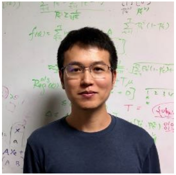
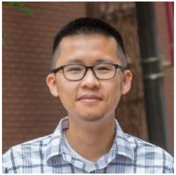
Refined Regret for Adversarial MDPs with Linear Function Approximation

(Published as a conference paper at ICML 2023)

Yan Dai[1] Haipeng Luo[2] Chen-Yu Wei[3] Julian Zimmert[4]





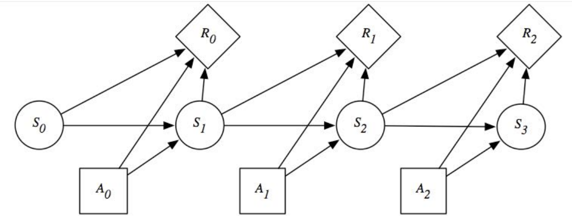
[1]IIIS, Tsinghua [2]USC [3]University of Virginia [4]Google Research

Table of Contents

Adversarial Markov Decision Process (AMDP)

Algorithm Interaction Protocol in AMDP

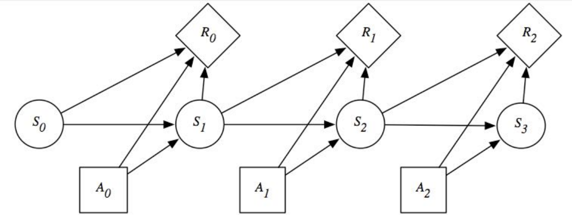
for #episode $k = 1, 2, \ldots, K$ **do**
  Agent reset to an initial state $s_1 \in \mathcal{S}_1$ ▷ Let $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_{H+1}$.
  for #step $h = 1, 2, \ldots, H$ **do**
    Agent picks an action $a_h \in \mathcal{A}$ ▷ Sample from **policy**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



Adversarial Markov Decision Process (AMDP)

Algorithm Interaction Protocol in AMDP

```
for #episode k = 1, 2, ..., K do
    Agent reset to an initial state s_1 ∈ S_1          ▷ Let S = S_1 ∪ S_2 ∪ ··· ∪ S_{H+1}.
    for #step h = 1, 2, ..., H do
        Agent picks an action a_h ∈ A                  ▷ Sample from policy π_k : S → △(A).
        Agent observes loss ℓ_{k,h}(s_h, a_h)          ▷ Loss ℓ depends on #episode k!
        Agent transits to s_{h+1} ~ P(· | s_h, a_h)    ▷ Transition P independent to k.
```



- Agent essentially decides K **policies** $\{\pi_k : \mathcal{S} \to \triangle(\mathcal{A})\}_{k=1}^{K}$.

Agent's Goal?

For the k-th episode, define **V-function** of policy $\pi\colon \mathcal{S} \to \triangle(\mathcal{A})$ as

$$V_k^\pi(s_1) = \mathbb{E}\left[\sum_{h=1}^{H} \ell_k(s_h, a_h) \middle| a_h \sim \pi_k(\cdot \mid s_h), s_{h+1} \sim \mathbb{P}(\cdot \mid s_h, a_h)\right].$$

Agent's Goal?

For the k-th episode, define **V-function** of policy $\pi\colon \mathcal{S} \to \triangle(\mathcal{A})$ as

$$V_k^\pi(s_1) = \mathbb{E}\left[\sum_{h=1}^{H} \ell_k(s_h, a_h) \middle| a_h \sim \pi_k(\cdot \mid s_h), s_{h+1} \sim \mathbb{P}(\cdot \mid s_h, a_h)\right].$$

The agent minimizes the expected total loss $\mathbb{E}[\sum_{k=1}^K V_k^{\pi_k}(s_1)]$. Or equivalently, minimize the **total regret**:

$$\mathcal{R}_K \triangleq \mathbb{E}\left[\sum_{k=1}^{K} V_k^{\pi_k}(s_1)\right] - \min_{\pi^*\colon \mathcal{S}\to\triangle(\mathcal{A})} \left\{\sum_{k=1}^{K} V_k^{\pi^*}(s_1)\right\}.$$

Agent's Goal?

For the k-th episode, define **V-function** of policy $\pi\colon \mathcal{S} \to \triangle(\mathcal{A})$ as

$$V_k^\pi(s_1) = \mathbb{E}\left[\sum_{h=1}^{H} \ell_k(s_h, a_h) \middle| a_h \sim \pi_k(\cdot \mid s_h), s_{h+1} \sim \mathbb{P}(\cdot \mid s_h, a_h)\right].$$

The agent minimizes the expected total loss $\mathbb{E}[\sum_{k=1}^K V_k^{\pi_k}(s_1)]$. Or equivalently, minimize the **total regret**:

$$\mathcal{R}_K \triangleq \mathbb{E}\left[\sum_{k=1}^{K} V_k^{\pi_k}(s_1)\right] - \min_{\pi^*\colon \mathcal{S}\to\triangle(\mathcal{A})} \left\{\sum_{k=1}^{K} V_k^{\pi^*}(s_1)\right\}.$$

	Full Information	Bandit Feedback
Known Transition	$\widetilde{\mathcal{O}}(H\sqrt{\mathbf{K}})$ [Zimin and Neu, 2013]	$\widetilde{\mathcal{O}}(\sqrt{HSA}\sqrt{\mathbf{K}})$ [Zimin and Neu, 2013]
Unknown Transition	$\widetilde{\mathcal{O}}(HS\sqrt{A}\sqrt{\mathbf{K}})$ [Rosenberg and Mansour, 2019]	$\widetilde{\mathcal{O}}(HS\sqrt{A}\sqrt{\mathbf{K}})$ [Jin et al., 2020]

Table: Previous Results on AMDP (w/o Function Approximation)

(K: No. of episodes; H: No. of steps; S: Size of $\mathcal{S}$; A: Size of $\mathcal{A}$)

AMDP with Linear Function Approximation

What if $\mathcal{S}$ can be prohibitively large?

AMDP with Linear Function Approximation

What if $\mathcal{S}$ can be prohibitively large?
Linear-Q AMDP: $\forall k \in [K]$, $\pi \colon \mathcal{S} \to \triangle(\mathcal{A})$, $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$Q_k^\pi(s,a) \triangleq \ell_k(s,a) + \mathop{\mathbb{E}}_{s' \sim \mathbb{P}(\cdot|s,a),\ a' \sim \pi(\cdot|s')} \left[Q_k^\pi(s',a') \right] \text{ is linear,}$$

i.e., $Q_k^\pi(s,a) = \langle \phi(s,a), \theta_k^\pi \rangle$ where $\phi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is **known**.

AMDP with Linear Function Approximation

What if $\mathcal{S}$ can be prohibitively large?

Linear-Q AMDP: $\forall k \in [K]$, $\pi\colon \mathcal{S} \to \triangle(\mathcal{A})$, $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$Q_k^\pi(s,a) \triangleq \ell_k(s,a) + \mathop{\mathbb{E}}_{s' \sim \mathbb{P}(\cdot|s,a),\ a' \sim \pi(\cdot|s')} \left[Q_k^\pi(s',a') \right] \text{ is linear,}$$

i.e., $Q_k^\pi(s,a) = \langle \phi(s,a), \theta_k^\pi \rangle$ where $\phi\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is **known**.

Some stronger variants of Linear-Q AMDP:

Linear MDP.	$\mathbb{P}(s' \mid s,a) = \langle \phi(s,a), \nu(s') \rangle$	(ϕ known but ν unknown).
Linear-Mixture MDP.	$\mathbb{P}(s' \mid s,a) = \langle \psi(s' \mid s,a), \nu \rangle$	(ψ known but ν unknown).
Linear Kernel MDP.	$\mathbb{P}(s' \mid s,a) = \langle \phi(s,a), M, \psi(s') \rangle$	(ϕ, ψ known but M unknown).

Previous Results on Linear-Q AMDPs

Setting	Assumption	Regret	
Linear-Q AMDP (with Simulator)	None	$\widetilde{\mathcal{O}}(d^{2/3}H^2\mathbf{K^{2/3}})$	[Luo et al., 2021a]
	Exploratory Policy	$\widetilde{\mathcal{O}}(\text{poly}(d,H)(\mathbf{K}/\lambda_0)^{1/2})$	[Luo et al., 2021a]
	None	$\widetilde{\mathcal{O}}(A^{1/2}d^{1/2}H^3\mathbf{K^{1/2}})$	(**This paper!**)
	None	$\widetilde{\mathcal{O}}(d^{1/2}H^3\mathbf{K^{1/2}})$	(**This paper!**)

Table: Previous Results on Linear-Q AMDPs.

(d: Dim. of ϕ; A: Size of $\mathcal{A}$; λ_0: Property of exploratory policy.)

Previous Results on Linear-Q AMDPs

Setting	Assumption	Regret	
Linear-Q AMDP (with Simulator)	None	$\widetilde{\mathcal{O}}(d^{2/3}H^2\mathbf{K}^{2/3})$	[Luo et al., 2021a]
	Exploratory Policy	$\widetilde{\mathcal{O}}(\text{poly}(d,H)(\mathbf{K}/\lambda_0)^{1/2})$	[Luo et al., 2021a]
	None	$\widetilde{\mathcal{O}}(A^{1/2}d^{1/2}H^3\mathbf{K}^{1/2})$	(**This paper!**)
	None	$\widetilde{\mathcal{O}}(d^{1/2}H^3\mathbf{K}^{1/2})$	(**This paper!**)

Table: Previous Results on Linear-Q AMDPs.

(d: Dim. of ϕ; A: Size of $\mathcal{A}$; λ_0: Property of exploratory policy.)

The first to get $\widetilde{\mathcal{O}}(\sqrt{K})$ regret w/o additional assumptions!

Previous Results on Other Variants

Setting	Assumption	Regret	
Linear-Mixture AMDP	**Full Information**	$\widetilde{\mathcal{O}}(dH\mathbf{K}^{1/2})$	[He et al., 2022]
	None	$\widetilde{\mathcal{O}}(d\mathbf{S}^2\mathbf{K}^{1/2}+\sqrt{H\mathbf{S}A}\mathbf{K}^{1/2})$	[Zhao et al., 2022]
Linear AMDP	**Known Transition**	$\widetilde{\mathcal{O}}(\text{poly}(d,H)(\mathbf{K}/\lambda_0)^{1/2})$	[Neu and Olkhovskaya, 2021]
	None	$\widetilde{\mathcal{O}}(d^2H^4\mathbf{K}^{14/15})$	[Luo et al., 2021b]
	Exploratory Policy	$\widetilde{\mathcal{O}}(\text{poly}(d,H)(\mathbf{K}/\lambda_0^{2/3})^{6/7})$	[Luo et al., 2021a]
	None	$\widetilde{\mathcal{O}}(\text{poly}(A,d,H)\mathbf{K}^{8/9})$	(**This paper!**)

Table: Previous Results on Other Variants of Linear-Q AMDPs.
(d: Dim. of ϕ; A: Size of $\mathcal{A}$; λ_0: Property of exploratory policy.)

Previous Results on Other Variants

Setting	Assumption	Regret	
Linear-Mixture AMDP	**Full Information**	$\widetilde{\mathcal{O}}(dH\mathbf{K}^{1/2})$	[He et al., 2022]
	None	$\widetilde{\mathcal{O}}(d\mathbf{S}^2\mathbf{K}^{1/2} + \sqrt{HSA}\mathbf{K}^{1/2})$	[Zhao et al., 2022]
Linear AMDP	**Known Transition**	$\widetilde{\mathcal{O}}(\text{poly}(d, H)(\mathbf{K}/\lambda_0)^{1/2})$	[Neu and Olkhovskaya, 2021]
	None	$\widetilde{\mathcal{O}}(d^2H^4\mathbf{K}^{14/15})$	[Luo et al., 2021b]
	Exploratory Policy	$\widetilde{\mathcal{O}}(\text{poly}(d, H)(\mathbf{K}/\lambda_0^{2/3})^{6/7})$	[Luo et al., 2021a]
	None	$\widetilde{\mathcal{O}}(\text{poly}(A, d, H)\mathbf{K}^{8/9})$	(**This paper!**)

Table: Previous Results on Other Variants of Linear-Q AMDPs.
(d: Dim. of ϕ; A: Size of $\mathcal{A}$; λ_0: Property of exploratory policy.)

Greatly outperform previous works on Linear AMDPs!

Overview of Our Algorithms

3 Algorithms, 3 New Techniques.

1. **Algorithm 1**: $\widetilde{\mathcal{O}}(\sqrt{AdH^6K})$ in Linear-Q AMDPs
   - FTRL w/ Log-Barrier on **Arbitrary** Losses.
2. **Algorithm 2**: $\widetilde{\mathcal{O}}(\sqrt{\quad dH^6K})$ in Linear-Q AMDPs
   - Magnitude-Reduced Estimator for **Any** Random Variable.
3. **Algorithm 3**: $\widetilde{\mathcal{O}}(\text{poly}(A, d, H)K^{8/9})$ in Linear AMDPs:
   - **Relative Concentration** Bounds for Stochastic Matrices.

Recap of FTRL Framework

Follow-the-Regularized-Leader (FTRL) Framework: For any loss estimation sequence $\{\hat{\ell}_t\}_{t=1}^T$, calculate actions $\{x_t \in \triangle(\mathcal{A})\}_{t=1}^T$ as

$$x_t = \mathop{\arg\min}_{x \in \triangle(\mathcal{A})} \left\{ \eta \left\langle x, \sum_{\tau=1}^{t-1} \ell_\tau \right\rangle + \Psi(x) \right\}, \quad t = 1, 2, \ldots, T.$$

Recap of FTRL Framework

Follow-the-Regularized-Leader (FTRL) Framework: For any loss estimation sequence $\{\hat{\ell}_t\}_{t=1}^T$, calculate actions $\{x_t \in \triangle(\mathcal{A})\}_{t=1}^T$ as

$$x_t = \operatorname*{arg\,min}_{x \in \triangle(\mathcal{A})} \left\{ \eta \left\langle x, \sum_{\tau=1}^{t-1} \ell_\tau \right\rangle + \Psi(x) \right\}, \quad t = 1, 2, \ldots, T.$$

Lemma (Classical Regret Guarantee on FTRL; Informal)

For "good enough" Ψ and ***losses such that*** $\hat{\ell}_{t,a} \geq -1/\eta$ ***for all*** $t = 1, 2, \ldots, T$ ***and*** $a \in \mathcal{A}$, *Eq. (1) holds for any fixed* $y \in \triangle(\mathcal{A})$.

$$\sum_{t=1}^T \langle x_t - y, \hat{\ell}_t \rangle \leq \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^T \sum_{a \in \mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \qquad (1)$$

What's the Issue?

In [Luo et al., 2021b], the final regret bound consists of

$$\widetilde{\mathcal{O}}\left(\beta K + \frac{1}{\eta} + \frac{\gamma}{\beta} K + \frac{\beta}{\gamma}\right),$$

where η is learning rate of FTRL, β is bonus coefficient, and γ is regularization factor (so the estimated loss $\hat{\ell} \in [-\gamma^{-1}, \gamma^{-1}]$).

What's the Issue?

In [Luo et al., 2021b], the final regret bound consists of

$$\widetilde{\mathcal{O}}\left(\beta K + \frac{1}{\eta} + \frac{\gamma}{\beta} K + \frac{\beta}{\gamma}\right),$$

where η is learning rate of FTRL, β is bonus coefficient, and γ is regularization factor (so the estimated loss $\hat{\ell} \in [-\gamma^{-1}, \gamma^{-1}]$).

How to get $\widetilde{\mathcal{O}}(\sqrt{K})$ regret?

What's the Issue?

In [Luo et al., 2021b], the final regret bound consists of

$$\widetilde{\mathcal{O}}\left(\beta K + \frac{1}{\eta} + \frac{\gamma}{\beta} K + \frac{\beta}{\gamma}\right),$$

where η is learning rate of FTRL, β is bonus coefficient, and γ is regularization factor (so the estimated loss $\hat{\ell} \in [-\gamma^{-1}, \gamma^{-1}]$).

How to get $\widetilde{\mathcal{O}}(\sqrt{K})$ regret?

Set $\beta = K^{-1/2}$ and $\eta = K^{-1/2} \Longrightarrow$ we need $\gamma = K^{-1}$!

What's the Issue?

In [Luo et al., 2021b], the final regret bound consists of

$$\widetilde{\mathcal{O}}\left(\beta K + \frac{1}{\eta} + \frac{\gamma}{\beta} K + \frac{\beta}{\gamma}\right),$$

where η is learning rate of FTRL, β is bonus coefficient, and γ is regularization factor (so the estimated loss $\hat{\ell} \in [-\gamma^{-1}, \gamma^{-1}]$).

How to get $\widetilde{\mathcal{O}}(\sqrt{K})$ regret?

Set $\beta = K^{-1/2}$ and $\eta = K^{-1/2} \Longrightarrow$ we need $\gamma = K^{-1}$!
But... we also need $\hat{\ell} \geq -1/\eta = -\sqrt{K}$ to ensure Eq. (1).

What's the Issue?

In [Luo et al., 2021b], the final regret bound consists of

$$\widetilde{\mathcal{O}}\left(\beta K + \frac{1}{\eta} + \frac{\gamma}{\beta} K + \frac{\beta}{\gamma}\right),$$

where η is learning rate of FTRL, β is bonus coefficient, and γ is regularization factor (so the estimated loss $\hat{\ell} \in [-\gamma^{-1}, \gamma^{-1}]$).

How to get $\widetilde{\mathcal{O}}(\sqrt{K})$ regret?

Set $\beta = K^{-1/2}$ and $\eta = K^{-1/2} \Longrightarrow$ we need $\gamma = K^{-1}$!
But... we also need $\hat{\ell} \geq -1/\eta = -\sqrt{K}$ to ensure Eq. (1).
So we essentially need $\gamma^{-1} \leq \eta^{-1}$ – that's why [Luo et al., 2021b] set $\beta = K^{-1/3}$, $\eta = K^{-2/3}$, $\gamma = K^{-2/3}$ for $\widetilde{\mathcal{O}}(K^{2/3})$ regret. ☹

How to Resolve?

Lemma (Classical Regret Guarantee on FTRL; Informal)

*For "good enough" Ψ and **losses such that** $\hat{\ell}_{t,a} \ge -1/\eta$ **for all** $t = 1, 2, \ldots, T$ **and** $a \in \mathcal{A}$, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \le \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \quad (1)$$

How to Resolve?

Lemma (Classical Regret Guarantee on FTRL; Informal)

*For "good enough" Ψ and **losses such that $\hat{\ell}_{t,a} \geq -1/\eta$ for all $t = 1, 2, \ldots, T$ and $a \in \mathcal{A}$**, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \leq \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \quad (1)$$

Lemma (Our Regret Guarantee on FTRL; Informal)

*For **log-barrier** Ψ (defined as $\Psi(x) = \sum_{a \in \mathcal{A}} \ln x_a^{-1}$) and **any real loss vectors $\ell_1, \ell_2, \ldots, \hat{\ell}_t$**, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

How to Resolve?

Lemma (Classical Regret Guarantee on FTRL; Informal)

*For "good enough" Ψ and **losses such that $\hat{\ell}_{t,a} \ge -1/\eta$ for all $t = 1, 2, \ldots, T$ and $a \in \mathcal{A}$**, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \le \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \qquad (1)$$

Lemma (Our Regret Guarantee on FTRL; Informal)

*For **log-barrier** Ψ (defined as $\Psi(x) = \sum_{a \in \mathcal{A}} \ln x_a^{-1}$) and **any real loss vectors** $\ell_1, \ell_2, \ldots, \hat{\ell}_t$, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

In this way, we no longer need $\gamma^{-1} \le \eta^{-1}$ and get the **first-ever $\widetilde{\mathcal{O}}(K^{1/2})$ regret** via $\beta = K^{-1/2}$, $\eta = K^{-1/2}$, $\gamma = K^{-1/2}$! ☺

Downside of the Previous Approach?

We can only use the **log-barrier** regularizer $\Psi(x) = \sum_{a \in \mathcal{A}} \ln x_a^{-1}$.

Downside of the Previous Approach?

We can only use the **log-barrier** regularizer $\Psi(x) = \sum_{a\in\mathcal{A}} \ln x_a^{-1}$. Compared to the original choice **negative-entropy** regularizer $\Psi(x) = \sum_{a\in\mathcal{A}} x_a \ln x_a$, it has **unavoidable poly**(A) **factors**!

Downside of the Previous Approach?

We can only use the **log-barrier** regularizer $\Psi(x) = \sum_{a\in\mathcal{A}} \ln x_a^{-1}$. Compared to the original choice **negative-entropy** regularizer $\Psi(x) = \sum_{a\in\mathcal{A}} x_a \ln x_a$, it has **unavoidable poly(A) factors**!

Lemma (Classical Regret Guarantee on FTRL; Informal)

For "good enough" Ψ and ***losses such that $\hat{\ell}_{t,a} \ge -1/\eta$ for all $t = 1, 2, \ldots, T$ and $a \in \mathcal{A}$****, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \le \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a\in\mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \tag{1}$$

Can we still use the original lemma (to use negative-entropy Ψ and avoid poly(A)), but instead reducing the magnitude of $\hat{\ell}$?

Downside of the Previous Approach?

We can only use the **log-barrier** regularizer $\Psi(x) = \sum_{a \in \mathcal{A}} \ln x_a^{-1}$. Compared to the original choice **negative-entropy** regularizer $\Psi(x) = \sum_{a \in \mathcal{A}} x_a \ln x_a$, it has **unavoidable poly(A) factors**!

Lemma (Classical Regret Guarantee on FTRL; Informal)

*For "good enough" Ψ and **losses such that $\hat{\ell}_{t,a} \ge -1/\eta$ for all $t = 1, 2, \ldots, T$ and $a \in \mathcal{A}$**, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \le \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \tag{1}$$

Can we still use the original lemma (to use negative-entropy Ψ and avoid poly(A)), but instead reducing the magnitude of $\hat{\ell}$? **Yes!**

Magnitude-Reduced Estimator for Any R.V.

Lemma (Magnitude-Reduced Estimator; Informal)

*For any random variable Z **unbounded from below**, the estimator*

$$\hat{Z} \triangleq Z - (Z)_- + \mathbb{E}[(Z)_-] \text{ where } (Z)_- \triangleq \min\{Z, 0\} \text{ ensures}$$

1. ***(Expectation Invariance)*** $\mathbb{E}[\hat{Z}] = \mathbb{E}[Z]$;
2. ***(Same-Order 2nd Moment)*** $\mathbb{E}[\hat{Z}^2] \le 4\,\mathbb{E}[Z^2]$;
3. ***(Bounded from Below)*** $\hat{Z} \ge \mathbb{E}[(Z)_-]$.

Magnitude-Reduced Estimator for Any R.V.

Lemma (Magnitude-Reduced Estimator; Informal)

*For any random variable Z **unbounded from below**, the estimator*

$$\hat{Z} \triangleq Z - (Z)_- + \mathbb{E}[(Z)_-] \text{ where } (Z)_- \triangleq \min\{Z, 0\} \text{ ensures}$$

1. *(**Expectation Invariance**)* $\mathbb{E}[\hat{Z}] = \mathbb{E}[Z]$;
2. *(**Same-Order 2nd Moment**)* $\mathbb{E}[\hat{Z}^2] \le 4\,\mathbb{E}[Z^2]$;
3. *(**Bounded from Below**)* $\hat{Z} \ge \mathbb{E}[(Z)_-]$.

$\mathbb{E}[(Z)_-]$ is **much larger** than the smallest possible value of Z.

Magnitude-Reduced Estimator for Any R.V.

Lemma (Magnitude-Reduced Estimator; Informal)

*For any random variable Z **unbounded from below**, the estimator*

$$\hat{Z} \triangleq Z - (Z)_- + \mathbb{E}[(Z)_-] \text{ where } (Z)_- \triangleq \min\{Z, 0\} \text{ ensures}$$

1. *(**Expectation Invariance**)* $\mathbb{E}[\hat{Z}] = \mathbb{E}[Z]$;
2. *(**Same-Order 2nd Moment**)* $\mathbb{E}[\hat{Z}^2] \le 4\,\mathbb{E}[Z^2]$;
3. *(**Bounded from Below**)* $\hat{Z} \ge \mathbb{E}[(Z)_-]$.

$\mathbb{E}[(Z)_-]$ is **much larger** than the smallest possible value of Z.

Lemma

After applying the magnitude-reduced estimator to $\hat{\ell}$, the range of $\hat{\ell}$ moves from $[-\gamma^{-1}, \gamma^{-1}]$ to $[-\gamma^{-1/2}, \gamma^{-1}]$!

Magnitude-Reduced Estimator for Any R.V. (Cont'd)

Lemma

After applying the magnitude-reduced estimator to $\hat{\ell}$, the range of $\hat{\ell}$ moves from $[-\gamma^{-1}, \gamma^{-1}]$ to $[-\gamma^{-1/2}, \gamma^{-1}]$!

Lemma (Classical Regret Guarantee on FTRL; Informal)

*For "good enough" Ψ and **losses such that** $\hat{\ell}_{t,a} \ge -1/\eta$ **for all** $t = 1, 2, \ldots, T$ **and** $a \in \mathcal{A}$, Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \le \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \tag{1}$$

$\Longrightarrow$ we only need $\gamma^{-1/2} \le \eta^{-1}$ instead of $\gamma^{-1} \le \eta^{-1}$!

Magnitude-Reduced Estimator for Any R.V. (Cont'd)

Lemma

After applying the magnitude-reduced estimator to $\hat{\ell}$, the range of $\hat{\ell}$ moves from $[-\gamma^{-1}, \gamma^{-1}]$ to $[-\gamma^{-1/2}, \gamma^{-1}]$!

Lemma (Classical Regret Guarantee on FTRL; Informal)

For "good enough" Ψ and ***losses such that*** $\hat{\ell}_{t,a} \geq -1/\eta$ ***for all*** $t = 1, 2, \ldots, T$ ***and*** $a \in \mathcal{A}$, *Eq. (1) holds for any fixed $y \in \triangle(\mathcal{A})$.*

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \leq \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} x_{t,a} \hat{\ell}_{t,a}^2. \qquad (1)$$

$\Longrightarrow$ we only need $\gamma^{-1/2} \leq \eta^{-1}$ instead of $\gamma^{-1} \leq \eta^{-1}$!
Still setting $\beta = K^{-1/2}$, $\eta = K^{-1/2}$, $\gamma = K^{-1/2}$ **gives $\widetilde{\mathcal{O}}(K^{1/2})$ regret & removes poly(A)** (as we use negative-entropy Ψ)! ☺

Summary

This paper studies AMDPs with Linear Function Approximation:

- In Linear-Q AMDPs (with simulators), we achieve the **first-ever $\widetilde{\mathcal{O}}(\sqrt{K})$ regret** in two different ways:

Summary

This paper studies AMDPs with Linear Function Approximation:

- In Linear-Q AMDPs (with simulators), we achieve the **first-ever $\widetilde{\mathcal{O}}(\sqrt{K})$ regret** in two different ways:
  1. Via **new analysis for FTRL w/ Log-Barrier Regularizer**.
     Pro: Easy to use. No much modifications needed! ☺
     Con: Only log-barrier Ψ. Unavoidable poly(A) factors! ☹

Summary

This paper studies AMDPs with Linear Function Approximation:

- In Linear-Q AMDPs (with simulators), we achieve the **first-ever $\widetilde{\mathcal{O}}(\sqrt{K})$ regret** in two different ways:
  1. Via **new analysis for FTRL w/ Log-Barrier Regularizer**.
     Pro: Easy to use. No much modifications needed! ☺
     Con: Only log-barrier Ψ. Unavoidable poly(A) factors! ☹
  2. Via **applying magnitude-reduced estimators to $\hat{\ell}$**.
     Pro: Can use any regularizer, e.g., negative-entropy. ☺
     Con: $\mathbb{E}[(Z)_-]$ is only calculable with simulators! ☹

Summary

This paper studies AMDPs with Linear Function Approximation:

- In Linear-Q AMDPs (with simulators), we achieve the **first-ever $\widetilde{\mathcal{O}}(\sqrt{K})$ regret** in two different ways:
  1. Via **new analysis for FTRL w/ Log-Barrier Regularizer**.
     Pro: Easy to use. No much modifications needed! ☺
     Con: Only log-barrier Ψ. Unavoidable poly(A) factors! ☹
  2. Via **applying magnitude-reduced estimators to $\hat{\ell}$**.
     Pro: Can use any regularizer, e.g., negative-entropy. ☺
     Con: $\mathbb{E}[(Z)_-]$ is only calculable with simulators! ☹
- In Linear AMDPs, we get $\widetilde{\mathcal{O}}(K^{8/9})$ regret via a **new relative concentration bound** for stochastic matrices (in appendix).

Concluding Remarks

1. People now do better than our $\widetilde{\mathcal{O}}(K^{8/9})$ on Linear AMDPs:
   - Linear AMDP w/ Unknown Transition & Bandit Feedback **(our setup)**: $\widetilde{\mathcal{O}}(K^{6/7})$ [Sherman et al., 2023b] and $\widetilde{\mathcal{O}}(K^{4/5})$ [Kong et al., 2023] (requires the existence of an exploratory policy, but no polynomial dependency on λ_0 presents).

Concluding Remarks

1. People now do better than our $\widetilde{\mathcal{O}}(K^{8/9})$ on Linear AMDPs:
    - Linear AMDP w/ Unknown Transition & Bandit Feedback **(our setup)**: $\widetilde{\mathcal{O}}(K^{6/7})$ [Sherman et al., 2023b] and $\widetilde{\mathcal{O}}(K^{4/5})$ [Kong et al., 2023] (requires the existence of an exploratory policy, but no polynomial dependency on λ_0 presents).
    - Linear AMDP w/ Unknown Transition & Full Information **(weaker setup)**: $\widetilde{\mathcal{O}}(K^{1/2})$ [Sherman et al., 2023a].

Concluding Remarks

1. People now do better than our $\widetilde{\mathcal{O}}(K^{8/9})$ on Linear AMDPs:
   - Linear AMDP w/ Unknown Transition & Bandit Feedback **(our setup)**: $\widetilde{\mathcal{O}}(K^{6/7})$ [Sherman et al., 2023b] and $\widetilde{\mathcal{O}}(K^{4/5})$ [Kong et al., 2023] (requires the existence of an exploratory policy, but no polynomial dependency on λ_0 presents).
   - Linear AMDP w/ Unknown Transition & Full Information **(weaker setup)**: $\widetilde{\mathcal{O}}(K^{1/2})$ [Sherman et al., 2023a].
2. Our relative concentration result for stochastic matrices is further improved by [Liu et al., 2023] ($\widetilde{\mathcal{O}}(\gamma^{-2}) \Rightarrow \widetilde{\mathcal{O}}(\gamma^{-1})$).

Thank you for listening!

Questions are more than welcomed.

References I

He, J., Zhou, D., and Gu, Q. (2022).
Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps.
In *International Conference on Artificial Intelligence and Statistics*, pages 4259–4280. PMLR.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. (2020).
Learning adversarial markov decision processes with bandit feedback and unknown transition.
In *International Conference on Machine Learning*, pages 4860–4869. PMLR.

Kong, F., Zhang, X., Wang, B., and Li, S. (2023).
Improved regret bounds for linear adversarial mdps via linear optimization.
arXiv preprint arXiv:2302.06834.

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

References II

Luo, H., Wei, C.-Y., and Lee, C.-W. (2021a).
Policy optimization in adversarial mdps: Improved exploration via dilated bonuses.
Advances in Neural Information Processing Systems, 34:22931–22942.

Luo, H., Wei, C.-Y., and Lee, C.-W. (2021b).
Policy optimization in adversarial mdps: Improved exploration via dilated bonuses.
arXiv preprint arXiv:2107.08346.

Neu, G. and Olkhovskaya, J. (2020).
Efficient and robust algorithms for adversarial linear contextual bandits.
In *Conference on Learning Theory*, pages 3049–3068. PMLR.

Neu, G. and Olkhovskaya, J. (2021).
Online learning in mdps with linear function approximation and bandit feedback.
Advances in Neural Information Processing Systems, 34:10407–10417.

Rosenberg, A. and Mansour, Y. (2019).
Online convex optimization in adversarial markov decision processes.
In *International Conference on Machine Learning*, pages 5478–5486. PMLR.

References III

Sherman, U., Cohen, A., Koren, T., and Mansour, Y. (2023a).
Rate-optimal policy optimization for linear markov decision processes.
arXiv preprint arXiv:2308.14642.

Sherman, U., Koren, T., and Mansour, Y. (2023b).
Improved regret for efficient online reinforcement learning with linear function approximation.
In *International Conference on Machine Learning*. PMLR.

Zhao, C., Yang, R., Wang, B., and Li, S. (2022).
Learning adversarial linear mixture markov decision processes with bandit feedback and unknown transition.
In *The Eleventh International Conference on Learning Representations.*

Zimin, A. and Neu, G. (2013).
Online learning in episodic markovian decision processes by relative entropy policy search.
Advances in neural information processing systems, 26.

Appendix. Our Relative Concentration Bound

Lemma (New Covariance Concentration; Informal)

For a d-dimensional distribution $\mathcal{D}$ w/ covariance Σ, sampling $W = (\mathbf{4d}\log\frac{\mathbf{d}}{\delta})\gamma^{\mathbf{-2}}$ i.i.d. samples $\phi_1, \phi_2, \ldots, \phi_W$ from $\mathcal{D}$ ensures

$$\left(\hat{\Sigma}^\dagger\right)^{1/2} (\gamma I + \Sigma) \left(\hat{\Sigma}^\dagger\right)^{1/2} \in [(\mathbf{1} - \mathbf{2}\sqrt{\gamma})\mathbf{I}, (\mathbf{1} + \mathbf{2}\sqrt{\gamma})\mathbf{I}],$$

where $\hat{\Sigma}^\dagger = \left(\gamma I + \sum_{w=1}^{W} \phi_w \phi_w^T\right)^{-1}$.

Previous approach gives **additive bounds**, e.g., Matrix Geometric Resampling (MGR) by [Neu and Olkhovskaya, 2020] needs $\mathcal{O}(\epsilon^{\mathbf{-2}}\gamma^{\mathbf{-3}})$ samples for a $\hat{\Sigma}^\dagger$ s.t. $\left\|\hat{\Sigma}^\dagger - (\gamma I + \Sigma)^{-1}\right\|_2 \le \epsilon$.