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

USC University of Southern California

Follow-the-Perturbed-Leader (FTPL) for Adversarial Markov Decision Processes (AMDP) with Bandit Feedback

Yan Dai [1], Haipeng Luo [2], Liyu Chen [2]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University

[2] Computer Science Department, University of Southern California

Presented by Yan Dai

Our Contribution

1. Follow-the-Perturbed Leader (FTPL) is as good as other OMD-based algorithms

Our Contribution

1. Follow-the-Perturbed Leader (FTPL) is as good as other OMD-based algorithms
2. Show that FTPL can be easily generalized to various settings, giving:
   - A near-optimal algorithm for episodic AMDPs with delays, and
   - The first no-regret algorithm for weakly-communicating infinite-horizon AMDPs.

OMD (Online Mirror Descent) vs FTPL (Follow-the-Perturbed-Leader)

OMD (Online Mirror Descent) vs FTPL (Follow-the-Perturbed-Leader)

Online Mirror Descent

- Flexible in Algorithm Design
- Studied More in the Literature

Follow-the-Perturbed-Leader

- Easier to Implement
- More Computationally Efficient

OMD (Online Mirror Descent) vs FTPL (Follow-the-Perturbed-Leader)

Online Mirror Descent

- Flexible in Algorithm Design
- Studied More in the Literature

Follow-the-Perturbed-Leader

- Easier to Implement
- More Computationally Efficient

Table 1: Comparison between OMD- and FTPL-Based Algorithms for Episodic AMDPs [1]

OMD-Based		Transition	Feedback	FTPL-Based	
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{K})$	Known	Full-info	$\tilde{O}(H\sqrt{SAK})$	(Even-Dar et al., 2009)
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{SAK})$	Known	Bandit	$\tilde{O}(H^2\sqrt{AK}/\alpha)$	(Neu et al., 2010)
(Rosenberg & Mansour, 2019)	$\tilde{O}(H^2S\sqrt{AK})$	Unknown	Full-info	$\tilde{O}(H^{1.5}SA\sqrt{K})$	(Neu et al., 2012)
(Jin et al., 2020)	$\tilde{O}(H^2S\sqrt{AK})$	Unknown	Bandit	N/A	(no such algorithm)

[1] K is the number of episodes, H is the episode length, S is the number of states, and A is the number of actions. $\tilde{O}$ hides all logarithmic factors. α by Neu et al. (2012) is a parameter of a strong exploratory assumption.

OMD (Online Mirror Descent) vs FTPL (Follow-the-Perturbed-Leader)

Online Mirror Descent

- ■ Flexible in Algorithm Design
- ■ Studied More in the Literature
- □ Better Regret Guarantees

Follow-the-Perturbed-Leader

- ■ Easier to Implement
- ■ More Computationally Efficient
- □ Worse Regret Guarantee

Table 1: Comparison between OMD- and FTPL-Based Algorithms for Episodic AMDPs [1]

OMD-Based		Transition	Feedback	FTPL-Based	
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{K})$	Known	Full-info	$\tilde{O}(H\sqrt{SAK})$	(Even-Dar et al., 2009)
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{SAK})$	Known	Bandit	$\tilde{O}(H^2\sqrt{AK}/\alpha)$	(Neu et al., 2010)
(Rosenberg & Mansour, 2019)	$\tilde{O}(H^2 S\sqrt{AK})$	Unknown	Full-info	$\tilde{O}(H^{1.5}SA\sqrt{K})$	(Neu et al., 2012)
(Jin et al., 2020)	$\tilde{O}(H^2 S\sqrt{AK})$	Unknown	Bandit	N/A	(no such algorithm)

[1] K is the number of episodes, H is the episode length, S is the number of states, and A is the number of actions. $\tilde{O}$ hides all logarithmic factors. α by Neu et al. (2012) is a parameter of a strong exploratory assumption.

OMD (Online Mirror Descent) vs FTPL (Follow-the-Perturbed-Leader)

Online Mirror Descent

- ■ Flexible in Algorithm Design
- ■ Studied More in the Literature
- □ Better Regret Guarantees?

Follow-the-Perturbed-Leader

- ■ Easier to Implement
- ■ More Computationally Efficient
- □ Worse Regret Guarantee?

Table 1: Comparison between OMD- and FTPL-Based Algorithms for Episodic AMDPs [1]

OMD-Based		Transition	Feedback	FTPL-Based	
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{K})$	Known	Full-info	$\tilde{O}(H^2\sqrt{K})$	(Wang & Dong, 2020)
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{SAK})$	Known	Bandit	$\tilde{O}(H^2\sqrt{AK}/\alpha)$	(Neu et al., 2010)
(Rosenberg & Mansour, 2019)	$\tilde{O}(H^2S\sqrt{AK})$	Unknown	Full-info	$\tilde{O}(H^2S\sqrt{AK})$	(Wang & Dong, 2020)
(Jin et al., 2020)	$\tilde{O}(H^2S\sqrt{AK})$	Unknown	Bandit	N/A	(no such algorithm)

[1] K is the number of episodes, H is the episode length, S is the number of states, and A is the number of actions. $\tilde{O}$ hides all logarithmic factors. α by Neu et al. (2012) is a parameter of a strong exploratory assumption.

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

USC University of Southern California

OMD (Online Mirror Descent) vs FTPL (Follow-the-Perturbed-Leader)

Online Mirror Descent

- ■ Flexible in Algorithm Design
- ■ Studied More in the Literature
- □ Better Regret Guarantees ×

Follow-the-Perturbed-Leader

- ■ Easier to Implement
- ■ More Computationally Efficient
- □ Worse Regret Guarantee ×

Table 1: Comparison between OMD- and FTPL-Based Algorithms for Episodic AMDPs [1]

OMD-Based		Transition	Feedback	FTPL-Based	
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{K})$	Known	Full-info	$\tilde{O}(H^2\sqrt{K})$	(Wang & Dong, 2020)
(Zimin & Neu, 2013)	$\tilde{O}(H\sqrt{SAK})$	Known	Bandit	$\tilde{O}(H^{1.5}\sqrt{SAK})$	(**This paper**)
(Rosenberg & Mansour, 2019)	$\tilde{O}(H^2S\sqrt{AK})$	Unknown	Full-info	$\tilde{O}(H^2S\sqrt{AK})$	(Wang & Dong, 2020)
(Jin et al., 2020)	$\tilde{O}(H^2S\sqrt{AK})$	Unknown	Bandit	$\tilde{O}(H^2S\sqrt{AK})$	(**This paper**)

[1] K is the number of episodes, H is the episode length, S is the number of states, and A is the number of actions. $\tilde{O}$ hides all logarithmic factors. α by Neu et al. (2012) is a parameter of a strong exploratory assumption.

Technical Stuff

$$\mathcal{R}_K = \sum_{h=1}^{H} \mathbb{E}\left[\left\langle \mu_{\pi_k}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi^*}, \widehat{\ell_k} \right\rangle\right],$$

where $\mu_\pi^h(s,a) = \Pr\{s^h = s, a^h = a \mid \pi\}$ is the occupancy measure.

Technical Stuff

$$\mathcal{R}_K = \sum_{h=1}^{H} \mathbb{E}\left[\left\langle \mu_{\pi_k}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi^*}, \widehat{\ell_k} \right\rangle\right],$$

where $\mu_\pi^h(s,a) = \Pr\{s^h = s, a^h = a \mid \pi\}$ is the occupancy measure.

$$\mathcal{R}_K = \sum_{k=1}^{K} \mathbb{E}\left[\left\langle \mu_{\pi_k}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi_{k+1}}, \widehat{\ell_k} \right\rangle\right] + \sum_{k=1}^{K} \mathbb{E}\left[\left\langle \mu_{\pi_{k+1}}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi^*}^h, \widehat{\ell_k} \right\rangle\right].$$

Stability Term
(every single step controlled by (☆)?)

Error Term
(controlled by “be-the-leader” lemma)

Technical Stuff

$$\mathcal{R}_K = \sum_{h=1}^{H} \mathbb{E}\left[\left\langle \mu_{\pi_k}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi^*}, \widehat{\ell_k} \right\rangle\right],$$

where $\mu_\pi^h(s,a) = \Pr\{s^h = s, a^h = a \mid \pi\}$ is the occupancy measure.

$$\mathcal{R}_K = \underbrace{\sum_{k=1}^{K} \mathbb{E}\left[\left\langle \mu_{\pi_k}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi_{k+1}}, \widehat{\ell_k} \right\rangle\right]}_{\text{Stability Term}} + \underbrace{\sum_{k=1}^{K} \mathbb{E}\left[\left\langle \mu_{\pi_{k+1}}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi^*}^h, \widehat{\ell_k} \right\rangle\right]}_{\text{Error Term}}.$$

Stability Term
(every single step controlled by (☆)?)

Error Term
(controlled by "be-the-leader" lemma)

$$\mathbb{E}\left[\sum_{\pi\in\Pi} (p_k(\pi) - p_{k+1}(\pi)) \left\langle \mu_\pi, \widehat{\ell_k} \right\rangle\right] \le \eta \mathbb{E}\left[\sum_{\pi\in\Pi} p_k(\pi) \left\langle \mu_\pi, \widehat{\ell_k} \right\rangle^2\right] \quad (☆)$$

(Syrgkanis et al., 2016)
Lemma 10 [2]

Invalid when $\mu_\pi^h \notin \{0,1\}^d$ (non-binary feature)!

[2] $p_k(\pi)$ denotes the probability of playing π in the k-th episode.

Technical Stuff

$$\mathcal{R}_K = \sum_{h=1}^{H} \mathbb{E}\left[\left\langle \mu_{\pi_k}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi^*}, \widehat{\ell_k} \right\rangle\right],$$

where $\mu_\pi^h(s,a) = \Pr\{s^h = s, a^h = a \mid \pi\}$ is the occupancy measure.

$$\mathcal{R}_K = \underbrace{\sum_{k=1}^{K} \mathbb{E}\left[\left\langle \mu_{\pi_k}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi_{k+1}}, \widehat{\ell_k} \right\rangle\right]}_{\text{Stability Term}} + \underbrace{\sum_{k=1}^{K} \mathbb{E}\left[\left\langle \mu_{\pi_{k+1}}, \widehat{\ell_k} \right\rangle - \left\langle \mu_{\pi^*}^h, \widehat{\ell_k} \right\rangle\right]}_{\text{Error Term}}.$$

Stability Term
(every single step controlled by (■)!)

Error Term
(controlled by "be-the-leader" lemma)

$$\mathbb{E}\left[\sum_{\pi \in \Pi} (p_k(\pi) - p_{k+1}(\pi)) \left\langle \mu_\pi, \widehat{\ell_k} \right\rangle\right] \le \eta \mathbb{E}\left[\left(\sum_{h=1}^{H} \|\widehat{\ell_k^h}\|_1\right) \sum_{\pi \in \Pi} p_k(\pi) \left\langle \mu_\pi, \widehat{\ell_k} \right\rangle^1\right]$$

(■)
(This paper)
Lemma 3 [2]

Only loosen by H times.

[2] $p_k(\pi)$ denotes the probability of playing π in the k-th episode.

Beyond Episodic AMDPs

Beyond Episodic AMDPs

Feedback Delays? No Problem!

Table 2: Application to Episodic AMDP with Feedback Delays [3]

Algorithm	Regret
Delayed Hedge	$\tilde{O}(H^2 S\sqrt{AK} + H^{1.5}\sqrt{S\mathfrak{D}})$
Delayed UOB-FTRL	$\tilde{O}(H^2 S\sqrt{AK} + H^{1.5} SA\sqrt{\mathfrak{D}})$
Delayed UOB-REPS	$\tilde{O}(H^2 S\sqrt{AK} + H^{5/4}(SA)^{1/4}\sqrt{\mathfrak{D}})$
This paper	$\tilde{O}(H^2 S\sqrt{AK} + H^{1.5} SA\sqrt{\mathfrak{D}})$

[3] $\mathfrak{D}$ is the total feedback delay. The first three OMD-based algorithms are all designed by Jin et al. (2022). Our algorithm is based on the second one.

Whether we can design FTPL-based algorithms using the "delay-adapted" loss estimator introduced by the third algorithm is left for future research.

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

USC University of Southern California

Beyond Episodic AMDPs

Feedback Delays? No Problem!

Table 2: Application to Episodic AMDP with Feedback Delays [3]

Algorithm	Regret
Delayed Hedge	$\tilde{O}(H^2 S\sqrt{AK} + H^{1.5}\sqrt{S\mathfrak{D}})$
Delayed UOB-FTRL	$\tilde{O}(H^2 S\sqrt{AK} + H^{1.5} SA\sqrt{\mathfrak{D}})$
Delayed UOB-REPS	$\tilde{O}(H^2 S\sqrt{AK} + H^{5/4}(SA)^{1/4}\sqrt{\mathfrak{D}})$
This paper	$\tilde{O}(H^2 S\sqrt{AK} + H^{1.5} SA\sqrt{\mathfrak{D}})$

[3] $\mathfrak{D}$ is the total feedback delay. The first three OMD-based algorithms are all designed by Jin et al. (2022). Our algorithm is based on the second one.

Whether we can design FTPL-based algorithms using the "delay-adapted" loss estimator introduced by the third algorithm is left for future research.

Infinite Horizon? Also Okay!

Table 3: Application to Infinite-Horizon AMDPs [4]

Algorithm	Regret
Neu et al. (2014)	$\tilde{O}(\tau^{1.5}\sqrt{AT})$ (Ergodic)
Dekel & Hazan (2013)	$\tilde{O}(S^3 AT^{2/3})$ (Deterministic)
This paper	$\tilde{O}(A^{1/2}(SD)^{2/3}T^{5/6})$ (Commu)
Dekel et al. (2014)	$\Omega(S^{1/3}T^{2/3})$ (Commu)

[4] For infinite-horizon AMDPs, assumptions about transitions are needed.

- Ergodic: the mixing time τ exists (*strong* assumption).
- Deterministic: all transitions are non-random (*strong* assumption).
- Communicating: the diameter D exists (the *weakest* assumption).

Hence, our paper considers the weakest communicating assumption and is the first to achieve a "no-regret" guarantee under bandit feedback.

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

USC University of Southern California

Thank You for Listening!

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

References

(I) OMD-Based Algorithms for Episodic AMDPs

- Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.
- Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486. PMLR, 2019.
- Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2020.

(II) FTPL-Based Algorithms for Episodic AMDPs

- Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.
- Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *COLT 2010 - The 23rd Conference on Learning Theory*, pages 231–243. Omnipress, 2010.
- Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813. PMLR, 2012.
- Yuanhao Wang and Kefan Dong. Refined analysis of fpl for adversarial markov decision processes. arXiv:2008.09251, 2020.
- Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert Schapire. Efficient algorithms for adversarial contextual learning. In International Conference on Machine Learning, pages 2159–2168. PMLR, 2016.

(III) Episodic AMDPs with Feedback Delays

- Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial mdp with delayed bandit feedback. arXiv:2201.13172, 2022.

(IV) Infinite-Horizon AMDPs

- Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. IEEE Transactions on Automatic Control, 59:676–691, 2014.
- Ofer Dekel and Elad Hazan. Better rates for any adversarial deterministic mdp. In International Conference on Machine Learning, pages 675–683. PMLR, 2013.
- Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In Proceedings of the forty-sixth annual ACM symposium on Theory of computing, pages 459–467, 2014.