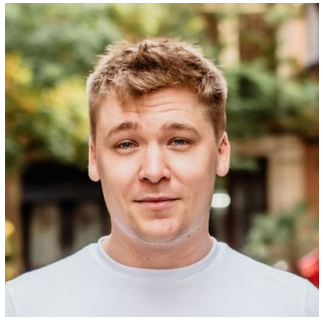
Oberwolfach, August 2026

An Improved Algorithm for Adversarial Linear Contextual Bandits via Reduction

Tim van Erven

UNIVERSITY OF AMSTERDAM

Joint work with:



Jack Mayo

Julia Olkhovskaya

Chen-Yu Wei

Outline

Setting

1. Adversarial Bandits
2. Adversarial Linear Contextual Bandits

Main Results

3. General Setting
4. First-order Bounds, Initial Setting

Approach: Reduction to Non-contextual Linear Bandits

5. Reduction: The Basic Idea
6. Approximating Ω
7. Side-Result: Efficient Robust Linear Bandit Algorithm
8. Controlling the Difference between Ψ and $\hat{\Psi}$
9. Restricting π_t to be a Mixture of Linear Policies

Adversarial Bandits

For $t = 1, \ldots, T$:

1. Learner choses (randomized) arm $a_t \in \{1, \ldots, K\}$
2. Loss value $\ell_t(a_t)$ is revealed

Regret w.r.t. arm a: $$R_T(a) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \sum_{t=1}^{T} \ell_t(a)$$

Adversarial Bandits

For $t = 1, \ldots, T$:

1. Learner choses (randomized) arm $a_t \in \{1, \ldots, K\}$
2. Loss value $\ell_t(a_t)$ is revealed

Regret w.r.t. arm a: $$R_T(a) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \sum_{t=1}^{T} \ell_t(a)$$

- Oblivious adversary: losses $\ell_t(a)$ fixed a priori for all t, a
- Expectation w.r.t. learner's randomness

Adversarial Contextual Bandits

For $t = 1, \ldots, T$:

1. **Context** $X_t \in \mathbb{R}^p$ is revealed
2. Learner choses (randomized) arm $a_t \in \{1, \ldots, K\}$
3. Loss value $\ell_t(a_t)$ is revealed

Regret w.r.t. policy π:

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi)\right]$$

$$\ell_t(\pi) = \mathop{\mathbb{E}}_{a \sim \pi(X_t)}[\ell_t(a)] \quad \text{for } \pi(X_t) \in \Delta_K$$

Adversarial Contextual Bandits

For $t = 1, \ldots, T$:

1. **Context** $X_t \in \mathbb{R}^p$ is revealed
2. Learner choses (randomized) arm $a_t \in \{1, \ldots, K\}$
3. Loss value $\ell_t(a_t)$ is revealed

Regret w.r.t. policy π:

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi)\right]$$

$$\ell_t(\pi) = \mathop{\mathbb{E}}_{a \sim \pi(X_t)}[\ell_t(a)] \quad \text{for } \pi(X_t) \in \Delta_K$$

Adversarial losses, stochastic contexts [Neu and Olkhovskaya, 2020]:

- **Linear losses:** $\ell_t(a) = \langle X_t, \theta_{t,a} \rangle \in [0, 1]$, where $\theta_{t,a}$ fixed a priori
- **I.i.d. contexts**: $X_t \sim \mathcal{D}$

Adversarial Contextual Bandits

For $t = 1, \ldots, T$:

1. **Context** $X_t \in \mathbb{R}^p$ is revealed
2. Learner choses (randomized) arm $a_t \in \{1, \ldots, K\}$
3. Loss value $\ell_t(a_t)$ is revealed

Regret w.r.t. policy π:

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi)\right]$$

$$\ell_t(\pi) = \mathop{\mathbb{E}}_{a \sim \pi(X_t)}[\ell_t(a)] \quad \text{for } \pi(X_t) \in \Delta_K$$

Adversarial losses, stochastic contexts [Neu and Olkhovskaya, 2020]:

- **Linear losses:** $\ell_t(a) = \langle X_t, \theta_{t,a} \rangle \in [0, 1]$, where $\theta_{t,a}$ fixed a priori
- **I.i.d. contexts**: $X_t \sim \mathcal{D}$
- (Opposite setting with fixed loss function and adversarial contexts also considered in literature.)

Adversarial Contextual Bandits More Abstractly I

Incorporate contexts $X_t \in \mathbb{R}^p$ into actions a such that $\ell_t(a) = \langle X_t, \theta_{t,a} \rangle = \langle a, \theta_t \rangle$:

$$\theta_t = \begin{pmatrix} \theta_{t,1} \\ \theta_{t,2} \\ \vdots \\ \theta_{t,K} \end{pmatrix} \in \mathbb{R}^{p \times K} \qquad a = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ X_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{p \times K}$$

Then every round we receive a **random action set** (with K actions):

$$\mathcal{A}_t = \left\{ \begin{pmatrix} X_t \\ \\ \end{pmatrix}, \begin{pmatrix} \\ X_t \\ \end{pmatrix}, \ldots, \begin{pmatrix} \\ \\ X_t \end{pmatrix} \right\} \subset \mathbb{R}^{p \times K}$$

Adversarial Contextual Bandits More Abstractly II

For $t = 1, \ldots, T$:

1. Draw **action set** $\mathcal{A}_t \subset \mathbb{R}^d$ i.i.d. from $\mathcal{D}$
2. Learner choses (randomized) action $a_t \in \mathcal{A}_t$
3. Loss value $\ell_t(a_t) = \langle a_t, \theta_t \rangle \in [0, 1]$ is revealed

Regret w.r.t. policy π: $$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi)\right]$$

$$\ell_t(\pi) = \langle \pi(\mathcal{A}_t), \theta_t \rangle \quad \text{for } \pi(\mathcal{A}_t) \in \text{conv}(\mathcal{A}_t)$$

Adversarial Contextual Bandits More Abstractly II

For $t = 1, \ldots, T$:

1. Draw **action set** $\mathcal{A}_t \subset \mathbb{R}^d$ i.i.d. from $\mathcal{D}$
2. Learner choses (randomized) action $a_t \in \mathcal{A}_t$
3. Loss value $\ell_t(a_t) = \langle a_t, \theta_t \rangle \in [0, 1]$ is revealed

Regret w.r.t. policy π: $$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi)\right]$$

$$\ell_t(\pi) = \langle \pi(\mathcal{A}_t), \theta_t \rangle \quad \text{for } \pi(\mathcal{A}_t) \in \mathrm{conv}(\mathcal{A}_t)$$

Optimal policy is linear:

$$\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \langle \pi(\mathcal{A}_t), \theta_t \rangle\right] = \min_{\pi} \mathbb{E}\left[\langle \pi(\mathcal{A}), \sum_{t=1}^{T} \theta_t \rangle\right]$$

$$\pi^*(\mathcal{A}) = \pi_\phi(\mathcal{A}) := \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle \quad \text{for } \phi = \sum_{t=1}^{T} \theta_t$$

Adversarial Contextual Bandits More Abstractly II

For $t = 1, \ldots, T$:

1. Draw **action set** $\mathcal{A}_t \subset \mathbb{R}^d$ i.i.d. from $\mathcal{D}$
2. Learner choses (randomized) action $a_t \in \mathcal{A}_t$
3. Loss value $\ell_t(a_t) = \langle a_t, \theta_t \rangle \in [0, 1]$ is revealed

Regret w.r.t. policy π: $$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi)\right]$$

$$\ell_t(\pi) = \langle \pi(\mathcal{A}_t), \theta_t \rangle \quad \text{for } \pi(\mathcal{A}_t) \in \mathrm{conv}(\mathcal{A}_t)$$

Optimal policy is linear:

$$\min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \langle \pi(\mathcal{A}_t), \theta_t \rangle\right] = \min_{\pi} \mathbb{E}\left[\langle \pi(\mathcal{A}), \sum_{t=1}^{T} \theta_t \rangle\right]$$

$$\pi^*(\mathcal{A}) = \pi_\phi(\mathcal{A}) := \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle \quad \text{for } \phi = \sum_{t=1}^{T} \theta_t$$

Linear optimization oracle:

- Returns $\pi_\phi(\mathcal{A}) = \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle$ for any $\phi \in \mathbb{R}^d$ and $\mathcal{A}$

Outline

Setting

1. Adversarial Bandits
2. Adversarial Linear Contextual Bandits

Main Results

3. General Setting
4. First-order Bounds, Initial Setting

Approach: Reduction to Non-contextual Linear Bandits

5. Reduction: The Basic Idea
6. Approximating Ω
7. Side-Result: Efficient Robust Linear Bandit Algorithm
8. Controlling the Difference between Ψ and $\hat{\Psi}$
9. Restricting π_t to be a Mixture of Linear Policies

Main Results I: General Setting

- d: dimension of the actions, $\mathcal{A}_t \subset \mathbb{R}^d$
- K: maximum number of actions, $|\mathcal{A}_t| \leq K$
- Simulator: free access to independent samples $\mathcal{A} \sim \mathcal{D}$

Algorithm	**Regret**[1]	**Runtime**	**Simulator**
Dai et al. [2023]	$\min\{d\sqrt{T}, \sqrt{dT \log K}\}$	$\mathrm{poly}(d, K, T)$	yes
Liu et al. [2023]	$d\sqrt{T}$	$K \cdot T^{\Omega(d)}$	no
Liu et al. [2023]	$d^2\sqrt{T}$	$\mathrm{poly}(d, K, T)$	no
Ours	$\min\{d^2\sqrt{T}, \sqrt{d^3 T \log K}\}$	$\mathrm{poly}(d, T)$ plus $\mathrm{poly}(d, T)$ **oracle calls**	no
Ours	$d\sqrt{T}$	$\mathrm{poly}(d, T)$ plus $\mathrm{poly}(d, T)$ **oracle calls**	yes

[1] Up to poly-logarithmic factors in d and T

Main Results I: General Setting

- d: dimension of the actions, $\mathcal{A}_t \subset \mathbb{R}^d$
- K: maximum number of actions, $|\mathcal{A}_t| \leq K$
- Simulator: free access to independent samples $\mathcal{A} \sim \mathcal{D}$

Algorithm	Regret[1]	Runtime	Simulator
Dai et al. [2023]	$\min\{d\sqrt{T}, \sqrt{dT \log K}\}$	$\text{poly}(d, K, T)$	yes
Liu et al. [2023]	$d\sqrt{T}$	$K \cdot T^{\Omega(d)}$	no
Liu et al. [2023]	$d^2\sqrt{T}$	$\text{poly}(d, K, T)$	no
Ours	$\min\{d^2\sqrt{T}, \sqrt{d^3 T \log K}\}$	$\text{poly}(d, T)$ plus $\text{poly}(d, T)$ **oracle calls**	no
Ours	$d\sqrt{T}$	$\text{poly}(d, T)$ plus $\text{poly}(d, T)$ **oracle calls**	yes

For semi-bandit feedback:

Neu and Valko [2014]	$(dT)^{2/3}$	$\text{poly}(d, T)$ plus $\text{poly}(d, T)$ **oracle calls**	no

[1] Up to poly-logarithmic factors in d and T

Main Results I: General Setting

- d: dimension of the actions, $\mathcal{A}_t \subset \mathbb{R}^d$
- K: maximum number of actions, $|\mathcal{A}_t| \leq K$
- Simulator: free access to independent samples $\mathcal{A} \sim \mathcal{D}$

Algorithm	Regret[1]	Runtime	Simulator
Dai et al. [2023]	$\min\{d\sqrt{T}, \sqrt{dT \log K}\}$	$\text{poly}(d, K, T)$	yes
Liu et al. [2023]	$d\sqrt{T}$	$K \cdot T^{\Omega(d)}$	no
Liu et al. [2023]	$d^2\sqrt{T}$	$\text{poly}(d, K, T)$	no
Ours	$\min\{d^2\sqrt{T}, \sqrt{d^3 T \log K}\}$	$\text{poly}(d, T)$ plus $\text{poly}(d, T)$ **oracle calls**	no
Ours	$d\sqrt{T}$	$\text{poly}(d, T)$ plus $\text{poly}(d, T)$ **oracle calls**	yes

- For combinatorial problems, Oracle can be cheap $O(d)$ while $K = 2^{\Omega(d)}$ is exponentially large
  - Example: shortest path with d edges

[1] Up to poly-logarithmic factors in d and T

Main Results I: General Setting

- d: dimension of the actions, $\mathcal{A}_t \subset \mathbb{R}^d$
- K: maximum number of actions, $|\mathcal{A}_t| \leq K$
- Simulator: free access to independent samples $\mathcal{A} \sim \mathcal{D}$

Algorithm	Regret	Runtime	Simulator
Dai et al. [2023]	$\min\{d\sqrt{T}, \sqrt{dT \log K}\}$	$\text{poly}(d, K, T)$	yes
Liu et al. [2023]	$d\sqrt{T}$	$K \cdot T^{\Omega(d)}$	no
Liu et al. [2023]	$d^2\sqrt{T}$	$\text{poly}(d, K, T)$	no
Ours	$\min\{d^2\sqrt{T}, \sqrt{d^3 T \log K}\}$	$\text{poly}(d, T)$ plus $\text{poly}(d, T)$ **oracle calls**	no
Ours	$d\sqrt{L^*}$	$\text{poly}(d, T)$ plus $\text{poly}(d, T)$ **oracle calls**	yes

$$L^* = \min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \langle \pi(\mathcal{A}_t), \theta_t \rangle\right] \leq T$$

Main Results II: First-order Bounds, Initial Setting

- p: dimension of contexts, $X_t \in \mathbb{R}^p$
- K: number of actions, $|\mathcal{A}_t| = K$
- Simulator: free access to independent context samples $X \sim \mathcal{D}$

$$L^* = \min_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \langle \pi(\mathcal{A}_t), \theta_t \rangle\right]$$

Algorithm	Regret[1]	Runtime	Simulator	Note
Neu and Olkhovskaya [2020]	$\sqrt{KpT}$	$\mathrm{poly}(p, K, T)$	yes	
Olkhovskaya et al. [2023]	$K\sqrt{pL^*}$	$\Theta\left(T\left(\frac{T}{K^2 p}\right)^{Kp}\right)$	no	
Olkhovskaya et al. [2023]	$K\sqrt{pL^*}$	$\mathrm{poly}(p, K, T)$	yes	$\star$
Ours	$Kp\sqrt{L^*}$	$\mathrm{poly}(p, K, T)$	yes	

Strong assumption $\star$: contexts X_t have log-concave distribution

[1] Up to poly-logarithmic factors

Outline

Setting

1. Adversarial Bandits
2. Adversarial Linear Contextual Bandits

Main Results

3. General Setting
4. First-order Bounds, Initial Setting

Approach: Reduction to Non-contextual Linear Bandits

5. Reduction: The Basic Idea
6. Approximating Ω
7. Side-Result: Efficient Robust Linear Bandit Algorithm
8. Controlling the Difference between Ψ and $\hat{\Psi}$
9. Restricting π_t to be a Mixture of Linear Policies

Reduction: The Basic Idea

Expected loss for policy π in round t is:

$$\mathop{\mathbb{E}}_{\mathcal{A}_t}\left[\langle \pi(\mathcal{A}_t), \theta_t \rangle\right] = \left\langle \mathop{\mathbb{E}}_{\mathcal{A}_t}[\pi(\mathcal{A}_t)], \theta_t \right\rangle = \left\langle \Psi(\pi), \theta_t \right\rangle,$$

where $\Psi(\pi)$ **is the mean action** for π:

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \in \mathbb{R}^d.$$

Reduction: The Basic Idea

Expected loss for policy π in round t is:

$$\mathop{\mathbb{E}}_{\mathcal{A}_t}\left[\langle \pi(\mathcal{A}_t), \theta_t \rangle\right] = \left\langle \mathop{\mathbb{E}}_{\mathcal{A}_t}[\pi(\mathcal{A}_t)], \theta_t \right\rangle = \left\langle \Psi(\pi), \theta_t \right\rangle,$$

where $\Psi(\pi)$ **is the mean action** for π:

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \in \mathbb{R}^d.$$

Possibilities for expected loss:

$$\langle y, \theta_t \rangle \qquad \text{for } y \in \Omega = \{\Psi(\pi) \mid \pi \in \Pi\}$$

Reduction: The Basic Idea

Expected loss for policy π in round t is:

$$\mathop{\mathbb{E}}_{\mathcal{A}_t}\left[\langle \pi(\mathcal{A}_t), \theta_t \rangle\right] = \left\langle \mathop{\mathbb{E}}_{\mathcal{A}_t}[\pi(\mathcal{A}_t)], \theta_t \right\rangle = \left\langle \Psi(\pi), \theta_t \right\rangle,$$

where $\Psi(\pi)$ **is the mean action** for π:

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \in \mathbb{R}^d.$$

Possibilities for expected loss:

$$\langle y, \theta_t \rangle \qquad \text{for } y \in \Omega = \{\Psi(\pi) \mid \pi \in \Pi\}$$

Reduction:

1. Let linear bandit algorithm choose $y_t \in \Omega$ in round t
2. Play π_t such that $\Psi(\pi_t) = y_t$
3. Provide unbiased loss estimate $\langle \pi_t(\mathcal{A}_t), \theta_t \rangle$ as feedback to linear bandit algorithm

Reduction: The Basic Idea

Expected loss for policy π in round t is:

$$\mathop{\mathbb{E}}_{\mathcal{A}_t}\left[\langle \pi(\mathcal{A}_t), \theta_t \rangle\right] = \left\langle \mathop{\mathbb{E}}_{\mathcal{A}_t}[\pi(\mathcal{A}_t)], \theta_t \right\rangle = \left\langle \Psi(\pi), \theta_t \right\rangle,$$

where $\Psi(\pi)$ **is the mean action** for π:

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \in \mathbb{R}^d.$$

Possibilities for expected loss:

$$\langle y, \theta_t \rangle \qquad \text{for } y \in \Omega = \{\Psi(\pi) \mid \pi \in \Pi\}$$

Reduction:

1. Let linear bandit algorithm choose $y_t \in \Omega$ in round t
2. Play π_t such that $\Psi(\pi_t) = y_t$
3. Provide unbiased loss estimate $\langle \pi_t(\mathcal{A}_t), \theta_t \rangle$ as feedback to linear bandit algorithm

NB Hanna et al. [2023] introduced this reduction for different setting of **stochastic** linear contextual bandits, but their techniques do not carry over.

Approximating Ω

Reduction:

1. Let linear bandit algorithm choose $y_t \in \Omega$ in round t
2. Play π_t such that $\Psi(\pi_t) = y_t$
3. Provide unbiased loss estimate $\langle \pi_t(\mathcal{A}_t), \theta_t \rangle$ as feedback to linear bandit algorithm

$$\Omega = \{\Psi(\pi) \mid \pi \in \Pi\}, \qquad \Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})]$$

Issue: Ψ and Ω depend on **unknown distribution $\mathcal{D}$ of $\mathcal{A}$**

Approximating Ω

Reduction:

1. Let linear bandit algorithm choose $y_t \in \Omega$ in round t
2. Play π_t such that $\Psi(\pi_t) = y_t$
3. Provide unbiased loss estimate $\langle \pi_t(\mathcal{A}_t), \theta_t \rangle$ as feedback to linear bandit algorithm

$$\Omega = \{\Psi(\pi) \mid \pi \in \Pi\}, \qquad \Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})]$$

Issue: Ψ and Ω depend on **unknown distribution $\mathcal{D}$ of $\mathcal{A}$**

Using simulator: Given separate sample $\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_N$ from $\mathcal{D}$:

$$\hat{\Omega} = \{\hat{\Psi}(\pi) \mid \pi \in \Pi\}, \qquad \hat{\Psi}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \pi(\tilde{\mathcal{A}}_i)$$

Approximating Ω

Reduction:

1. Let linear bandit algorithm choose $y_t \in \hat{\Omega}$ in round t
2. Play π_t such that $\hat{\Psi}(\pi_t) \approx y_t$
3. Provide biased loss estimate $\langle \pi_t(\mathcal{A}_t), \theta_t \rangle$ as feedback to linear bandit algorithm

$$\Omega = \{\Psi(\pi) \mid \pi \in \Pi\}, \qquad \Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})]$$

Using simulator: Given separate sample $\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_N$ from $\mathcal{D}$:

$$\hat{\Omega} = \{\hat{\Psi}(\pi) \mid \pi \in \Pi\}, \qquad \hat{\Psi}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \pi(\tilde{\mathcal{A}}_i)$$

- Need computationally efficient adversarial linear bandit algorithm that is **robust to biased stochastic feedback**

Approximating Ω

Reduction:

1. Let linear bandit algorithm choose $y_t \in \hat{\Omega}$ in round t
2. Play π_t such that $\hat{\Psi}(\pi_t) \approx y_t$
3. Provide biased loss estimate $\langle \pi_t(\mathcal{A}_t), \theta_t \rangle$ as feedback to linear bandit algorithm

$$\Omega = \{\Psi(\pi) \mid \pi \in \Pi\}, \qquad \Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})]$$

Using simulator: Given separate sample $\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_N$ from $\mathcal{D}$:

$$\hat{\Omega} = \{\hat{\Psi}(\pi) \mid \pi \in \Pi\}, \qquad \hat{\Psi}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \pi(\tilde{\mathcal{A}}_i)$$

Proposition

Linear optimization over $\hat{\Omega}$ costs N oracle calls: for any $\phi \in \mathbb{R}^d$,

$$\hat{\Psi}(\pi_\phi) = \frac{1}{N} \sum_{i=1}^{N} \arg\min_{a \in \tilde{\mathcal{A}}_i} \langle a, \phi \rangle \in \arg\min_{v \in \hat{\Omega}} \langle v, \phi \rangle$$

Side-Result: Efficient Robust Linear Bandit Algorithm

Definition (α-misspecification-robust linear bandit algorithm)

Given random feedback $c_t \in [0,1]$ with bias at most some **known** $\epsilon \geq 0$:

$$\left|\mathbb{E}_t[c_t] - \langle y_t, \theta_t \rangle\right| \leq \epsilon,$$

the algorithm achieves regret at most

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle y_t, \theta_t \rangle\right] \leq \min_{y \in \hat{\Omega}} \sum_{t=1}^{T} \langle y, \theta_t \rangle + \tilde{O}\left(d\sqrt{T} + \alpha\sqrt{d}\epsilon T\right).$$

- [Liu et al., 2024a]: optimal $\alpha = 1$, but runtime scales with number of actions K
- New alg: $\alpha = \sqrt{d}$, and runtime poly(d, T) + poly(d, T) oracle calls
  - Version of continuous exponential weights similar to Ito et al. [2020] with bonuses to handle misspecification like [Lee et al., 2020, Zimmert and Lattimore, 2022, Liu et al., 2024b]
  - Can implement efficiently given linear optimization oracle over $\hat{\Omega}$
  - Also achieves first-order bound

Controlling the Difference between ψ and $\hat{\psi}$

Are we there yet?

Controlling the Difference between Ψ and $\hat{\Psi}$

Suppose, for $t = 1, \ldots, T$ and $\pi \in \{\pi^*, \pi_t\}$:

uniform convergence: $|\langle \Psi(\pi), \theta_t \rangle - \langle \hat{\Psi}(\pi), \theta_t \rangle| \leq \gamma_1$ (w.h.p.)

approximate play: $|\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \gamma_2$ (deterministic)

Controlling the Difference between Ψ and $\hat{\Psi}$

Suppose, for $t = 1, \ldots, T$ and $\pi \in \{\pi^*, \pi_t\}$:

uniform convergence: $|\langle \Psi(\pi), \theta_t \rangle - \langle \hat{\Psi}(\pi), \theta_t \rangle| \leq \gamma_1$ (w.h.p.)

approximate play: $|\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \gamma_2$ (deterministic)

This would resolve the following **two remaining issues:**

1. Controlling feedback bias:

$$|\mathop{\mathbb{E}}_{\mathcal{A}_t}[\langle \pi_t(\mathcal{A}_t), \theta_t \rangle] - \langle y_t, \theta_t \rangle| \leq |\langle \Psi(\pi_t), \theta_t \rangle - \langle \hat{\Psi}(\pi_t), \theta_t \rangle|$$
$$+ |\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \gamma_1 + \gamma_2 =: \epsilon$$

Controlling the Difference between Ψ and $\hat{\Psi}$

Suppose, for $t = 1, \ldots, T$ and $\pi \in \{\pi^*, \pi_t\}$:

uniform convergence: $|\langle \Psi(\pi), \theta_t \rangle - \langle \hat{\Psi}(\pi), \theta_t \rangle| \leq \gamma_1$ (w.h.p.)

approximate play: $|\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \gamma_2$ (deterministic)

This would resolve the following **two remaining issues:**

1. Controlling feedback bias:

$$|\mathop{\mathbb{E}}_{\mathcal{A}_t}[\langle \pi_t(\mathcal{A}_t), \theta_t \rangle] - \langle y_t, \theta_t \rangle| \leq |\langle \Psi(\pi_t), \theta_t \rangle - \langle \hat{\Psi}(\pi_t), \theta_t \rangle|$$
$$+ |\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \gamma_1 + \gamma_2 =: \epsilon$$

2. Linear bandit gives regret bound w.r.t. $y^* \in \hat{\Omega}$ instead of Ω:

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(\langle y_t, \theta_t \rangle - \langle y^*, \theta_t \rangle\right)\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \left(\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle \hat{\Psi}(\pi^*), \theta_t \rangle\right)\right] - T\gamma_2$$
$$\geq \mathbb{E}\left[\sum_{t=1}^{T} \left(\langle \Psi(\pi_t), \theta_t \rangle - \langle \Psi(\pi^*), \theta_t \rangle\right)\right] - 2T\gamma_1 - T\gamma_2$$

Uniform Convergence

To show: with high probability, $|\langle \Psi(\pi), \theta_t \rangle - \langle \hat{\Psi}(\pi), \theta_t \rangle| \leq \gamma_1$ for $\pi \in \{\pi^*, \pi_t\}$, $t = 1, \ldots, T$.

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \qquad \hat{\Psi}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \pi(\tilde{\mathcal{A}}_i)$$

Lemma (Uniform Convergence over Linear Policies)

Let $\pi_\phi(\mathcal{A}) := \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle$ be a linear policy. Then, w.p. $\geq 1 - \delta$,

$$\sup_{\phi} \left| \langle \Psi(\pi_\phi), \theta_t \rangle - \langle \hat{\Psi}(\pi_\phi), \theta_t \rangle \right| \leq \gamma_1 := 2\sqrt{\frac{2d \ln(NK^2)}{N}} + \sqrt{\frac{2 \ln(4/\delta)}{N}}.$$

Uniform Convergence

To show: with high probability, $|\langle \Psi(\pi), \theta_t \rangle - \langle \hat{\Psi}(\pi), \theta_t \rangle| \leq \gamma_1$ for $\pi \in \{\pi^*, \pi_t\}$, $t = 1, \ldots, T$.

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \qquad \hat{\Psi}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \pi(\tilde{A}_i)$$

Lemma (Uniform Convergence over Linear Policies)

Let $\pi_\phi(\mathcal{A}) := \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle$ be a linear policy. Then, w.p. $\geq 1 - \delta$,

$$\sup_{\phi} \left| \langle \Psi(\pi_\phi), \theta_t \rangle - \langle \hat{\Psi}(\pi_\phi), \theta_t \rangle \right| \leq \gamma_1 := 2\sqrt{\frac{2d \ln(NK^2)}{N}} + \sqrt{\frac{2 \ln(4/\delta)}{N}}.$$

- Union bound over $t = 1, \ldots, T$ is cheap: δ/T instead of δ

Uniform Convergence

To show: with high probability, $|\langle \Psi(\pi), \theta_t \rangle - \langle \hat{\Psi}(\pi), \theta_t \rangle| \leq \gamma_1$ for $\pi \in \{\pi^*, \pi_t\}$, $t = 1, \ldots, T$.

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \qquad \hat{\Psi}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \pi(\tilde{A}_i)$$

Lemma (Uniform Convergence over Linear Policies)

Let $\pi_\phi(\mathcal{A}) := \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle$ be a linear policy. Then, w.p. $\geq 1 - \delta$,

$$\sup_{\phi} \left| \langle \Psi(\pi_\phi), \theta_t \rangle - \langle \hat{\Psi}(\pi_\phi), \theta_t \rangle \right| \leq \gamma_1 := 2\sqrt{\frac{2d \ln(NK^2)}{N}} + \sqrt{\frac{2 \ln(4/\delta)}{N}}.$$

- Union bound over $t = 1, \ldots, T$ is cheap: δ/T instead of δ
- We know π^* is always a linear policy, but algorithm's choices π_t may not be! **Problem!**

Uniform Convergence

To show: with high probability, $|\langle \Psi(\pi), \theta_t \rangle - \langle \hat{\Psi}(\pi), \theta_t \rangle| \leq \gamma_1$ for $\pi \in \{\pi^*, \pi_t\}$, $t = 1, \ldots, T$.

$$\Psi(\pi) = \mathop{\mathbb{E}}_{\mathcal{A}}[\pi(\mathcal{A})] \qquad \hat{\Psi}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \pi(\tilde{A}_i)$$

Lemma (Uniform Convergence over Linear Policies)

Let $\pi_\phi(\mathcal{A}) := \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle$ be a linear policy. Then, w.p. $\geq 1 - \delta$,

$$\sup_{\phi} \left| \langle \Psi(\pi_\phi), \theta_t \rangle - \langle \hat{\Psi}(\pi_\phi), \theta_t \rangle \right| \leq \gamma_1 := 2\sqrt{\frac{2d \ln(NK^2)}{N}} + \sqrt{\frac{2 \ln(4/\delta)}{N}}.$$

- Union bound over $t = 1, \ldots, T$ is cheap: δ/T instead of δ
- We know π^* is always a linear policy, but algorithm's choices π_t may not be! **Problem!**
- Can we solve this by extending to uniform convergence over all policies π? No, does not hold! So need to ensure π_t is a (mixture of) linear policies.

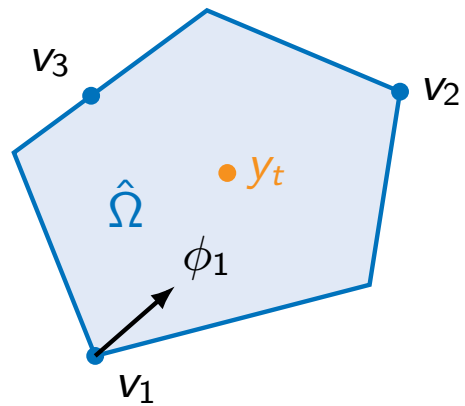
Restricting π_t to be a Mixture of Linear Policies

Proposition (Repeated)

Linear optimization over $\hat{\Omega}$ costs N oracle calls: for any $\phi \in \mathbb{R}^d$,

$$\hat{\Psi}(\pi_\phi) = \frac{1}{N}\sum_{i=1}^{N} \arg\min_{a \in \tilde{\mathcal{A}}_i} \langle a, \phi \rangle \in \arg\min_{v \in \hat{\Omega}} \langle v, \phi \rangle$$

- Linear policies only reach the **boundary** of $\hat{\Omega}$
- y_t can be anywhere in $\hat{\Omega}$



$$v_j = \hat{\Psi}(\pi_{\phi_j})$$

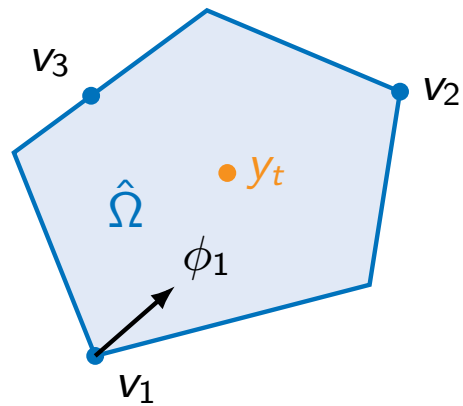
Restricting π_t to be a Mixture of Linear Policies

Proposition (Repeated)

Linear optimization over $\hat{\Omega}$ costs N oracle calls: for any $\phi \in \mathbb{R}^d$,

$$\hat{\Psi}(\pi_\phi) = \frac{1}{N}\sum_{i=1}^{N} \arg\min_{a \in \tilde{\mathcal{A}}_i} \langle a, \phi \rangle \in \arg\min_{v \in \hat{\Omega}} \langle v, \phi \rangle$$

- Linear policies only reach the **boundary** of $\hat{\Omega}$
- y_t can be anywhere in $\hat{\Omega}$



$$v_j = \hat{\Psi}(\pi_{\phi_j})$$

Solution: Let π_t play π_{ϕ_j} for $j \sim \mathrm{Unif}\{1, \ldots, M\}$, so that

$$\hat{\Psi}(\pi_t) = \frac{1}{M}\sum_{j=1}^{M} \hat{\Psi}(\pi_{\phi_j}) \approx y_t.$$

Finding Directions ϕ_j with Frank-Wolfe

Frank-Wolfe to minimize any convex f over $\hat{\Omega}$ (with $z_0 = 0$):

$$v_j \in \arg\min_{v \in \hat{\Omega}} \langle v, \nabla f(z_{j-1}) \rangle$$
$$z_j = (1 - \eta_j) z_{j-1} + \eta_j v_j$$

Theorem (e.g. Bubeck, 2015)

Suppose f is β-smooth w.r.t. $\|\cdot\|$ and let $R = \sup_{u,w \in \hat{\Omega}} \|u - w\|$ be the diameter of $\hat{\Omega}$. Let $\eta_j = \frac{2}{j+1}$. Then

$$f(z_M) - f(z^*) \leq \frac{2\beta R^2}{M + 2}.$$

Finding Directions ϕ_j with Frank-Wolfe

Frank-Wolfe to minimize any convex f over $\hat{\Omega}$ (with $z_0 = 0$):

$$v_j \in \arg\min_{v \in \hat{\Omega}} \langle v, \nabla f(z_{j-1}) \rangle$$
$$z_j = (1 - \eta_j) z_{j-1} + \eta_j v_j$$

Lemma

Let $f(z) = \frac{1}{2}\|z - y_t\|^2$ *for* $y_t \in \hat{\Omega}$, *so that* $\beta = 1$ *and* $f(z^*) = 0$, *and let* $\eta_j = \frac{1}{j}$. *Then* $z_M = \frac{1}{M}\sum_{j=1}^{M} v_j$ *and*

$$\frac{1}{2}\|z_M - y_t\|^2 = f(z_M) \leq \frac{R^2}{2M}.$$

Finding Directions ϕ_j with Frank-Wolfe

Frank-Wolfe to minimize any convex f over $\hat{\Omega}$ (with $z_0 = 0$):

$$v_j \in \arg\min_{v \in \hat{\Omega}} \langle v, \nabla f(z_{j-1}) \rangle$$
$$z_j = (1 - \eta_j) z_{j-1} + \eta_j v_j$$

Lemma

Let $f(z) = \frac{1}{2}\|z - y_t\|^2$ *for* $y_t \in \hat{\Omega}$, *so that* $\beta = 1$ *and* $f(z^*) = 0$, *and let* $\eta_j = \frac{1}{j}$. *Then* $z_M = \frac{1}{M}\sum_{j=1}^{M} v_j$ *and*

$$\frac{1}{2}\|z_M - y_t\|^2 = f(z_M) \leq \frac{R^2}{2M}.$$

Finding Directions ϕ_j with Frank-Wolfe

Frank-Wolfe to minimize any convex f over $\hat{\Omega}$ (with $z_0 = 0$):

$$\begin{aligned}\phi_j &= \nabla f(z_{j-1}) = z_{j-1} - y_t \\ v_j &\in \arg\min_{v \in \hat{\Omega}} \langle v, \phi_j \rangle \\ z_j &= (1 - \eta_j) z_{j-1} + \eta_j v_j\end{aligned}$$

Lemma

Let $f(z) = \frac{1}{2}\|z - y_t\|^2$ *for* $y_t \in \hat{\Omega}$, *so that* $\beta = 1$ *and* $f(z^*) = 0$, *and let* $\eta_j = \frac{1}{j}$. *Then* $z_M = \frac{1}{M}\sum_{j=1}^{M} v_j$ *and*

$$\frac{1}{2}\|z_M - y_t\|^2 = f(z_M) \leq \frac{R^2}{2M}.$$

Finding Directions ϕ_j with Frank-Wolfe

Frank-Wolfe to minimize any convex f over $\hat{\Omega}$ (with $z_0 = 0$):

$$\begin{aligned} \phi_j &= \nabla f(z_{j-1}) = z_{j-1} - y_t \\ v_j &= \hat{\Psi}(\pi_{\phi_j}) \in \arg\min_{v \in \hat{\Omega}} \langle v, \phi_j \rangle \\ z_j &= (1 - \eta_j) z_{j-1} + \eta_j v_j \end{aligned}$$

Lemma

Let $f(z) = \frac{1}{2}\|z - y_t\|^2$ for $y_t \in \hat{\Omega}$, so that $\beta = 1$ and $f(z^*) = 0$, and let $\eta_j = \frac{1}{j}$. Then $z_M = \frac{1}{M}\sum_{j=1}^{M} \hat{\Psi}(\pi_{\phi_j})$ and

$$\frac{1}{2}\|z_M - y_t\|^2 = f(z_M) \leq \frac{R^2}{2M}.$$

Let π_t play π_{ϕ_j} for $j \sim \mathrm{Unif}\{1, \ldots, M\}$. Then $\hat{\Psi}(\pi_t) = z_M$, so

$$\|\hat{\Psi}(\pi_t) - y_t\| \leq \frac{R}{\sqrt{M}}.$$

- Each Frank-Wolfe iteration costs N oracle calls, so MN calls per round t

Finding Directions ϕ_j with Frank-Wolfe

Let π_t play π_{ϕ_j} for $j \sim \mathrm{Unif}\{1, \ldots, M\}$. Then $\hat{\Psi}(\pi_t) = z_M$, so

$$\|\hat{\Psi}(\pi_t) - y_t\| \leq \frac{R}{\sqrt{M}}.$$

- Each Frank-Wolfe iteration costs N oracle calls, so MN calls per round t

Approximate play with a mixture of linear policies:

$$|\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \|\hat{\Psi}(\pi_t) - y_t\| \cdot \|\theta_t\| \leq \frac{R\|\theta_t\|}{\sqrt{M}}$$

Finding Directions ϕ_j with Frank-Wolfe

Let π_t play π_{ϕ_j} for $j \sim \mathrm{Unif}\{1, \ldots, M\}$. Then $\hat{\Psi}(\pi_t) = z_M$, so

$$\|\hat{\Psi}(\pi_t) - y_t\| \leq \frac{R}{\sqrt{M}}.$$

- Each Frank-Wolfe iteration costs N oracle calls, so MN calls per round t

Approximate play with a mixture of linear policies:

$$|\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \|\hat{\Psi}(\pi_t) - y_t\| \cdot \|\theta_t\| \leq \frac{R\|\theta_t\|}{\sqrt{M}}$$

- Without loss of generality, can assume $\|a\| \leq 1$ for all $a \in \mathcal{A}$ a.s., and $\|\theta_t\| \leq \sqrt{d}$ [Wei et al., 2021]
- Then $R = \mathrm{diam}(\hat{\Omega}) \leq \max_i \mathrm{diam}(\tilde{\mathcal{A}}_i) \leq 2$, so, taking $M = N$,

$$|\langle \hat{\Psi}(\pi_t), \theta_t \rangle - \langle y_t, \theta_t \rangle| \leq \gamma_2 := 2\sqrt{\frac{d}{N}} \leq \gamma_1$$

with N^2 oracle calls per round t

Putting It All Together

Theorem

With $O(TN^2)$ simulator calls in total, we obtain

$$R_T(\pi) = \tilde{O}\Big(d\sqrt{T} + d^{3/2}T\sqrt{\frac{\log(NKT)}{N}}\Big).$$

Proof.

- α-misspecification-robust linear bandit algorithm with misspecification ϵ gives linear bandit regret bounded by

$$\tilde{O}\big(d\sqrt{T} + \alpha\sqrt{d}\epsilon T\big)$$

  with $\alpha = \sqrt{d}$ for an efficient algorithm.
- In the reduction to linear bandits we obtained misspecification $\epsilon = \gamma_1 + \gamma_2 \leq 2\gamma_1$ for $\gamma_1 = O\big(\sqrt{d\log(NKT)/N}\big)$ (issue 1)
- We pay an additive $2\gamma_1 T + \gamma_2 T \leq 3\gamma_1 T$ overhead because the linear bandit gives regret bound w.r.t. $\hat{\Omega}$ instead of Ω (issue 2)

Putting It All Together

Theorem

With $O(TN^2)$ simulator calls in total, we obtain

$$R_T(\pi) = \tilde{O}\left(d\sqrt{T} + d^{3/2}T\sqrt{\frac{\log(NKT)}{N}}\right).$$

With simulator access: Take N large enough to make the second term negligible:

$$R_T(\pi) = \tilde{O}(d\sqrt{T})$$

Putting It All Together

Theorem

With $O(TN^2)$ simulator calls in total, we obtain

$$R_T(\pi) = \tilde{O}\left(d\sqrt{T} + d^{3/2}T\sqrt{\frac{\log(NKT)}{N}}\right).$$

With simulator access: Take N large enough to make the second term negligible:

$$R_T(\pi) = \tilde{O}(d\sqrt{T})$$

Without simulator access:

- Run in epochs of lengths 2^i for $i = 0, 1, 2, 3, \ldots$
- In epoch i use $N = \Theta(2^i)$ samples from all previous epochs to construct $\hat{\Omega}$

$$R_T(\pi) = \tilde{O}\left(d\sqrt{T} + d^{3/2}\sqrt{T\log K}\right)$$

- Since $N \leq T$: $\mathrm{poly}(d, T)$ time plus $\mathrm{poly}(d, T)$ oracle calls in total

Conclusion

Highlights:

- First algorithm for this setting that handles combinatorial action sets efficiently, using only a linear optimization oracle for each action set
- Efficient reduction from contextual to (misspecified) non-contextual linear bandits
- Handle resulting misspecification in linear bandit algorithm

Open Questions:

- Improve regret to $\tilde{O}(d\sqrt{T})$ with polynomial-time algorithm without a simulator?
- Obtain a first-order bound $\text{poly}(d)\sqrt{L^*}$ without a simulator?
- Improve first-order bound to $\tilde{O}(\sqrt{pKL^*})$ in initial setting?

References

An Improved Algorithm for Adversarial Linear Contextual Bandits via Reduction

- Published version: **NeurIPS 2025**
- Strengthened newer version 2026 (this talk): **arXiv:2508.11931**

Upcoming: in-probability results

Probabilistic Guarantees for Adversarial Linear Contextual Bandits
Guyon Phoelich, Julia Olkhovskaya, Tim van Erven

References I

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–358, 2015.

Y. Dai, H. Luo, C.-Y. Wei, and J. Zimmert. Refined regret for adversarial MDPs with linear function approximation. In *International Conference on Machine Learning*, pages 6726–6759. PMLR, 2023.

O. A. Hanna, L. Yang, and C. Fragouli. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1791–1821. PMLR, 2023.

S. Ito, S. Hirahara, T. Soma, and Y. Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33:2028–2038, 2020.

C.-W. Lee, H. Luo, C.-Y. Wei, and M. Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs. *Advances in neural information processing systems*, 33:15522–15533, 2020.

H. Liu, C.-Y. Wei, and J. Zimmert. Bypassing the simulator: Near-optimal adversarial linear contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2023.

H. Liu, A. Tajdini, A. Wagenmaker, and C.-Y. Wei. Corruption-robust linear bandits: Minimax optimality and gap-dependent misspecification. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

References II

H. Liu, C.-Y. Wei, and J. Zimmert. Towards optimal regret in adversarial linear MDPs with bandit feedback. In *The Twelfth International Conference on Learning Representations*, 2024b.

G. Neu and J. Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3049–3068. PMLR, 09–12 Jul 2020. URL `https://proceedings.mlr.press/v125/neu20b.html`.

G. Neu and M. Valko. Online combinatorial optimization with stochastic decision sets and adversarial losses. *Advances in Neural Information Processing Systems*, 27, 2014.

J. Olkhovskaya, J. Mayo, T. van Erven, G. Neu, and C.-Y. Wei. First-and second-order bounds for adversarial linear contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2023.

C.-Y. Wei, M. Jafarnia Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward MDPs with linear function approximation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

J. Zimmert and T. Lattimore. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Conference on Learning Theory*, pages 3285–3312. PMLR, 2022.