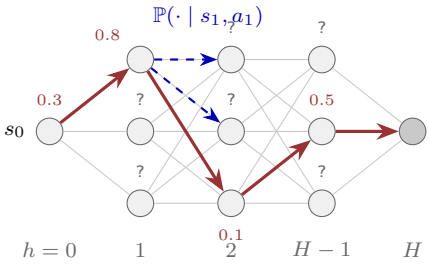
Introducing Feedback Delays to Non-Delayed Adversarial MDPs

Optimal $\sqrt{S}$ and $\sqrt{K}$ Dependencies in Tabular and Linear AMDPs

Aarush Kulkarni, Khang Nguyen, William Chang, **Yan Dai**

Episodic Adversarial MDPs



- K episodes each with length H. States $\mathcal{S} = \{s_0\} \cup \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_{H-1} \cup \{s_H\}$, initial state s_0, terminal state s_H. Actions $\mathcal{A}$
- At state $s_h \in \mathcal{S}_h$: pick action $a_h \in \mathcal{A}$, incur and observe **loss** $\ell_k(s_h, a_h) \in [0, 1]$, transit to $s_{h+1} \sim \mathbb{P}(\cdot \mid s_h, a_h)$ $(\in \mathcal{S}_{h+1})$
- Transition $\mathbb{P}$: fixed but **unknown**
- Losses $\ell_1, \ldots, \ell_K$: obliviously **adversarial**
- **Bandit feedback**: only see incurred losses
- **Policy** $\pi \colon \mathcal{S} \to \triangle(\mathcal{A})$ (can depend on k)

Goal: Compete with the Best Fixed Policy

- V_k^π: expected total loss of policy π in episode k (under the true transition $\mathbb{P}$)

Regret: total loss compared with the best fixed policy in hindsight

$$\mathcal{R}_K := \sum_{k=1}^{K} V_k^{\pi_k} - \min_{\pi} \sum_{k=1}^{K} V_k^{\pi}$$

$S = H = 1$: the A-armed bandit, $\Theta(\sqrt{AK})$. In general: $\Omega(H\sqrt{SAK})$ [JABJ18]

- $Q_k^\pi(s,a)$: expected loss of π from (s,a) to the end of episode k
- $q^\pi(s,a)$: probability that π visits (s,a) (occupancy measure)

Linear AMDPs: infinitely many states, known features $\phi(s,a) \in \mathbb{R}^d$

$\mathbb{P}(s' \mid s,a) = \langle \phi(s,a), \psi(s') \rangle$, $\ell_k(s,a) = \langle \phi(s,a), \zeta_{k,h} \rangle$. Known: $\Omega(dH^{3/2}\sqrt{K})$ [ZGS21]

Known Results: Stuck at $S\sqrt{AK}$ (tabular) and $K^{3/4}$ (linear)

Tabular AMDPs

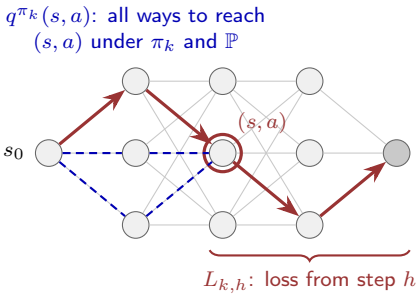
$H^{3/2}SA^{1/4}K^{3/4}$	[RM19]
$HS\sqrt{AK}$	[Jin+20] open problem
$H^2S\sqrt{AK}$	[LWL21]
$HS\sqrt{AK}$	[DLC22]
$H^2S\sqrt{AK}$	[DWZ23]
$HS\sqrt{AK}$	[LT26] concurrent
$H^{5/2}\sqrt{SAK}$	**ours**
lower bound: $H\sqrt{SAK}$	[JABJ18]

Linear AMDPs

$K^{14/15}$	[LWL21]
$K^{8/9}$	[DLWZ23]
$K^{6/7}$	[SKM23]
$K^{4/5}$	[KZWL24]
$K^{3/4}$	[LWZ24]
$\sqrt{K}$ (inefficient)	[LWZ24]
$\sqrt{K}$, efficient	[LWZ26] concurrent
$\sqrt{K}$, **efficient**	**ours**
lower bound: $\sqrt{K}$	[ZGS21]

Why were all of them stuck?

Evaluate the Current Policy π_k



- Evaluate the current π_k: estimate $Q_k^{\pi_k}(s,a)$ at **every** (s,a), from a **single** trajectory
- If the trajectory visits (s,a) at step h: we see $L_{k,h}$, the loss from step h to the end
- **Importance weighting** is unbiased:

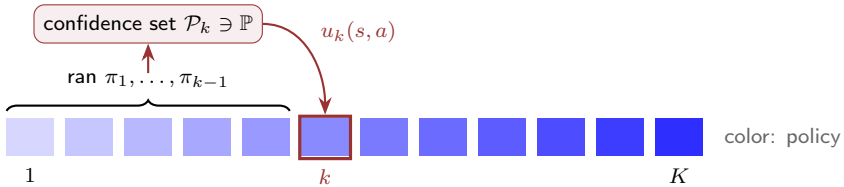
$$\widehat{Q}_k(s,a) = \frac{\mathbb{1}[\text{visit } (s,a)] \cdot L_{k,h}}{q^{\pi_k}(s,a)}$$

(i.e., $\mathbb{E}_{k-1}[\widehat{Q}_k(s,a)] = Q_k^{\pi_k}(s,a),\ \forall (s,a)$)

- But $q^{\pi_k}(s,a)$ depends on the **unknown** $\mathbb{P}$!

How to estimate $q^{\pi_k}(s,a)$ with a changing policy π_k and unknown transition $\mathbb{P}$?

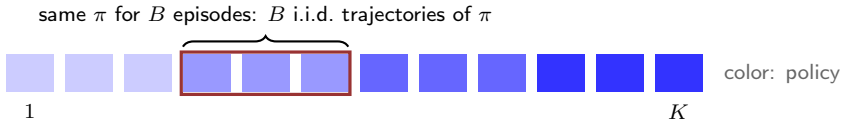
Idea 1: Relying on the Past Brings Linear-in-S Dependency



- Past episodes ran **other** policies, which may rarely visit where π_k visits
- Idea: estimate $\mathbb{P}$ as a confidence set $\mathcal{P}_k$, and take the most favorable $P' \in \mathcal{P}_k$
  (optimism: $\bar{q}_k(s,a) := \max_{P' \in \mathcal{P}_k} q^{P',\pi_k}(s,a) \geq q^{\pi_k}(s,a)$ [Jin+20])
- Confidence radius at (s,a) is $\approx \sqrt{\mathbb{P}(s' \mid s,a)/N(s,a)}$ after $N(s,a)$ visits
- $\Longrightarrow$ estimation error $\sum_k \sum_{s,a} \left(\bar{q}_k - q^{\pi_k}\right)(s,a) \approx HS\sqrt{AK}$

The width of $\mathcal{P}_k$ brings a linear-in-S dependency!

Idea 2: Freezing the Future Brings $K^{3/4}$ Dependency

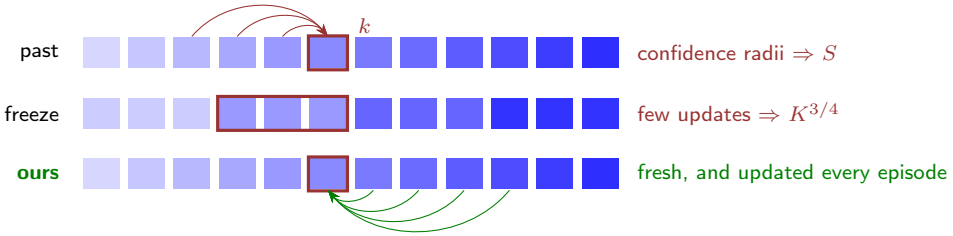


- Play the same π for an epoch of B episodes
- Estimate q^π from its own B trajectories: no mismatch, no $\mathcal{P}_k$ [DLWZ23; LWZ24]
- But only update policies for K/B times $\Longrightarrow$ regret $\gtrsim \sqrt{BK}$
- Best efficient algorithm [LWZ24]: $B \approx \sqrt{K} \Longrightarrow \sqrt{BK} = K^{3/4}$

Slow updates brings the sub-optimal $K^{3/4}$ term!

Evaluate π_k without relying on the past or freezing the future?

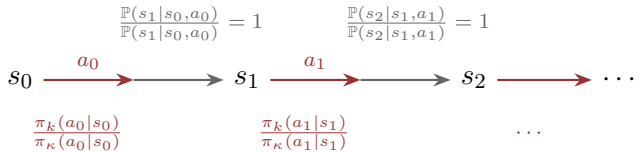
Evaluate π_k while Updating



- Estimate π_k's Q with **later** episodes while still updating $\pi_{k+1}, \pi_{k+2}, \ldots$
- Problem 1: episode $\kappa > k$ runs π_κ, not π_k? π_κ is **fresh** $\longrightarrow$ rejection sampling
  (cannot do the same in Idea 1: π_k was computed **from** historical trajectories)
- Problem 2: π_k's estimate is ready only after $\pi_{k+1}, \pi_{k+2}, \ldots \longrightarrow$ treat it as **delayed feedback**, introduced voluntarily and on purpose

Rejection Sampling

- Episode $\kappa > k$ runs π_κ. Can its trajectory serve as a trajectory of π_k?



- Likelihood ratio on actions: ratio of the two policies $\Longrightarrow$ **known!**
- Likelihood ratio on states: the same $\mathbb{P}$ under both $\Longrightarrow$ **cancels!**
- $\Longrightarrow$ Accept episode $\kappa > k$ w.p. $\alpha_{k,\kappa} := \frac{1}{2}\prod_h \frac{\pi_k(a_h|s_h)}{\pi_\kappa(a_h|s_h)}$ (if $\alpha_{k,\kappa} \le 1$)
- Given $\mathcal{F}_{\kappa-1}$: accepted w.p. $\frac{1}{2}$, and an accepted trajectory is a fresh trajectory of π_k
- $\Longrightarrow \Pr\{\pi_\kappa \text{ accepted and visits } (s,a)\} = q^{\pi_k}(s,a)/2$, **regardless of history**

We want $\alpha_{k,\kappa} \le 1 \Longrightarrow$ policies must change slowly.

Policy Optimization: A Bandit Algorithm on Each State

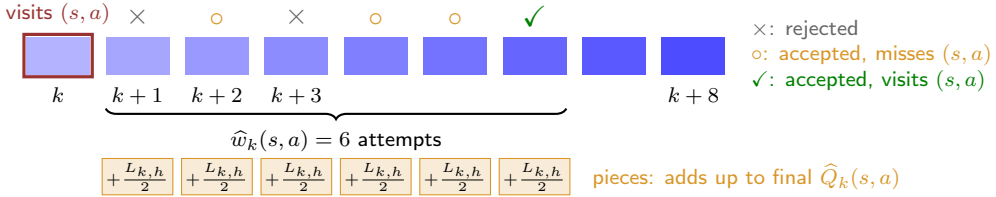
- Performance difference lemma [KL02] ($q^\star$: occupancy of the best policy $\pi^\star$)

$$\mathcal{R}_K = \sum_s q^\star(s) \underbrace{\sum_{k=1}^{K} \sum_a \big(\pi_k(a \mid s) - \pi^\star(a \mid s)\big) Q_k^{\pi_k}(s,a)}_{\text{a bandit at } s\text{: arm } a \text{ has loss } Q_k^{\pi_k}(s,a) \text{ in episode } k}$$

- Run Exponential Weights (EW) [FS97] at s: $\pi_k(a \mid s) \propto \exp\big(-\eta \sum_{\kappa<k} \widehat{Q}_\kappa(s,a)\big)$
  (η: learning rate; $\widehat{Q}_\kappa$: estimation of the Q-function $Q_\kappa^{\pi_\kappa}$)
- EW's regret at s: $\frac{\log A}{\eta} + \eta \sum_k \sum_a \pi_k(a \mid s)\, \mathbb{E}\big[\widehat{Q}_k(s,a)^2\big]$ **but weighted by $q^\star$**
- Bonuses [LWL21]: Feed $\widehat{Q}_k - B_k$ into EW for some B_k $\Longrightarrow$ **now weighted by q^{π_k}**

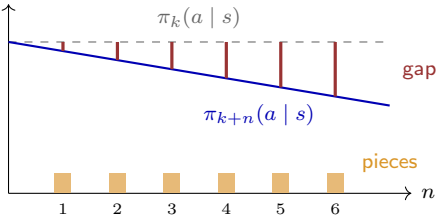
Want: still do importance weighting by $1/q^{\pi_k}(s,a)$, s.t. $\mathbb{E}\big[\widehat{Q}_k(s,a)^2\big] \lesssim \frac{H^2}{q^{\pi_k}(s,a)}$

Estimate $1/q^{\pi_k}$ by Geometric Resampling with Rejection (GRR)



- Episode k visits (s,a) at step $h \implies$ start an estimate
- Each later episode $\kappa > k$ is an **attempt**, accepted w.p. $\alpha_{k,\kappa}$
- Success: accepted **and** visits (s,a) (happens w.p. $q^{\pi_k}(s,a)/2$ every time)
- $\widehat{w}_k(s,a)$: #attempts until success (Geometric Resampling [NB16] with Rejection)
- $\widehat{Q}_k(s,a) := \mathbb{1}[\text{visit}] \cdot \frac{L_{k,h}}{2} \cdot \widehat{w}_k(s,a)$: unbiased and $\mathbb{E}\big[\widehat{Q}_k(s,a)^2\big] \le \frac{2H^2}{q^{\pi_k}(s,a)}$
- $\widehat{Q}_k \approx H/q^{\pi_k}$ is big (changes π_k a lot) feed it in small **pieces**, $L_{k,h}/2$ per attempt
- Starts GRR w.p. q^{π_k}, lasts $\approx 2/q^{\pi_k} \implies \widetilde{\mathcal{O}}(1)$ unfinished estimates per (s,a)

Bounding the Overhead due to Introduced Feedback Delays



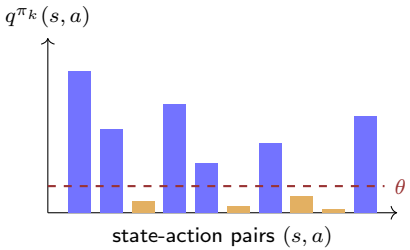
- The n-th piece of $\widehat{Q}_k$ reaches EW in episode $k+n$: **delayed** feedback [Jin+22; LRS23]
- Delay cost: the pieces are weighted by π_{k+n} instead of π_k; the difference:

$$\sum_{n=1}^{\widehat{w}_k(s,a)} \frac{L_{k,h}}{2} \big(\pi_k(a \mid s) - \pi_{k+n}(a \mid s)\big)$$

- $\log \pi(a \mid s)$ moves by $\widetilde{\mathcal{O}}(\eta H^2)$ per episode $\implies$ gap $\lesssim \eta H^2 n\, \pi_k(a \mid s)$
- Visits w.p. q^{π_k}, then $\mathbb{E}[\widehat{w}^2] \lesssim \frac{1}{(q^{\pi_k})^2} \implies$ delay cost $= \widetilde{\mathcal{O}}(\eta H^3) \frac{\pi_k(a|s)}{q^{\pi_k}(s,a)}$

Same as the EW regret (only H times larger)!

Rarely Visited Pairs? Thresholding on Occupancy Measures



- GRR needs $\alpha_{k,\kappa} \le 1$ for $\approx \frac{2}{q^{\pi_k}(s,a)}$ episodes, but policies stay that close for $\approx \frac{1}{\eta H^3}$ rounds
- **Thresholding:** If $q^{\pi_k}(s,a) \ge \theta \approx \eta H^3$, estimate. Otherwise: rarely visited $\implies$ estimation error $\le H$ is negligible (q^{π_k} is unknown: use confidence sets and pessimism [LWL21])

$$b_k(s) = \sum_a \pi_k(a \mid s) \left(\frac{\eta H^3}{q^{\pi_k}(s,a)} \mathbb{1}[q^{\pi_k}(s,a) \ge \theta] + H \cdot \mathbb{1}[q^{\pi_k}(s,a) < \theta] \right)$$

- Overhead is only $\sum_s q^{\pi_k}(s) b_k(s) \le \eta H^3 SA + H\theta SA \approx \eta H^4 SA$ per episode

Tabular AMDPs: $\mathcal{R}_K = \widetilde{\mathcal{O}}\left(H^{5/2}\sqrt{SAK} + H^4 S^2 A\right)$ **w.h.p. (optimal in** S, A, K**)**

Linear AMDPs: Matrix GRR

Linear AMDPs: infinitely many states, known features $\phi(s,a) \in \mathbb{R}^d$

$\mathbb{P}(s' \mid s,a) = \langle \phi(s,a), \psi(s') \rangle$, $\ell_k(s,a) = \langle \phi(s,a), \zeta_{k,h} \rangle$ for unknown ψ and ζ.

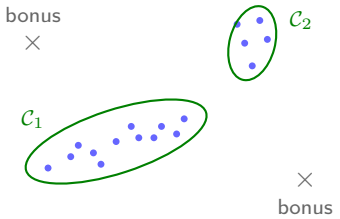
- $Q_k^\pi(s,a)$ is linear in $\phi(s,a)$. Covariance of π at layer h: $\Sigma_h^\pi := \mathbb{E}_\pi\left[\phi_h \phi_h^\top\right]$
- Linear importance weighting [NO21]: $\widehat{Q}_k(s,a) = \phi(s,a)^\top (\Sigma_h^{\pi_k})^{-1} \phi_{k,h} L_{k,h}$

	key fact	term from $\kappa = k+n$
GRR	$\frac{1}{q} = \frac{1}{2}\sum_{n=1}^{\infty}\left(1-\frac{q}{2}\right)^{n-1}$	$1 - \mathbb{1}[\kappa \text{ accepted, visits } (s,a)]$
Matrix GRR	$\Sigma^{-1} = \frac{1}{2}\sum_{n=1}^{\infty}\left(I-\frac{\Sigma}{2}\right)^{n-1}$	$I - \mathbb{1}[\kappa \text{ accepted}]\, \phi_{\kappa,h}\phi_{\kappa,h}^\top$

- Issue: never "succeed" $\Longrightarrow$ truncate after $M \gtrsim 1/\lambda_{\min}(\Sigma_h^{\pi_k})$ attempts **(large!)**
- Also: one estimate changes **all** (s,a)'s $\Longrightarrow$ large M means π **changes fast**

Small $\lambda_{\min}$ = a rarely visited direction $\longrightarrow$ thresholding again!

Regions: No Rarely Visited Direction



- Dots: layer-h features of the last $\sqrt{K}$ episodes
- Group them into $\mathcal{O}(\log K)$ ellipsoids $\mathcal{C}_1, \mathcal{C}_2, \ldots$: each takes at least half of the remaining dots (elliptical potential lemma [APS11])
- $\mathcal{C}_j$'s empirical covariance $\widehat{\Sigma}_j$ has $\lambda_{\min} \gtrsim \widehat{p}_j$ (fraction of dots in $\mathcal{C}_j$) $\Longrightarrow$ truncate after $\approx \frac{d}{\widehat{p}_j} \leq \sqrt{K}$ attempts

- MGRR in each region: $\widetilde{\mathcal{O}}(d)$ unfinished estimates per region $\Longrightarrow$ π changes slowly
- Dots outside all regions: cover using bonuses (like rarely visted pairs in tabular)
- $\eta \approx K^{-1/2} \Longrightarrow$ π close for $\sqrt{K}$ episodes $\Longrightarrow$ outside regions' bonus $\approx K^{-1/2}$

Linear AMDPs: $\mathcal{R}_K = \widetilde{\mathcal{O}}\big(d^3H^4\sqrt{K}\big)$ w.h.p. (efficient & optimal in K)

More in the Paper

Actual delays: losses of episode k arrive $\mathfrak{d}_k$ episodes later, total delay $\mathfrak{D} := \sum_k \mathfrak{d}_k$

Our analysis already allows delays: start an estimate once its losses arrive

	best known	**ours**
Tabular	$HS\sqrt{AK} + H\sqrt{\mathfrak{D}}$ [Zie+25]	$H^{5/2}\sqrt{SAK} + H^3\sqrt{\mathfrak{D}}$
Linear	$K^{3/4} + \mathfrak{D}^{3/4}$, needs a simulator [LRS23]	$d^3H^4\sqrt{K} + d^{3/2}H^3\sqrt{\mathfrak{D}}$

Lower bounds: tabular $H\sqrt{SAK} + H\sqrt{\mathfrak{D}}$ [LRM22] and linear $dH^{3/2}\sqrt{K} + H\sqrt{\mathfrak{D}}$ [ZGS21]

Open questions

- Optimal H and d? ($H^{5/2}$ vs H and d^3H^4 vs $dH^{3/2}$; [LT26]: optimal H but $S\sqrt{AK}$)
- Other places where a changing policy must be evaluated, e.g., multi-agent RL?

Evaluate the current policy with future episodes, as if feedback was delayed – now introduced by us!

tabular: first $\sqrt{SAK}$ linear: first efficient $\sqrt{K}$ extensions to delays: $\sqrt{\mathfrak{D}}$

Thank you!

https://arxiv.org/abs/2611.17214/
yandai20@mit.edu https://yandaichn.github.io/

References I

[DLC22] Yan Dai, Haipeng Luo, and Liyu Chen. "Follow-the-Perturbed-Leader for Adversarial Markov Decision Processes with Bandit Feedback". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 11437–11449.

[DLWZ23] Yan Dai, Haipeng Luo, Chen-Yu Wei, and Julian Zimmert. "Refined Regret for Adversarial MDPs with Linear Function Approximation". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 6726–6759.

[APS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems* 24 (2011).

[DWZ23] Christoph Dann, Chen-Yu Wei, and Julian Zimmert. "Best of Both Worlds Policy Optimization". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 6968–7008.

[FS97] Yoav Freund and Robert E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[JABJ18] Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. "Is Q-Learning Provably Efficient?" In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[Jin+20] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. "Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4860–4869.

[Jin+22] Tiancheng Jin, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. "Near-Optimal Regret for Adversarial MDP with Delayed Bandit Feedback". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.

References II

[KL02] Sham Kakade and John Langford. "Approximately Optimal Approximate Reinforcement Learning". In: *Proceedings of the 19th International Conference on Machine Learning*. 2002, pp. 267–274.

[KZWL24] Fang Kong, XiangCheng Zhang, Baoxiang Wang, and Shuai Li. "Improved Regret Bounds for Linear Adversarial MDPs via Linear Optimization". In: *Transactions on Machine Learning Research* (2024). ISSN: 2835-8856. URL: https://openreview.net/forum?id=KcmWZSk53y.

[LRM22] Tal Lancewicki, Aviv Rosenberg, and Yishay Mansour. "Learning Adversarial Markov Decision Processes with Delayed Feedback". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2022, pp. 7281–7289.

[LRS23] Tal Lancewicki, Aviv Rosenberg, and Dmitry Sotnikov. "Delay-Adapted Policy Optimization and Improved Regret for Adversarial MDP with Delayed Bandit Feedback". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 18482–18534.

[LT26] Mingyi Li and Taira Tsuchiya. "Closing the Horizon Gap in Policy Optimization for Adversarial MDPs". In: *arXiv preprint arXiv:2610.12362* (2026).

[LWL21] Haipeng Luo, Chen-Yu Wei, and Chung-Wei Lee. "Policy Optimization in Adversarial MDPs: Improved Exploration via Dilated Bonuses". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 22931–22942.

[LWZ24] Haolin Liu, Chen-Yu Wei, and Julian Zimmert. "Towards Optimal Regret in Adversarial Linear MDPs with Bandit Feedback". In: *The Twelfth International Conference on Learning Representations*. 2024.

[LWZ26] Haolin Liu, Chen-Yu Wei, and Hanbin Zhou. "A Computationally Efficient and Rate-Optimal Algorithm for Adversarial Linear MDPs with (Aggregate) Bandit Feedback". In: *arXiv preprint arXiv:2610.30801* (2026).

References III

[NB16] Gergely Neu and Gábor Bartók. "Importance Weighting Without Importance Weights: An Efficient Algorithm for Combinatorial Semi-Bandits". In: *Journal of Machine Learning Research* 17.154 (2016), pp. 1–21.

[NO21] Gergely Neu and Julia Olkhovskaya. "Online Learning in MDPs with Linear Function Approximation and Bandit Feedback". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 10407–10417.

[RM19] Aviv Rosenberg and Yishay Mansour. "Online Stochastic Shortest Path with Bandit Feedback and Unknown Transition Function". In: *Advances in Neural Information Processing Systems* 32 (2019).

[SKM23] Uri Sherman, Tomer Koren, and Yishay Mansour. "Improved Regret for Efficient Online Reinforcement Learning with Linear Function Approximation". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 31117–31150.

[ZGS21] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. "Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes". In: *Conference on Learning Theory*. PMLR. 2021, pp. 4532–4576.

[Zie+25] Lukas Zierahn, Dirk van der Hoeven, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. "A Unified Analysis of Nonstochastic Delayed Feedback for Combinatorial Semi-Bandits, Linear Bandits, and MDPs". In: *Journal of Machine Learning Research* 26.104 (2025), pp. 1–60.