清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

USC University of Southern California

Refined Regret for Adversarial MDPs with Linear Function Approximation

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

[1] IIIS, Tsinghua University
[2] University of Southern California
[3] IDSS, MIT
[4] Google Research

Presented by Yan Dai

IDSS | MIT

Google Research

Problem Setup & Related Work

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q_{k,h}^{\pi}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q_{k,h+1}^{\pi}(s',a')]$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q_{k,h}^{\pi}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q_{k,h+1}^{\pi}(s',a')]$
  - That is, $Q_{k,h}^{\pi}(s,a) = \langle \phi(s,a), \theta_{k,h}^{\pi} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta_{k,h}^{\pi}$ is *unknown*

Problem Setup & Related Work

- **Adversarial MDP:** MDP with <u>*time-varying*</u> losses $\ell_{k,h}(s,a)$ but <u>*stationary*</u> transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with <u>*linear*</u> Q-function $Q_{k,h}^{\pi}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q_{k,h+1}^{\pi}(s',a')]$
  - That is, $Q_{k,h}^{\pi}(s,a) = \langle \phi(s,a), \theta_{k,h}^{\pi} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is <u>*known and stationary*</u> but $\theta_{k,h}^{\pi}$ is <u>*unknown*</u>
- **Linear AMDP:** Linear-Q AMDP with <u>*linear*</u> transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^\pi_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_\pi[Q^\pi_{k,h+1}(s',a')]$
  - That is, $Q^\pi_{k,h}(s,a) = \langle \phi(s,a), \theta^\pi_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^\pi_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^K$ and a stationary comparator π^*

Table 1: Comparison with related works on **Linear-Q AMDPs (with a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	**Regret**
Luo et al. (2021)	None	$\tilde{O}(d^{2/3}H^2K^{2/3})$
This paper	**None**	$\tilde{O}(A^{1/2}d^{1/2}H^3K^{1/2})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^{\pi}_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q^{\pi}_{k,h+1}(s',a')]$
  - That is, $Q^{\pi}_{k,h}(s,a) = \langle \phi(s,a), \theta^{\pi}_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^{\pi}_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^{K}$ and a stationary comparator π^*

Table 1: Comparison with related works on **Linear-Q AMDPs (with a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	Regret
Luo et al. (2021)	None	$\tilde{O}(d^{2/3}H^2K^{2/3})$
	Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{1/2})$
This paper	None	$\tilde{O}(A^{1/2}d^{1/2}H^3K^{1/2})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^\pi_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_\pi[Q^\pi_{k,h+1}(s',a')]$
  - That is, $Q^\pi_{k,h}(s,a) = \langle \phi(s,a), \theta^\pi_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^\pi_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^K$ and a stationary comparator π^*

Table 1: Comparison with related works on **Linear-Q AMDPs (with a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	Regret
Luo et al. (2021)	None	$\tilde{O}(d^{2/3}H^2K^{2/3})$
	Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{1/2})$
This paper	**None**	$\tilde{O}(A^{1/2}d^{1/2}H^3K^{1/2})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^{\pi}_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q^{\pi}_{k,h+1}(s',a')]$
    - That is, $Q^{\pi}_{k,h}(s,a) = \langle \phi(s,a), \theta^{\pi}_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^{\pi}_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^K$ and a stationary comparator π^*

Table 1: Comparison with related works on **Linear-Q AMDPs (with a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	Regret
Luo et al. (2021)	None	$\tilde{O}(d^{2/3}H^2K^{2/3})$
	Exploratory Policy π_0	$\tilde{O}(\text{poly}(d,H)(K/\lambda_0)^{1/2})$
This paper	**None**	$\tilde{O}(A^{1/2}d^{1/2}H^3K^{1/2})$
This paper	**None**	$\tilde{O}(d^{1/2}H^3K^{1/2})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^{\pi}_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q^{\pi}_{k,h+1}(s',a')]$
  - That is, $Q^{\pi}_{k,h}(s,a) = \langle \phi(s,a), \theta^{\pi}_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^{\pi}_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^{K}$ and a stationary comparator π^*

Table 2: Comparison with related works on **Linear AMDPs (without a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	**Regret**
Neu & Olkhovskaya (2021)	**Known Transition**	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{1/2})$
	Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{6/7})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^\pi_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_\pi[Q^\pi_{k,h+1}(s',a')]$
  - That is, $Q^\pi_{k,h}(s,a) = \langle \phi(s,a), \theta^\pi_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^\pi_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^K$ and a stationary comparator π^*

Table 2: Comparison with related works on **Linear AMDPs (without a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	Regret
Neu & Olkhovskaya (2021)	Known Transition & Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{1/2})$
	Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{6/7})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^{\pi}_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q^{\pi}_{k,h+1}(s',a')]$
  - That is, $Q^{\pi}_{k,h}(s,a) = \langle \phi(s,a), \theta^{\pi}_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^{\pi}_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^K$ and a stationary comparator π^*

Table 2: Comparison with related works on **Linear AMDPs (without a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	Regret
Neu & Olkhovskaya (2021)	Known Transition & Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{1/2})$
Luo et al. (2021)	None	$\tilde{O}(d^2 H^4 K^{14/15})$
	Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{6/7})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^{\pi}_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q^{\pi}_{k,h+1}(s',a')]$
  - That is, $Q^{\pi}_{k,h}(s,a) = \langle \phi(s,a), \theta^{\pi}_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^{\pi}_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^K$ and a stationary comparator π^*

Table 2: Comparison with related works on **Linear AMDPs (without a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	Regret
Neu & Olkhovskaya (2021)	Known Transition & Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{1/2})$
Luo et al. (2021)	None	$\tilde{O}(d^2H^4K^{14/15})$
	Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{6/7})$

Problem Setup & Related Work

- **Adversarial MDP:** MDP with *time-varying* losses $\ell_{k,h}(s,a)$ but *stationary* transitions $\mathbb{P}_h(s'|s,a)$
- **Linear-Q AMDP:** AMDP with *linear* Q-function $Q^{\pi}_{k,h}(s,a) = \ell_{k,h}(s,a) + \mathbb{E}_{\pi}[Q^{\pi}_{k,h+1}(s',a')]$
  - That is, $Q^{\pi}_{k,h}(s,a) = \langle \phi(s,a), \theta^{\pi}_{k,h} \rangle$ where $\phi(s,a) \in \mathbb{R}^d$ is *known and stationary* but $\theta^{\pi}_{k,h}$ is *unknown*
- **Linear AMDP:** Linear-Q AMDP with *linear* transitions: $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \nu(s') \rangle$
- **Regret**: Expected difference between losses collected by $\{\pi_k\}_{k=1}^K$ and a stationary comparator π^*

Table 2: Comparison with related works on **Linear AMDPs (without a simulator)**; $\tilde{O}$ hides all logarithmic factors

	Assumption	Regret
Neu & Olkhovskaya (2021)	Known Transition & Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{1/2})$
Luo et al. (2021)	None	$\tilde{O}(d^2H^4K^{14/15})$
	Exploratory Policy π_0	$\tilde{O}(\mathrm{poly}(d,H)(K/\lambda_0)^{6/7})$
This paper	**None**	$\tilde{O}(\mathrm{poly}(d,A,H)K^{8/9})$

Technical Overview

Technical Overview

- **Refined Analysis of FTRL w/ Log-Barrier on arbitrary loss vectors $\{\ell_t \in \mathbb{R}^A\}_{t=1}^T$: (no longer require $\ell_{t,i} \ge -1/\eta$!)**

  Actions $\left\{x_t \in \Delta^{[A]}\right\}_{t=1}^T$ are defined as: $$x_t = \operatorname*{arg\,min}_{x \in \Delta^{[A]}} \left\{\eta \left\langle x, \sum_{t'<t} \ell_{t'} \right\rangle + \Psi(x)\right\},$$ where $\Psi(x) = \sum_{i=1}^A \ln \frac{1}{x_i}$.

  Then the following holds for any comparator $y \in \Delta^{[A]}$:

$$\sum_{t=1}^T \langle x_t - y, \ell_t \rangle \le \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^A x_{t,i} \ell_{t,i}^2 .$$

Technical Overview

- **Refined Analysis of FTRL w/ Log-Barrier on arbitrary loss vectors $\{\ell_t \in \mathbb{R}^A\}_{t=1}^T$: (no longer require $\ell_{t,i} \geq -1/\eta$!)**

  Actions $\{x_t \in \Delta^{[A]}\}_{t=1}^T$ are defined as: $$x_t = \operatorname*{arg\,min}_{x \in \Delta^{[A]}} \left\{\eta \left\langle x, \sum_{t'<t} \ell_{t'} \right\rangle + \Psi(x)\right\}, \quad \text{where } \Psi(x) = \sum_{i=1}^A \ln \frac{1}{x_i}.$$

  Then the following holds for any comparator $y \in \Delta^{[A]}$:
  $$\sum_{t=1}^T \langle x_t - y, \ell_t \rangle \leq \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^A x_{t,i} \ell_{t,i}^2 .$$

- **Magnitude Reduced Estimator:** For an arbitrary random variable Z that can be prohibitively negative, define
  $$\hat{Z} = Z - (Z)_- + \mathbb{E}[(Z)_-], \qquad \text{where } (Z)_- = \min(Z, 0).$$

  Then our Magnitude Reduced Estimator $\hat{Z}$ enjoys the following properties:
  - Preserve Expectation: $\mathbb{E}[\hat{Z}] = \mathbb{E}[Z] - \mathbb{E}[(Z)_-] + \mathbb{E}[(Z)_-] = \mathbb{E}[Z]$.
  - Similar Second Order Moment: $\mathbb{E}[\hat{Z}^2] \leq 2\mathbb{E}[Z^2] + 2(\mathbb{E}[(Z)_-])^2 = \mathcal{O}(\mathbb{E}[Z^2])$.
  - Bounded from Below: $\hat{Z} \geq \mathbb{E}[(Z)_-]$ as $Z - (Z)_- = 0$ when $Z < 0$ and $Z - (Z)_- = Z \geq 0$ when $Z \geq 0$.

Technical Overview

- **Refined Analysis of FTRL w/ Log-Barrier on arbitrary loss vectors $\{\ell_t \in \mathbb{R}^A\}_{t=1}^T$: (no longer require $\ell_{t,i} \ge -1/\eta$!)**

  Actions $\left\{x_t \in \Delta^{[A]}\right\}_{t=1}^T$ are defined as: $x_t = \operatorname*{arg\,min}_{x \in \Delta^{[A]}} \left\{\eta \left\langle x, \sum_{t'<t} \ell_{t'} \right\rangle + \Psi(x)\right\}$, where $\Psi(x) = \sum_{i=1}^A \ln \frac{1}{x_i}$.

  Then the following holds for any comparator $y \in \Delta^{[A]}$:

$$\sum_{t=1}^T \langle x_t - y, \ell_t \rangle \le \frac{\Psi(y) - \Psi(x_1)}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^A x_{t,i} \ell_{t,i}^2 .$$

- **Magnitude Reduced Estimator:** For an arbitrary random variable Z that can be prohibitively negative, define

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

  Then our Magnitude Reduced Estimator $\hat{Z}$ enjoys the following properties:

  - Preserve Expectation: $\mathbb{E}[\hat{Z}] = \mathbb{E}[Z] - \mathbb{E}[(Z)_-] + \mathbb{E}[(Z)_-] = \mathbb{E}[Z]$.
  - Similar Second Order Moment: $\mathbb{E}[\hat{Z}^2] \le 2\mathbb{E}[Z^2] + 2(\mathbb{E}[(Z)_-])^2 = \mathcal{O}(\mathbb{E}[Z^2])$.
  - Bounded from Below: $\hat{Z} \ge \mathbb{E}[(Z)_-]$ as $Z - (Z)_- = 0$ when $Z < 0$ and $Z - (Z)_- = Z \ge 0$ when $Z \ge 0$.

- **New Covariance Estimation Bound:** For a d-dim'l distribution w/ covariance Σ, samples $\{\phi_i\}_{i=1}^W$ ensures (w.p. $1-\delta$):

$$\left(\hat{\Sigma}^+\right)^{1/2} (\gamma I + \Sigma) \left(\hat{\Sigma}^+\right)^{1/2} \in [(1 - 2\sqrt{\gamma})I, (1 + 2\sqrt{\gamma})I], \qquad \text{where } \hat{\Sigma}^+ = \left(\gamma I + \sum_{i=1}^W \phi_i \phi_i^T\right)^{-1}, \qquad \text{given } W \ge \left(4 d \log \frac{d}{\delta}\right) \gamma^{-2}.$$

Institute for Interdisciplinary Information Sciences, Tsinghua University

Thank You for Listening!

Email: yan-dai20@mails.tsinghua.edu.cn

References

- Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. Advances in Neural Information Processing Systems, 34:22931–22942, 2021.
- Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. Advances in Neural Information Processing Systems, 34: 10407–10417, 2021.