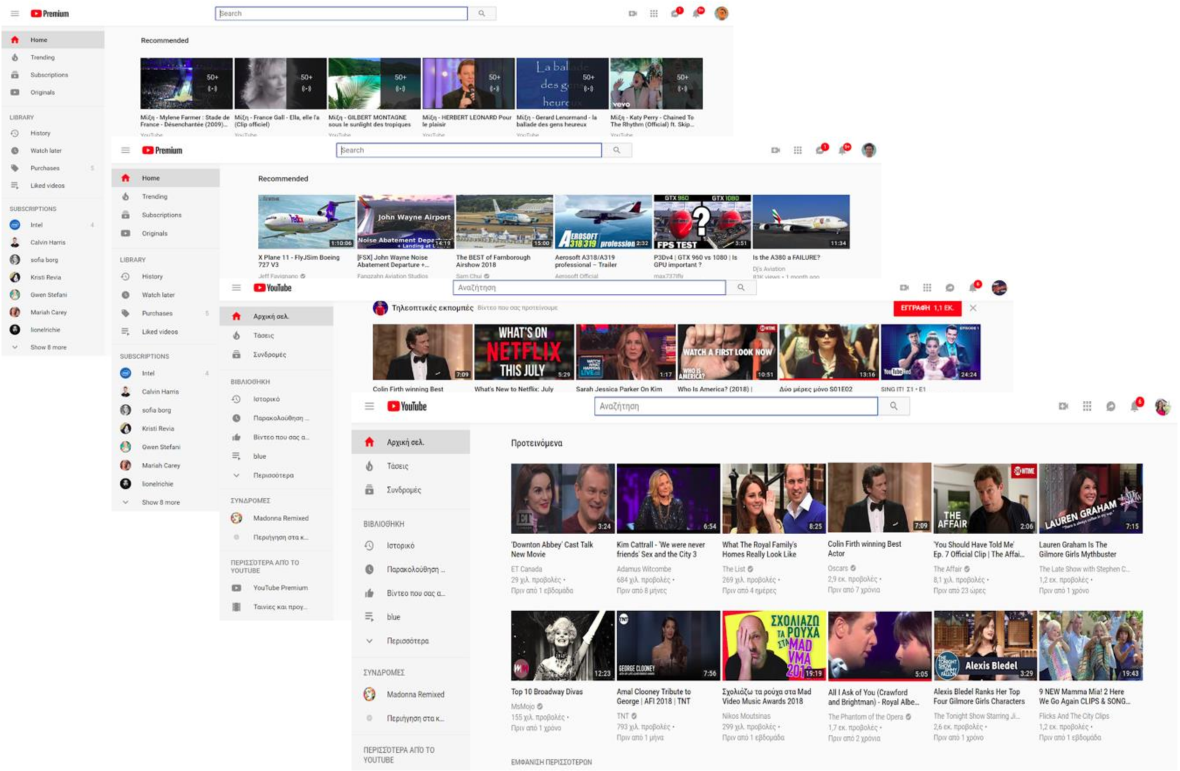
Duke University

Joint work with Qinxun Bai, Yiteng Zhang, Maria Dimakopoulou, Shi Dong, Qi Sun, Zhengyuan Zhou

Concurrent Reinforcement Learning

- Multi-agent Learning in the same environment

Google AI robot farm



Web services

Concurrent Reinforcement Learning Framework

Markov decision process (Γ aggregated states, N agents, S states, A actions)

Aggregate state-action pairs whose values are close

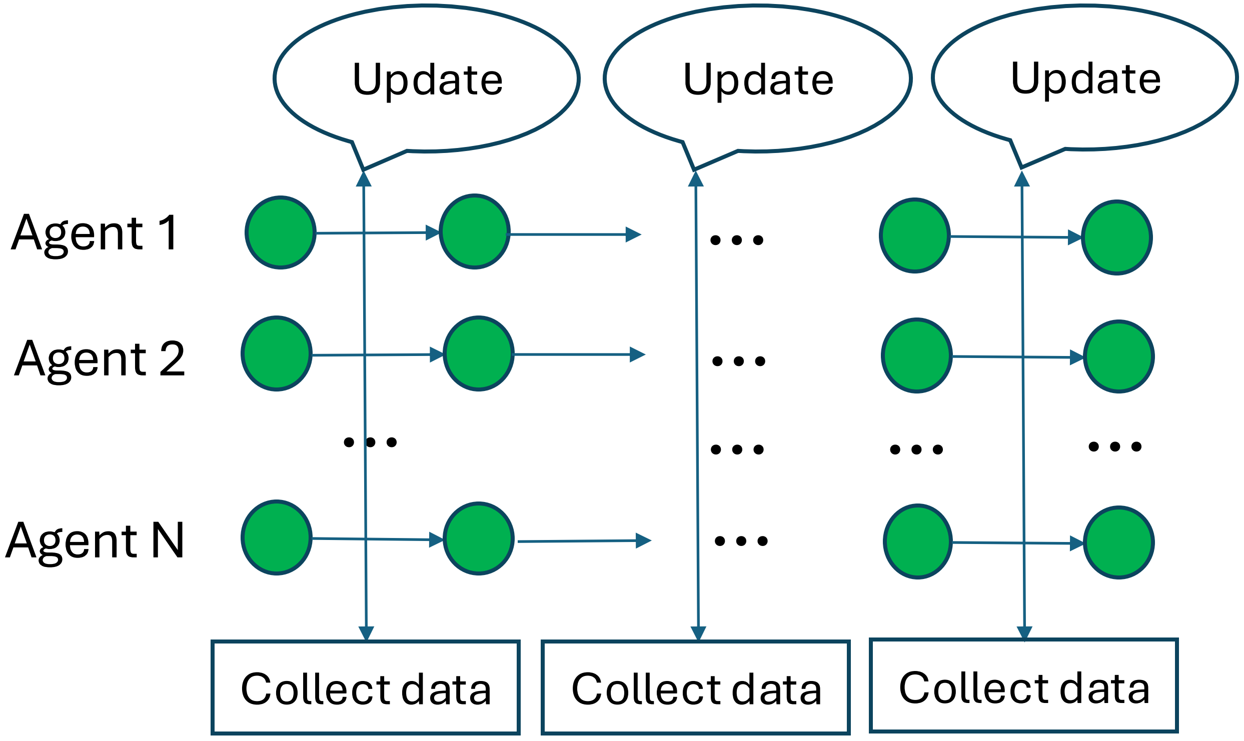
- reduce the computational cost when SA is large

Randomized least-squares value iteration (RLSVI) (Osband et al., 2019)

- injects Gaussian noise into the rewards
- learn a randomized value function from the perturbed dataset

Finite-horizon and infinite horizon cases: worst-case regret bound

Finite-horizon case



Algorithm 1: keep all historical data

- worst case regret bound: $\tilde{O}\left(H^{\frac{5}{2}}\Gamma\sqrt{KN}\right)$
- space complexity: $O(KHN)$

Algorithm 2: keep only the historical data from last episode

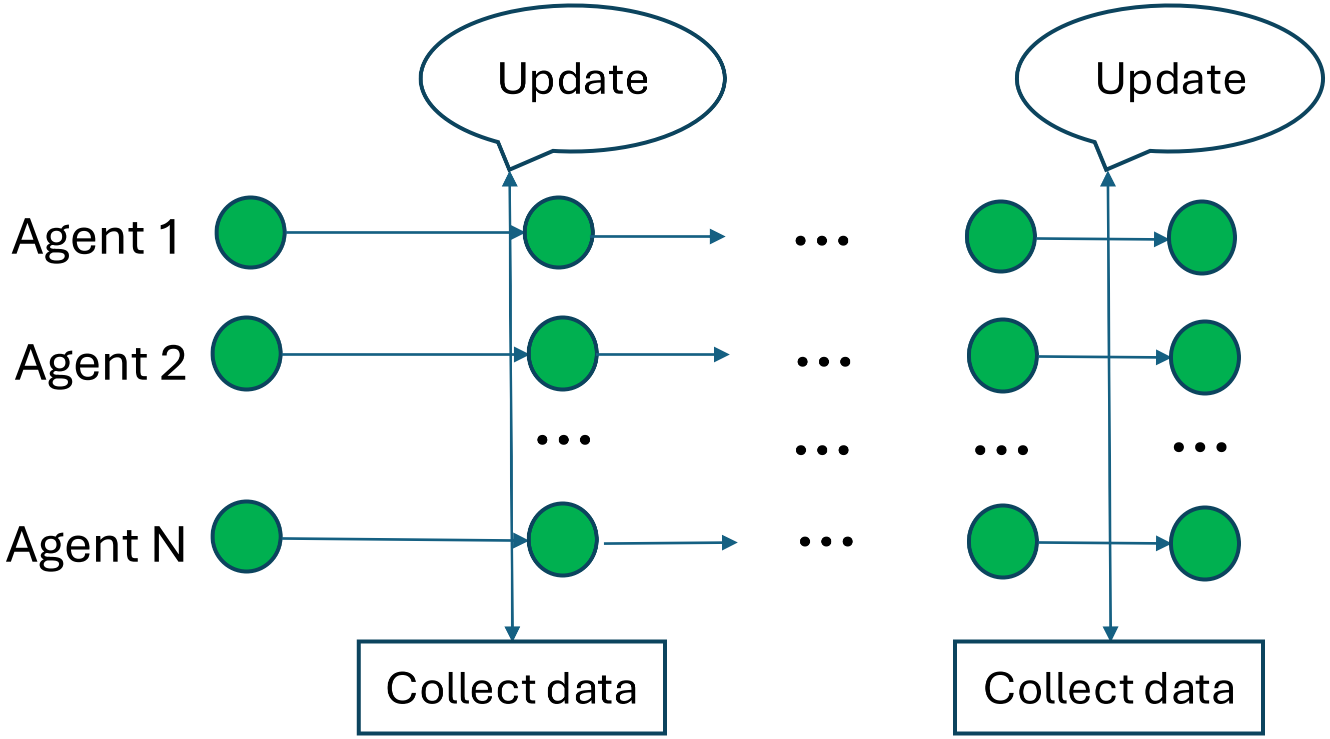
- Worst case regret bound: $\tilde{O}\left(KH^{\frac{5}{2}}\Gamma\sqrt{N}\right)$
- Space complexity: $O(HN)$

Finite-horizon: K-episode, H-horizon, N agents

Comparison with worst-case regret bounds from single-agent RLSVI

- (Russo, 2019): worst-case regret $\tilde{O}\left(H^3 S^{\frac{3}{2}} \sqrt{AK}\right)$
- (Agrawal et al. 2021): worst-case regret $\tilde{O}\left(H^{\frac{5}{2}} S \sqrt{AK}\right)$
- Both algorithms keep all historical data in the buffer
- N=1, our algorithm 1 (keep all historical data) gives worst-case regret bound $\tilde{O}\left(H^{\frac{5}{2}} \Gamma \sqrt{K}\right)$
- matching with (Agrawal et al. 2021) if Γ=SA and $S \approx \sqrt{\Gamma}$

Infinite-horizon case



Algorithm 1: keep all historical data

- worst case regret bound: $\tilde{O}(\sqrt{TN})$

Algorithm 2: keep only the historical data from last pseudo-episode

- worst case regret bound: $\tilde{O}(T\sqrt{N})$

Infinite-horizon (N agents)
generate pseudo-episodes using geometric distribution

Comparison Table

Table 1. Comparison of regret bounds for various RLSVI/LSVI algorithms

Agent	Setup	Algorithm	Regret Bound	Regret-Type	Data Stored	Numerical
Single	Tabular	RLSVI (Russo, 2019)	$\tilde{O}(H^3 S^{3/2}\sqrt{AK})$	Worst-case	All-history	N/A
Single	Tabular	RLSVI (Agrawal et al., 2021)	$\tilde{O}(H^{5/2} S\sqrt{AK})$	Worst-case	All-history	N/A
Multi	Tabular	Concurrent RLSVI (Taiga et al., 2022)	N/A	Bayes	All-history	Synthetic
Multi	Linear Functional Approximation	Concurrent LSVI (Desai et al., 2018)	$\tilde{O}(H^2\sqrt{d^3 KN})$	Worst-case	All-history	N/A
Multi	Linear Functional Approximation	Concurrent LSVI (Min et al., 2023)	$\tilde{O}(H\sqrt{dKN})$	Worst-case	All-history	N/A
Multi	Tabular	Concurrent RLSVI (ours-1)	$\tilde{O}(H^{5/2}\sqrt{KN})$	Worst-case	All-history	N/A
Multi	Tabular	Concurrent RLSVI (ours-2)	$\tilde{O}(H^{5/2} K\sqrt{N})$	Worst-case	One episode	Synthetic

Numerical Results