Follow-the-Perturbed-Leader for Adversarial Markov Decision Processes with Bandit Feedback

Yan Dai (Institute for Interdisciplinary Information Sciences, Tsinghua University),
Haipeng Luo (Computer Science Department, University of Southern California),
Liyu Chen (Computer Science Department, University of Southern California)

NEURAL INFORMATION PROCESSING SYSTEMS

Debate between FTPL and OMD

Which works better for adversarial Markov Decision Processes (AMDPs)?

Follow-the-Perturbed-Leader (FTPL) and Online Mirror Descent (OMD) are two popular frameworks for AMDP algorithms. It was commonly believed that **FTPL-based algorithms have worse regret**, despite being easier to implement and computationally efficient (see Table 1 for a comparison).

Table 1: Comparison between ***previous*** 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. Better algorithms (ignoring poly(H) and logarithmic factors) marked in red.

Table 2: Comparison between OMD and FTPL Frameworks

Online Mirror Descent (OMD)	Follow-the-Perturbed-Leader (FTPL)
Flexible in algorithm design	Easier to implement
Studied more in the literature	More computationally efficient
Common Belief: FTPL-based algorithms have worse regret.	

Our Contribution

FTPL is as good as OMD in episodic AMDPs!

Inspired by the recent work (Wang & Dong, 2020), which showed that FTPL-based algorithms can get near-optimal regret guarantees for AMDPs with full information feedback, we question the common belief by asking:

Does FTPL really suffer from worse regret guarantees, compared with OMD?

Our paper **refutes** the common belief by designing FTPL-based algorithms for AMDPs with bandit feedback, which achieve the same regret guarantees as OMD-based ones (up to poly(H) and logarithmic factors, see Table 3).

Table 3: Comparison between ***current*** 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. Better algorithms (ignoring poly(H) and logarithmic factors) marked in red.

Takeaway: Our results show that FTPL, when applied to episodic AMDPs, is

- **as good as** OMD in terms of regret, while
- **much better than** OMD in terms of implementation difficulty and computational efficiency.

Our Contribution (Continued)

Going beyond episodic AMDPs by utilizing FTPL.

After showing that FTPL can be applied to the standard episodic AMDP setting, we also extend FTPL to more general settings.

The first setting is AMDPs with feedback delays, where the feedback of the k-th episode will only be delivered after d_k episodes and$\{d_k\}$ is arbitrary.

Table 4: Application to Episodic AMDP with Feedback Delays

Algorithm	Regret
Delayed Hedge (Jin et al., 2022)	$\tilde{O}(H^2S\sqrt{AK}+H^{1.5}\sqrt{S\mathfrak{D}})$
Delayed UOB-FTRL (Jin et al., 2022)	$\tilde{O}(H^2S\sqrt{AK}+H^{1.5}SA\sqrt{\mathfrak{D}})$
Delayed UOB-REPS (Jin et al., 2022)	$\tilde{O}(H^2S\sqrt{AK}+H^{5/4}(SA)^{1/4}\sqrt{\mathfrak{D}})$
This paper	$\tilde{O}(H^2S\sqrt{AK}+H^{1.5}SA\sqrt{\mathfrak{D}})$

In AMDPs with feedback delays, our FTPL-based algorithm is **as good as** its OMD-based counterpart, namely Delayed UOB-FTRL (Jin et al., 2022).

In infinite-horizon AMDPs, FTPL is even better. It gives **the first** no-regret algorithm for weakly communicating AMDPs with bandit feedback.

Table 5: Application to Infinite-Horizon AMDPs

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

As weakly-communication is known to be the weakest assumption under which an MDP can be learned, our assumption is the weakest.

Conclusion & Future Directions

FTPL allows better regret bounds & more flexibility than people imagined!

By FTPL, we get various algorithms for AMDPs with bandit feedback that

- **achieve near-optimal regret** (compared with OMD-based ones), and
- **are easy to implement** (only an offline planning problem is solved)

In some cases, our algorithms are even the first to be "no-regret".

Open problems:

1. Whether we can further improve our single-step stability lemma (see "Technical Details"). This is fruitful as it can give **the first** algorithm for AMDPs with bandit feedback, whose delay-related term is $\tilde{O}\left(H^{1.5}\sqrt{\mathfrak{D}}\right)$.
2. Whether we can adapt the "delay-adapted" loss estimator by Jin et al. (2022) for a better regret guarantee in AMDPs with feedback delays.
3. Whether we can improve our infinite-horizon AMDP algorithms.
4. Whether we can tackle adversarial transitions (instead of only losses).

Technical Details

A new "single-step" stability lemma.

A typical regret decomposition for algorithms in episodic AMDPs:

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

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

For FTPL-based algorithms, we usually further write the regret as:

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

where the first term is called the **stability term**, and the second, **error term**.

Error is usually controlled by the **"be-the-leader" lemma** (which is of order η^{-1}, where η is the parameter of perturbations; c.f., learning rates in OMD). Stability is controlled step-by-step, using a **"single-step" stability lemma**:

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

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

However, the lemma above (Syrgkanis et al., 2016, Lemma 10) only works when $\boldsymbol{\mu_\pi^h}$ **is binary** – which infers the transitions to be deterministic!

Main technical innovation: We show a new "single-step" stability lemma

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

where we bear an extra term related to the sum of ℓ_1-norms of all estimated losses (in violet). Roughly speaking, this term will cause the stability to be H times larger than before – however, it is now capable of **non-binary $\boldsymbol{\mu_\pi^h}$'s**!

An important future direction is to further enhance this "single-step" stability lemma. For example, can we modify the original lemma by Syrgkanis et al. (2016), such that the new version holds for non-binary features while only a constant overhead (independent of H and the losses) appears?

References

We omit the references only appearing in the tables. Please check our paper and/or slides for more details.

Yuanhao Wang and Kefan Dong (2020). "Refined Analysis of FPL for Adversarial Markov Decision Processes." In: arXiv preprint, arXiv:2008.09251.

Tiancheng Jin et al. (2022). "Near-optimal Regret for Adversarial MDP with Delayed Bandit Feedback.". In: arXiv preprint, arXiv:2201.13172.

Vasilis Syrgkanis et al. (2016). "Efficient Algorithms for Adversarial Contextual Learning.". In: International Conference on Machine Learning, PMLR, pp. 2159–2168.

Marge Simpson et al. (2013). "Lorem Ipsum." In: Advances in Neural Information Processing Systems 26. Ed. by Christopher J. C. Burges et al., pp. 27–29.