Refined Sample Complexity for Markov Games with Independent Linear Function Approximation

(Published as a conference paper at COLT 2024)

Yan Dai[1] Qiwen Cui[2] Simon S. Du[2]



[1]IIIS, Tsinghua University [2]University of Washington

Introduction

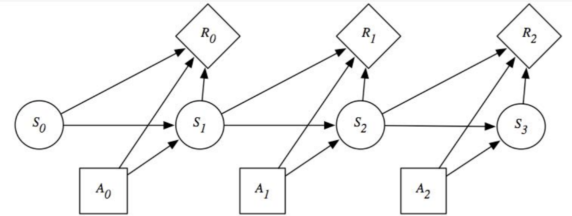
(Single-Agent) Reinforcement Learning

- Markov Decision Process (MDP): **Single** agent interacts for K episodes $\times$ H steps. **Single** state, **single** action action, **single** loss.

Algorithm Interaction Protocol in a MDP

---

for #episode $k = 1, 2, \dots, K$ **do**
    Agent reset to initial state $s_1 \in \mathcal{S}_1$ ▷ Assume $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_{H+1}$.
    for #step $h = 1, 2, \dots, H$ **do**
        Agent picks an action $a_h \in \mathcal{A}$ ▷ Sample from **policy** $\pi_k : \mathcal{S} \to \triangle(\mathcal{A})$.
        Agent observes loss $\ell(s_h, a_h)$
        Agent transits to $s_{h+1} \sim \mathbb{P}(\cdot \mid s_h, a_h)$



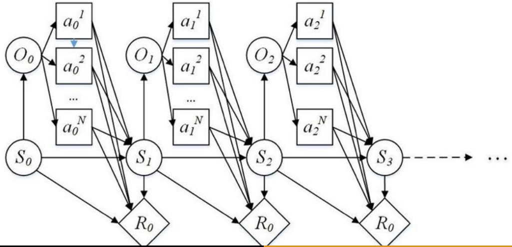
Multi-Agent Reinforcement Learning

- Markov Games (MG): **Multiple** agents interact for K episodes $\times$ H steps. **Single** state, **multiple** action, **multiple** loss.

Algorithm Interaction Protocol in a MG

---

for #episode $k = 1, 2, \ldots, K$ **do**
  Agents reset to initial state $s_1 \in \mathcal{S}_1$ ▷ Assume $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_{H+1}$.
  for #step $h = 1, 2, \ldots, H$ **do**
    Agents pick actions $a_h^1 \in \mathcal{A}^1, a_h^2 \in \mathcal{A}^2, \ldots, a_h^m \in \mathcal{A}^m$ ▷ Sample from a **joint policy** $\pi_k \colon \mathcal{S} \to \triangle(\mathcal{A}^1 \times \mathcal{A}^2 \times \cdots \times \mathcal{A}^m)$.
    Each agent observes loss $\ell^{\boldsymbol{i}}(s_h, a_h^1, a_h^2, \ldots, a_h^m)$ ▷ Loss depends on i
    Agent transits to $s_{h+1} \sim \mathbb{P}(\cdot \mid s_h, a_h^1, a_h^2, \ldots, a_h^m)$ ▷ Same new state s_{h+1}



Objective of Agents

Given *joint policy* $\pi \in \Pi = \{\pi \colon \mathcal{S} \to \triangle(\mathcal{A}^1 \times \mathcal{A}^2 \times \cdots \times \mathcal{A}^m)\}$, for each layer-$h$ state $s \in \mathcal{S}_h$, define *V-function* for each agent:

$$V_\pi^i(s) = \mathop{\mathbb{E}}_{(s_1,\mathbf{a}_1,s_2,\mathbf{a}_2,\ldots,s_H,\mathbf{a}_H)} \left[\sum_{h'=h}^{H} \ell^i(s_{h'}, \mathbf{a}_{h'}) \middle| s_h = s \right], \quad \forall i \in [m].$$

Objective of Agents

Given *joint policy* $\pi \in \Pi = \{\pi \colon \mathcal{S} \to \triangle(\mathcal{A}^1 \times \mathcal{A}^2 \times \cdots \times \mathcal{A}^m)\}$, for each layer-$h$ state $s \in \mathcal{S}_h$, define *V-function* for each agent:

$$V_\pi^i(s) = \mathop{\mathbb{E}}_{(s_1,\mathbf{a}_1,s_2,\mathbf{a}_2,\ldots,s_H,\mathbf{a}_H)} \left[\sum_{h'=h}^{H} \ell^i(s_{h'}, \mathbf{a}_{h'}) \middle| s_h = s \right], \quad \forall i \in [m].$$

Fixing $i \in [m]$, for opponents' policy π^{-i}, define *best response V*:

$$V_{\dagger,\pi^{-i}}^i(s) = \min_{\pi^i \in \Pi^i = \{\pi \colon \mathcal{S} \to \triangle(\mathcal{A}^i)\}} V_{\pi^i \circ \pi^{-i}}^i(s), \quad \forall i \in [m], s \in \mathcal{S}.$$

Objective of Agents

Given *joint policy* $\pi \in \Pi = \{\pi \colon \mathcal{S} \to \triangle(\mathcal{A}^1 \times \mathcal{A}^2 \times \cdots \times \mathcal{A}^m)\}$, for each layer-$h$ state $s \in \mathcal{S}_h$, define *V-function* for each agent:

$$V_\pi^i(s) = \mathop{\mathbb{E}}_{(s_1, \mathbf{a}_1, s_2, \mathbf{a}_2, \ldots, s_H, \mathbf{a}_H)} \left[\sum_{h'=h}^{H} \ell^i(s_{h'}, \mathbf{a}_{h'}) \middle| s_h = s \right], \quad \forall i \in [m].$$

Fixing $i \in [m]$, for opponents' policy π^{-i}, define *best response V*:

$$V_{\dagger,\pi^{-i}}^i(s) = \min_{\pi^i \in \Pi^i = \{\pi \colon \mathcal{S} \to \triangle(\mathcal{A}^i)\}} V_{\pi^i \circ \pi^{-i}}^i(s), \quad \forall i \in [m], s \in \mathcal{S}.$$

Policy $\pi \in \Pi$ is a ϵ*-Coarse Correlated Equilibrium* (ϵ*-CCE*) if

$$\max_{i \in [m]} \left\{ V_\pi^i(s_1) - V_{\dagger,\pi^{-i}}^i(s_1) \right\} \le \epsilon.$$

Agents **collaborate** to minimize #samples needed for finding an ϵ-CCE (*sample complexity*).

Previous Works on Linear Markov Games

Linear MG. $|\mathcal{S}| \gg 0$ but allows a d-dim'l linear structure s.t. every *Q-function* is linear in some known feature $\phi(s, a^i)$:

$$Q^i_{\pi^{-i}}(s, a^i) \triangleq \mathop{\mathbb{E}}_{a^{-i} \sim \pi^{-i}} \left[\ell^i(s, \mathbf{a}) + \mathop{\mathbb{E}}_{s' \sim \mathbb{P}(s, \mathbf{a})} \left[V^i(s') \right] \right],$$

where $V : \mathcal{S} \times [m] \to \mathbb{R}$ is an arbitrary next-layer V-function.

1. [Cui et al., 2023]: $\widetilde{\mathcal{O}}(\epsilon^{-4} d^4 H^{10} m^4)$.
2. [Wang et al., 2023]: $\widetilde{\mathcal{O}}(\epsilon^{-2} A^5_{\max} d^4 H^6 m^2)$.

Previous Works on Linear Markov Games

Linear MG. $|\mathcal{S}| \gg 0$ but allows a d-dim'l linear structure s.t. every *Q-function* is linear in some known feature $\phi(s, a^i)$:

$$Q^i_{\pi^{-i}}(s, a^i) \triangleq \mathop{\mathbb{E}}_{a^{-i}\sim\pi^{-i}}\left[\ell^i(s, \mathbf{a}) + \mathop{\mathbb{E}}_{s'\sim\mathbb{P}(s,\mathbf{a})}\left[V^i(s')\right]\right],$$

where $V : \mathcal{S} \times [m] \to \mathbb{R}$ is an arbitrary next-layer V-function.

1. [Cui et al., 2023]: $\widetilde{\mathcal{O}}(\epsilon^{-4} d^4 H^{10} m^4)$.
2. [Wang et al., 2023]: $\widetilde{\mathcal{O}}(\epsilon^{-2} A_{\max}^5 d^4 H^6 m^2)$.
3. [Fan et al., 2024] (concurrent): $\widetilde{\mathcal{O}}(\epsilon^{-2} d^2 H^6 m^2)$ (simulator).

Previous Works on Linear Markov Games

Linear MG. $|\mathcal{S}| \gg 0$ but allows a d-dim'l linear structure s.t. every *Q-function* is linear in some known feature $\phi(s, a^i)$:

$$Q^i_{\pi^{-i}}(s, a^i) \triangleq \mathop{\mathbb{E}}_{a^{-i}\sim\pi^{-i}}\left[\ell^i(s, \mathbf{a}) + \mathop{\mathbb{E}}_{s'\sim\mathbb{P}(s,\mathbf{a})}\left[V^i(s')\right]\right],$$

where $V : \mathcal{S} \times [m] \to \mathbb{R}$ is an arbitrary next-layer V-function.

1. [Cui et al., 2023]: $\widetilde{\mathcal{O}}(\epsilon^{-4} d^4 H^{10} m^4)$.
2. [Wang et al., 2023]: $\widetilde{\mathcal{O}}(\epsilon^{-2} A_{\max}^5 d^4 H^6 m^2)$.
3. [Fan et al., 2024] (concurrent): $\widetilde{\mathcal{O}}(\epsilon^{-2} d^2 H^6 m^2)$ (simulator).
4. (**Ours**): $\widetilde{\mathcal{O}}(\epsilon^{-2} m^4 d^5 H^6)$ – optimal ϵ^{-2} convergence, no poly$(A_{\max})$ dependency, no simulator! [1]

[1] We require a slightly stronger notion of linearity that transitions also are linear – see Linear MDPs vs Linear-Q MDPs in single-agent RL [Jin et al., 2020].

Our Algorithm

Main Insights

1. When designing the framework, **data-dependent (*i.e.*, random) estimators** for sub-optimality gaps can allow "good-in-expectation" plug-in algorithms.
2. When designing the plug-in algorithm, **action-dependent bonuses** can handle occasionally extreme estimation errors.

Data-Dep Sub-Opt Gap Est

Previous AVLPR Framework [Wang et al., 2023]

Algorithm AVLPR Framework (Informal) [Wang et al., 2023]

1: **for** $t = 1, 2, \ldots, T = \mathcal{O}(\epsilon^{-2})$ **do** ▷ Find an $\mathcal{O}(1/t)$-CCE with $\mathcal{O}(t^2)$ samples
2: Use potential function $\{\Psi_{t,h}^i\}_{t,h,i}$ to "lazily update" s.t. #updates $= \mathcal{O}(\log T)$.
3: **for** $h = H, H-1, \ldots, 1$ **do** ▷ Do policy improvement layer-by-layer
4: Call CCE-Approx_h for a $\tilde{\pi}_t$ s.t. $\text{SubOpt}^i(\tilde{\pi}_t, s) \le G_t^i(s)$ *w.h.p.*, where

$$G_t^i \text{ is deterministic s.t. } \sum_{i=1}^m \mathop{\mathbb{E}}_{s \sim_h \tilde{\pi}_t} \left[G_t^i(s)\right] \sim m\sqrt{1/t}.$$

5: Call V-Approx_h to estimate the current-layer V-function of $\tilde{\pi}_t$.

Issue? **Deterministic** sub-optimality gap estimation in Linear MGs

Previous AVLPR Framework [Wang et al., 2023]

Algorithm AVLPR Framework (Informal) [Wang et al., 2023]

for $t = 1, 2, \ldots, T = \mathcal{O}(\epsilon^{-2})$ **do** ▷ Find an $\mathcal{O}(1/t)$-CCE with $\mathcal{O}(t^2)$ samples
  Use potential function $\{\Psi_{t,h}^i\}_{t,h,i}$ to "lazily update" s.t. #updates $= \mathcal{O}(\log T)$.
  for $h = H, H-1, \ldots, 1$ **do** ▷ Do policy improvement layer-by-layer
    Call CCE-Approx_h for a $\tilde{\pi}_t$ s.t. $\text{SubOpt}^i(\tilde{\pi}_t, s) \le G_t^i(s)$ *w.h.p.*, where

$$G_t^i \text{ is deterministic s.t. } \sum_{i=1}^{m} \mathop{\mathbb{E}}_{s \sim_h \tilde{\pi}_t} \left[G_t^i(s)\right] \sim m\sqrt{1/t}.$$

    Call V-Approx_h to estimate the current-layer V-function of $\tilde{\pi}_t$.

Issue? **Deterministic** sub-optimality gap estimation in Linear MGs
$\Rightarrow$ **Open problem** of high-probability regret bounds for adversarial contextual linear bandits [Olkhovskaya et al., 2023]

Previous AVLPR Framework [Wang et al., 2023]

Algorithm AVLPR Framework (Informal) [Wang et al., 2023]

1: **for** $t = 1, 2, \ldots, T = \mathcal{O}(\epsilon^{-2})$ **do** ▷ Find an $\mathcal{O}(1/t)$-CCE with $\mathcal{O}(t^2)$ samples
2: Use potential function $\{\Psi^i_{t,h}\}_{t,h,i}$ to "lazily update" s.t. $\#\text{updates} = \mathcal{O}(\log T)$.
3: **for** $h = H, H-1, \ldots, 1$ **do** ▷ Do policy improvement layer-by-layer
4: Call $\text{CCE-A}\textsc{pprox}_h$ for a $\tilde{\pi}_t$ s.t. $\text{SubOpt}^i(\tilde{\pi}_t, s) \le G^i_t(s)$ *w.h.p.*, where

$$G^i_t \text{ is deterministic s.t. } \sum_{i=1}^m \mathop{\mathbb{E}}_{s \sim_h \tilde{\pi}_t} \left[G^i_t(s)\right] \sim m\sqrt{1/t}.$$

5: Call $\text{V-A}\textsc{pprox}_h$ to estimate the current-layer V-function of $\tilde{\pi}_t$.

Issue? **Deterministic** sub-optimality gap estimation in Linear MGs
⇒ **Open problem** of high-probability regret bounds for adversarial contextual linear bandits [Olkhovskaya et al., 2023]
⇒ Pure exploration deployed, resulting in **poly**$(A_{\max})$ **factors!**

Improved AVLPR Framework (**Ours**)

Algorithm Improved AVLPR Framework (Informal, **Ours**)

1: **for** $t = 1, 2, \ldots, T = \mathcal{O}(\epsilon^{-2})$ **do** ▷ Find an $\mathcal{O}(1/t)$-CCE with $\mathcal{O}(t^2)$ samples

2: Use potential function $\{\Psi^i_{t,h}\}_{t,h,i}$ to "lazily update" s.t. #updates $= \mathcal{O}(\log T)$.

3: **for** $h = H, H-1, \ldots, 1$ **do** ▷ Do policy improvement layer-by-layer

4: Call $\text{CCE-}\textsc{Approx}_h$ for a $\tilde{\pi}_t$ s.t. $\text{SubOpt}^i(\tilde{\pi}_t, s) \le \textsc{Gap}^i_t(s)$ *w.h.p.*, where

$$\textsc{Gap}^i_t \text{ is random variable s.t.} \sum_{i=1}^m \mathop{\mathbb{E}}_{s \sim_h \tilde{\pi}_t} \left[\mathop{\mathbb{E}}_{\textsc{Gap}} [\textsc{Gap}^i_t(s)] \right] \sim m\sqrt{1/t}.$$

5:

6: Call $\text{V-}\textsc{Approx}_h$ to estimate the current-layer V-function of $\tilde{\pi}_t$.

Improved AVLPR Framework (**Ours**)

Algorithm Improved AVLPR Framework (Informal, **Ours**)

1: **for** $t = 1, 2, \ldots, T = \mathcal{O}(\epsilon^{-2})$ **do** ▷ Find an $\mathcal{O}(1/t)$-CCE with $\mathcal{O}(t^2)$ samples
2: Use potential function $\{\Psi_{t,h}^i\}_{t,h,i}$ to "lazily update" s.t. #updates $= \mathcal{O}(\log T)$.
3: **for** $h = H, H-1, \ldots, 1$ **do** ▷ Do policy improvement layer-by-layer
4: Call CCE-Approx_h for a $\tilde{\pi}_t$ s.t. $\text{SubOpt}^i(\tilde{\pi}_t, s) \le \text{Gap}_t^i(s)$ *w.h.p.*, where

$$\text{Gap}_t^i \text{ is random variable s.t. } \sum_{i=1}^m \mathop{\mathbb{E}}_{s\sim_h \tilde{\pi}_t} \left[\mathop{\mathbb{E}}_{\text{Gap}} [\text{Gap}_t^i(s)] \right] \sim m\sqrt{1/t}.$$

5: Repeat Step 4 for $R = \mathcal{O}(\log \frac{1}{\delta})$ times, getting $(\tilde{\pi}_{t,r}, \text{Gap}_{t,r})_{r\in[R]}$. Set

$$(\tilde{\pi}_t(s), \text{Gap}_t(s)) \leftarrow (\tilde{\pi}_{t,r^*(s)}(s), \text{Gap}_{t,r^*(s)}(s)), \text{ where } r^*(s) = \mathop{\text{argmin}}_{r\in[R]} \sum_{i=1}^m \text{Gap}_{t,r}^i(s).$$

6: Call V-Approx_h to estimate the current-layer V-function of $\tilde{\pi}_t$.

Proposition. By Markov Inequality, Step 5 ensures *w.h.p.*
$\sum_{i=1}^m \text{Gap}_{t,r^*(s)}^i(s) \le 2\sum_{i=1}^m \mathbb{E}_{\text{Gap}}[\text{Gap}_t^i(s)], \forall s \in \mathcal{S}_h, i \in [m]$.

Why is Data-Dependent Sub-Optimality Gap Estimator Important?

- This removes the original assumption of $G_t^i(s)$ is deterministic.
- This bypasses the open problem of high-prob regret bound for adv. contextual linear bandits, avoiding poly($A_{\max}$) factors.
- This only causes $\mathcal{O}(\log \frac{1}{\delta}) = \widetilde{\mathcal{O}}(1)$ factor in sample complexity.

Action-Dependent Bonuses

CCE-Approx Subroutine

Objective. Find policy $\tilde{\pi}$ for layer $\mathcal{S}_h$ with $\mathcal{O}(\epsilon^{-2})$ samples s.t.

$$V_{\tilde{\pi}}^i(s) - V_{\dagger,\tilde{\pi}^{-i}}^i(s) \le \textsc{Gap}^i(s) \textit{ w.h.p.}, \quad \mathop{\mathbb{E}}_{s\sim\bar{\pi}}\left[\textsc{Gap}^i(s)\right] \lesssim \epsilon. \qquad (*)$$

CCE-Approx Subroutine

Objective. Find policy $\tilde{\pi}$ for layer $\mathcal{S}_h$ with $\mathcal{O}(\epsilon^{-2})$ samples s.t.

$$V_{\tilde{\pi}}^i(s) - V_{\dagger,\tilde{\pi}^{-i}}^i(s) \le \textsc{Gap}^i(s) \text{ w.h.p.}, \ \mathop{\mathbb{E}}_{s\sim\bar{\pi}}\left[\textsc{Gap}^i(s)\right] \lesssim \epsilon. \quad (*)$$

Regret-to-Sample-Complexity Conversion $\Rightarrow \forall\ i \in [m]$, do regret-minimization over $K = \mathcal{O}(\epsilon^{-2})$ episodes in an adversarial (other agents) contextual ($s \sim \bar{\pi}$) linear bandit (action be $\mathcal{A}^i$). If

$$\sum_{k=1}^K \mathop{\mathbb{E}}_{a^i\sim\pi_k^i(\cdot|s)}\left[L_k^i(s,a^i)\right] \le \widetilde{\textsc{Gap}}^i(s) \text{ w.h.p.}, \ \mathop{\mathbb{E}}_{s\sim\bar{\pi}}[\widetilde{\textsc{Gap}}^i(s)] = \widetilde{\mathcal{O}}(\sqrt{K}),$$

where $L_k^i(s,a^i) = \mathbb{E}_{a^{-i}\sim\pi_k^{-i}}\left[\ell^i(s,\mathbf{a}) + \mathbb{E}_{s'\sim\mathbb{P}(s,\mathbf{a})}[V^i(s')]\right]$, then setting $\tilde{\pi} = \frac{1}{K}\sum_{k=1}^K \pi_k$, $\textsc{Gap}^i(s) = \frac{1}{K}\widetilde{\textsc{Gap}}^i(s)$ ensures (*).

Challenge: Designing Bonuses to Cancel Est. Err.

To creaft $\widetilde{\text{GAP}}(s)$ for some $s \in \mathcal{S}_h$, we need to cancel the total estimation errors associated with the optimal action a^* on s.

1. **Classical Idea.** Use bonuses *w.h.p.* larger than estimation errors to cancel them.

Challenge: Designing Bonuses to Cancel Est. Err.

To creaft $\widetilde{\mathrm{GAP}}(s)$ for some $s \in \mathcal{S}_h$, we need to cancel the total estimation errors associated with the optimal action a^* on s.

1. **Classical Idea.** Use bonuses *w.h.p.* larger than estimation errors to cancel them. Suppose that $(\mathsf{EstErr}_k^i(s,a))_{k=1}^K$ is a stochastic process adapted to $(\mathcal{F}_k)_{k=0}^K$. Design $B_k^i(s,a)$ s.t. $\sum_{k=1}^K \mathsf{EstErr}_k^i(s,a^*) \le \sum_{k=1}^K B_k^i(s,a^*)$ *w.h.p.* for the unknown a^*, and $\sum_{k=1}^K \mathbb{E}_{a\sim\pi_k^i(\cdot|s)}[B_k^i(s,a)] = \widetilde{\mathcal{O}}(\sqrt{K})$.

Challenge: Designing Bonuses to Cancel Est. Err.

To creaft $\widetilde{\textsc{Gap}}(s)$ for some $s \in \mathcal{S}_h$, we need to cancel the total estimation errors associated with the optimal action a^* on s.

1. **Classical Idea.** Use bonuses *w.h.p.* larger than estimation errors to cancel them. Suppose that $(\mathsf{EstErr}_k^i(s,a))_{k=1}^K$ is a stochastic process adapted to $(\mathcal{F}_k)_{k=0}^K$. Design $B_k^i(s,a)$ s.t. $\sum_{k=1}^K \mathsf{EstErr}_k^i(s,a^*) \le \sum_{k=1}^K B_k^i(s,a^*)$ *w.h.p.* for the unknown a^*, and $\sum_{k=1}^K \mathbb{E}_{a\sim\pi_k^i(\cdot|s)}[B_k^i(s,a)] = \widetilde{\mathcal{O}}(\sqrt{K})$.
2. **Traditional Freedman.** As a^* is unknown, concentrate using $\sum_{k=1}^K \mathsf{EstErr}_k^i(s,a^*) \lesssim \sum_{k=1}^K \sqrt{\mathsf{Var}_k(\mathsf{EstErr}_k^i(s,a^*))} + \sup_{a\in\mathcal{A}^i} \max_{k\in[K]} |\mathsf{EstErr}_k^i(s,a)|$ – variance + magnitude.

Challenge: Designing Bonuses to Cancel Est. Err.

To creaft $\widetilde{\text{GAP}}(s)$ for some $s \in \mathcal{S}_h$, we need to cancel the total estimation errors associated with the optimal action a^* on s.

1. **Classical Idea.** Use bonuses *w.h.p.* larger than estimation errors to cancel them. Suppose that $(\mathsf{EstErr}_k^i(s,a))_{k=1}^K$ is a stochastic process adapted to $(\mathcal{F}_k)_{k=0}^K$. Design $B_k^i(s,a)$ s.t. $\sum_{k=1}^K \mathsf{EstErr}_k^i(s,a^*) \le \sum_{k=1}^K B_k^i(s,a^*)$ *w.h.p.* for the unknown a^*, and $\sum_{k=1}^K \mathbb{E}_{a\sim\pi_k^i(\cdot|s)}[B_k^i(s,a)] = \widetilde{\mathcal{O}}(\sqrt{K})$.
2. **Traditional Freedman.** As a^* is unknown, concentrate using $\sum_{k=1}^K \mathsf{EstErr}_k^i(s,a^*) \lesssim \sum_{k=1}^K \sqrt{\mathsf{Var}_k(\mathsf{EstErr}_k^i(s,a^*))} + \sup_{a\in\mathcal{A}^i} \max_{k\in[K]} |\mathsf{EstErr}_k^i(s,a)|$ – variance + magnitude.
3. **Issue.** Estimation errors on **rarely visited** (s,a) are large, *i.e.*, if $\mathsf{EstErr}_k^i(s,a) \le v^i(s,a), \forall k$, then $\sup_{a\in\mathcal{A}_i} v_k^i(s,a) = \widetilde{\mathcal{O}}(K)$, but on average, $\mathbb{E}_{a\sim\frac{1}{K}\sum_{k=1}^K \pi_k^i(s)}[v_k^i(s,a)] = \widetilde{\mathcal{O}}(\sqrt{K})$.

Action-Dependent Bonuses Technique

$\exists v^i(s,a) \geq B_k^i(s,a), \forall k,$ s.t. $\sup_{a \in \mathcal{A}_i} v^i(s,a) = \widetilde{\mathcal{O}}(K)$ (occasionally large)

but $\mathbb{E}_{a \sim \frac{1}{K}\sum_{k=1}^{K} \pi_k^i(s)} \left[\max_{k \in [K]} |\mathsf{EstErr}_k^i(s)|\right] = \widetilde{\mathcal{O}}(\sqrt{K})$ (on average moderate)

Action-Dependent Bonuses Technique

$$\exists v^i(s,a) \geq B_k^i(s,a), \forall k, \text{ s.t. } \sup_{a \in \mathcal{A}_i} v^i(s,a) = \widetilde{\mathcal{O}}(K) \quad \text{(occasionally large)}$$

$$\text{but} \quad \mathop{\mathbb{E}}_{a \sim \frac{1}{K}\sum_{k=1}^K \pi_k^i(s)} \left[\max_{k \in [K]} |\mathsf{EstErr}_k^i(s)| \right] = \widetilde{\mathcal{O}}(\sqrt{K}) \quad \text{(on average moderate)}$$

Action-Dependent Bonuses. Set bonuses such that $\forall a \in \mathcal{A}^i$:

$$B_k^i(s,a) \gtrsim \sum_{k=1}^K \sqrt{\mathsf{Var}_k(\mathsf{EstErr}_k^i(s,a))} + \frac{\max_{k \in [K]} |\mathsf{EstErr}_k^i(s,a)|}{K},$$

$$\Rightarrow \sum_{k=1}^K \mathsf{EstErr}_k^i(s,a^*) \leq \sum_{k=1}^K B_k^i(s,a^*) \text{ \textit{w.h.p.} regardless of } a^* \in \mathcal{A}^i,$$

$$\sum_{k=1}^K \mathop{\mathbb{E}}_{a \sim \pi_k^i(\cdot|s)} [B_k^i(s,a)] = \sum_{k=1}^K \mathop{\mathbb{E}}_{a \sim \pi_k^i(\cdot|s)} \left[\sqrt{\mathsf{Var}_k(\mathsf{EstErr}_k^i(s,a))} \right] + \underbrace{\widetilde{\mathcal{O}}(\sqrt{K})}_{\text{Replace } \widetilde{\mathcal{O}}(K)!} .$$

Other Techniques Adopted into This Paper

1. Magnitude-Reduced Estimator [Dai et al., 2023], moving loss estimations from $[-\widetilde{\mathcal{O}}(K), \widetilde{\mathcal{O}}(K)]$ to $[-\widetilde{\mathcal{O}}(\sqrt{K}), \widetilde{\mathcal{O}}(K)]$.
2. Adaptive Freedman Inequality [Zimmert and Lattimore, 2022], removing deterministic magnitude upper bounds in Freedman.
3. Refined Covariance Estimation Analysis [Liu et al., 2023], ensuring $\mathrm{Tr}(\widehat{\Sigma}^{-1/2}(\widehat{\Sigma} - \Sigma)) = \widetilde{\mathcal{O}}(n^{-1/2})$ where n is #samples.

Read our paper at https://arxiv.org/pdf/2402.07082v2 for details!

Questions are more than welcomed!

References I

Cui, Q., Zhang, K., and Du, S. (2023).
Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation.
In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 2651–2652. PMLR.

Dai, Y., Luo, H., Wei, C.-Y., and Zimmert, J. (2023).
Refined regret for adversarial mdps with linear function approximation.
In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 6726–6759. PMLR.

Fan, J., Han, Y., Zeng, J., Cai, J.-F., Wang, Y., Xiang, Y., and Zhang, J. (2024).
Rl in markov games with independent function approximation: Improved sample complexity bound under the local access model.
In *International Conference on Artificial Intelligence and Statistics*, pages 2035–2043. PMLR.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020).
Provably efficient reinforcement learning with linear function approximation.
In *Conference on Learning Theory*, pages 2137–2143. PMLR.

References II

Liu, H., Wei, C.-Y., and Zimmert, J. (2023).
Bypassing the simulator: Near-optimal adversarial linear contextual bandits.
arXiv preprint arXiv:2309.00814.

Olkhovskaya, J., Mayo, J., van Erven, T., Neu, G., and Wei, C.-Y. (2023).
First-and second-order bounds for adversarial linear contextual bandits.
arXiv preprint arXiv:2305.00832.

Wang, Y., Liu, Q., Bai, Y., and Jin, C. (2023).
Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation.
In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 2793–2848. PMLR.

Zimmert, J. and Lattimore, T. (2022).
Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits.
In *Conference on Learning Theory*, pages 3285–3312. PMLR.