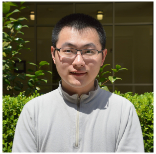
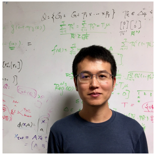
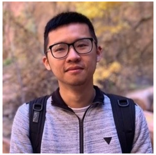
Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs

Mengxiao Zhang



joint with **Chung-Wei Lee**, **Haipeng Luo** and **Chen-Yu Wei**





USC University of Southern California

USC Viterbi School of Engineering

Adversarial Bandits

Adversarial Bandits

Multi-Armed Bandits (MAB)

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss
- goal: be competitive with the best fixed arm

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss
- goal: be competitive with the best fixed arm

Linear Bandits (LB)

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss
- goal: be competitive with the best fixed arm

Linear Bandits (LB)

- a convex action set Ω available

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss
- goal: be competitive with the best fixed arm

Linear Bandits (LB)

- a convex action set Ω available
- adversary decides the loss vectors

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss
- goal: be competitive with the best fixed arm

Linear Bandits (LB)

- a convex action set Ω available
- adversary decides the loss vectors
- learner sequentially chooses an action from Ω and observe its loss, which is its inner product with the loss vector

Adversarial Bandits

Multi-Armed Bandits (MAB)

- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss
- goal: be competitive with the best fixed arm

Linear Bandits (LB)

- a convex action set Ω available
- adversary decides the loss vectors
- learner sequentially chooses an action from Ω and observe its loss, which is its inner product with the loss vector
- goal: be competitive with the best fixed action in Ω

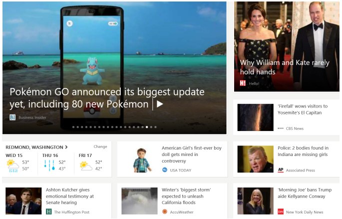
Adversarial Bandits

Multi-Armed Bandits (MAB)

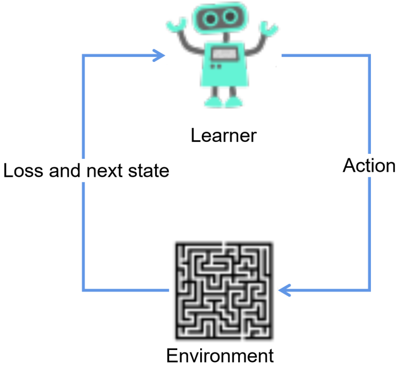
- d arms/actions available
- adversary decides the losses for each arm
- learner sequentially pull an arm and observes its loss
- goal: be competitive with the best fixed arm

Linear Bandits (LB) (e.g. news recommendation)

- a convex action set Ω available
- adversary decides the loss vectors
- learner sequentially chooses an action from Ω and observe its loss, which is its inner product with the loss vector
- goal: be competitive with the best fixed action in Ω

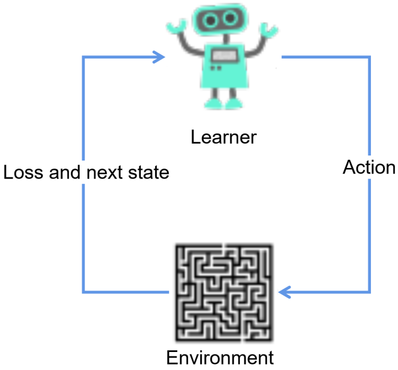


Adversarial Markov Decision Process with Bandit Feedback



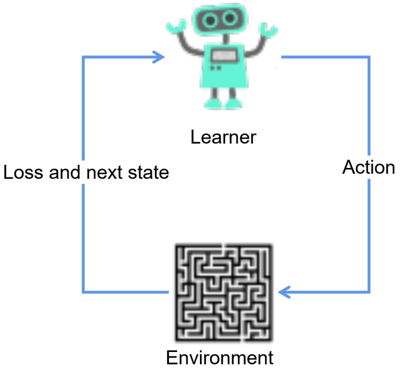
Adversarial Markov Decision Process with Bandit Feedback

- episodic finite time horizon, unknown transition



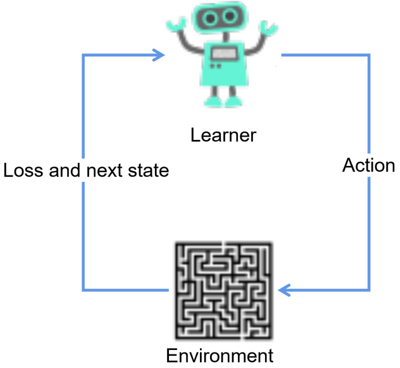
Adversarial Markov Decision Process with Bandit Feedback

- episodic finite time horizon, unknown transition
- loss is adversarially chosen by the environment



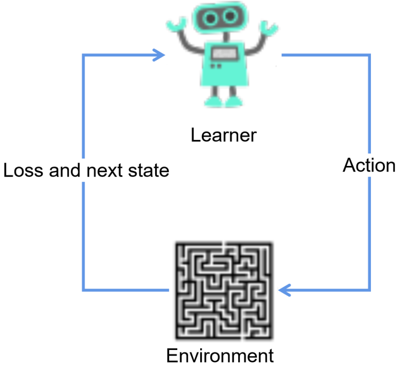
Adversarial Markov Decision Process with Bandit Feedback

- episodic finite time horizon, unknown transition
- loss is adversarially chosen by the environment
- learner sequentially chooses an action according to its current state, observe its loss, and transits to the next state



Adversarial Markov Decision Process with Bandit Feedback

- episodic finite time horizon, unknown transition
- loss is adversarially chosen by the environment
- learner sequentially chooses an action according to its current state, observe its loss, and transits to the next state
- goal: be competitive with the best fixed policy



From Expected Regret to High Probability Regret

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- $\textsc{Exp3}$: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- Exp3: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]
- SCRiBLe: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB [AHR12]

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- EXP3: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]
- SCRIBLE: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB [AHR12]

High probability regret bounds:

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- EXP3: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]
- SCRIBLE: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB [AHR12]

High probability regret bounds:

- EXP3.P, EXP3-IX: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB [ACFS02,N15]

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- EXP3: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]
- SCRIBLE: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB [AHR12]

High probability regret bounds:

- EXP3.P, EXP3-IX: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB [ACFS02,N15]
- GEOMETRICHEDGE.P: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB, an inefficient algorithm [BDHKRT08]

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- EXP3: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]
- SCRIBLE: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB [AHR12]

High probability regret bounds:

- EXP3.P, EXP3-IX: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB [ACFS02,N15]
- GEOMETRICHEDGE.P: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB, an inefficient algorithm [BDHKRT08]
- COMPEXP: $\tilde{\mathcal{O}}(T^{2/3})$ for LB, an efficient algorithm [BP16]

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- Exp3: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]
- SCRiBLe: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB [AHR12]

High probability regret bounds:

- Exp3.P, Exp3-IX: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB [ACFS02,N15]
- GeometricHedge.P: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB, an inefficient algorithm [BDHKRT08]
- CompEXP: $\tilde{\mathcal{O}}(T^{2/3})$ for LB, an efficient algorithm [BP16]
- $\tilde{\mathcal{O}}(\sqrt{T})$ high probability regret for LB under a set of conditions [AR09]

From Expected Regret to High Probability Regret

Expected regret bounds for bandit problems:

- EXP3: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB. [ACFS02]
- SCRIBLE: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB [AHR12]

High probability regret bounds:

- EXP3.P, EXP3-IX: $\tilde{\mathcal{O}}(\sqrt{T})$ for MAB [ACFS02,N15]
- GEOMETRICHEDGE.P: $\tilde{\mathcal{O}}(\sqrt{T})$ for LB, an inefficient algorithm [BDHKRT08]
- COMPEXP: $\tilde{\mathcal{O}}(T^{2/3})$ for LB, an efficient algorithm [BP16]
- $\tilde{\mathcal{O}}(\sqrt{T})$ high probability regret for LB under a set of conditions [AR09]

Open Problem (BDHKRT08, BP16, AR09): Whether $\tilde{\mathcal{O}}(\sqrt{T})$ high probability regret bound is achievable efficiently for general LB?

From Minimax Regret to Data-Dependent Regert Bounds

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

Data-dependent regret bounds: much better than minimax regret for "easy" instances

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

Data-dependent regret bounds: much better than minimax regret for "easy" instances

- small-loss bound: replace T by the loss of the best action in hindsight

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

Data-dependent regret bounds: much better than minimax regret for "easy" instances

- small-loss bound: replace T by the loss of the best action in hindsight
- variation bound: replace T by the variance of the loss vector

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

Data-dependent regret bounds: much better than minimax regret for "easy" instances

- small-loss bound: replace T by the loss of the best action in hindsight
- variation bound: replace T by the variance of the loss vector

Near-optimal small-loss high probability regret bounds:

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

Data-dependent regret bounds: much better than minimax regret for "easy" instances

- small-loss bound: replace T by the loss of the best action in hindsight
- variation bound: replace T by the variance of the loss vector

Near-optimal small-loss high probability regret bounds:

- achievable for MAB [N15]

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

Data-dependent regret bounds: much better than minimax regret for "easy" instances

- small-loss bound: replace T by the loss of the best action in hindsight
- variation bound: replace T by the variance of the loss vector

Near-optimal small-loss high probability regret bounds:

- achievable for MAB [N15]
- achievable for more general bandit problems with graph feedback. [LTS19]

From Minimax Regret to Data-Dependent Regert Bounds

Minimax regret bounds:

- both MAB and LB: $\tilde{\Theta}(\sqrt{T})$ [ACFS02,DP08]

Data-dependent regret bounds: much better than minimax regret for "easy" instances

- small-loss bound: replace T by the loss of the best action in hindsight
- variation bound: replace T by the variance of the loss vector

Near-optimal small-loss high probability regret bounds:

- achievable for MAB [N15]
- achievable for more general bandit problems with graph feedback. [LTS19]

> Open Problem (N15): Whether data-dependent high probability regret bound is achievable efficiently for general bandit problems?

Open Problem (BDHKRT08, BP16, AR09):

Near-optimal efficient + high-probability bound for LB

Open Problem (N15):

Near-optimal data-dependent + high-probability bound for bandits

This work:

Open Problem (BDHKRT08, BP16, AR09):

Near-optimal efficient + high-probability bound for LB

Open Problem (N15):

Near-optimal data-dependent + high-probability bound for bandits

This work:

- Near-optimal efficient + data-dependent + high-probability bound for LB

Open Problem (BDHKRT08, BP16, AR09):

Near-optimal efficient + high-probability bound for LB

Open Problem (N15):

Near-optimal data-dependent + high-probability bound for bandits

This work:

- Near-optimal efficient + data-dependent + high-probability bound for LB
- also achieves small-loss + high-probability regret bounds for adversarial episodic Markov Decision Process with bandit feedback and unknown transition function

Open Problem (BDHKRT08, BP16, AR09):

Near-optimal efficient + high-probability bound for LB

Open Problem (N15):

Near-optimal data-dependent + high-probability bound for bandits

This work:

- Near-optimal efficient + data-dependent + high-probability bound for LB
- also achieves small-loss + high-probability regret bounds for adversarial episodic Markov Decision Process with bandit feedback and unknown transition function
- uses unbiased estimators and relies on an increasing learning rate schedule, together with a strengthened Freedman's inequality and normal barriers.

High Probability Near-Optimal Data-Dependent Bound for LB

Setup

A convex set Ω is given to the learner
For $t = 1, \ldots, T$:

Setup

A convex set Ω is given to the learner

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

- the adversary decides a loss vector $\ell_t \in \mathbb{R}^d$

Setup

A convex set Ω is given to the learner
For $t = 1, \ldots, T$:

- the adversary decides a loss vector $\ell_t \in \mathbb{R}^d$
- the learner picks an **arm** $\widetilde{w}_t \in \Omega$ and incurs **loss** $\langle \widetilde{w}_t, \ell_t \rangle$

Setup

A convex set Ω is given to the learner
For $t = 1, \ldots, T$:

- the adversary decides a loss vector $\ell_t \in \mathbb{R}^d$
- the learner picks an **arm** $\widetilde{w}_t \in \Omega$ and incurs **loss** $\langle \widetilde{w}_t, \ell_t \rangle$
- the learner observes her **loss** $\langle \widetilde{w}_t, \ell_t \rangle$

Setup

A convex set Ω is given to the learner

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

- the adversary decides a loss vector $\ell_t \in \mathbb{R}^d$
- the learner picks an **arm** $\widetilde{w}_t \in \Omega$ and incurs **loss** $\langle \widetilde{w}_t, \ell_t \rangle$
- the learner observes her **loss** $\langle \widetilde{w}_t, \ell_t \rangle$

Goal: to be competitive w.r.t. a fixed action

$$\mathrm{Reg} \triangleq \sum_{t=1}^{T} \langle \widetilde{w}_t, \ell_t \rangle - \min_{u \in \Omega} \sum_{t=1}^{T} \langle u, \ell_t \rangle$$

Setup

A convex set Ω is given to the learner

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

- the adversary decides a loss vector $\ell_t \in \mathbb{R}^d$
- the learner picks an **arm** $\widetilde{w}_t \in \Omega$ and incurs **loss** $\langle \widetilde{w}_t, \ell_t \rangle$
- the learner observes her **loss** $\langle \widetilde{w}_t, \ell_t \rangle$

Goal: to be competitive w.r.t. a fixed action

$$\boxed{\mathrm{Reg} \triangleq \sum_{t=1}^{T} \langle \widetilde{w}_t, \ell_t \rangle - \min_{u \in \Omega} \sum_{t=1}^{T} \langle u, \ell_t \rangle}$$

Assumption: $|\langle w, \ell_t \rangle| \leq 1$ for all $w \in \Omega$

Recall: SCRIBLE

For each round $t = 1, 2, \ldots, T$

Recall: SCRIBLE

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$

Recall: SCRiBLe

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$
  - ψ: ν-self-concordant barrier over Ω
  - D_ψ: Bregman divergence with respect to ψ

Recall: SCRIBLE

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$
    - ψ: ν-self-concordant barrier over Ω
    - D_ψ: Bregman divergence with respect to ψ
- choose $\widetilde{w}_t$ from Dikin ellipsoid $\|\widetilde{w}_t - w_t\|_{H_t} = 1$ and observe $\langle \widetilde{w}_t, \ell_t \rangle$

Recall: SCRIBLE

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$
  - ψ: ν-self-concordant barrier over Ω
  - D_ψ: Bregman divergence with respect to ψ
- choose $\widetilde{w}_t$ from Dikin ellipsoid $\|\widetilde{w}_t - w_t\|_{H_t} = 1$ and observe $\langle \widetilde{w}_t, \ell_t \rangle$
  - $H_t = \nabla^2 \psi(w_t)$

Recall: SCRIBLE

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$
  - ψ: ν-self-concordant barrier over Ω
  - D_ψ: Bregman divergence with respect to ψ
- choose $\widetilde{w}_t$ from Dikin ellipsoid $\|\widetilde{w}_t - w_t\|_{H_t} = 1$ and observe $\langle \widetilde{w}_t, \ell_t \rangle$
  - $H_t = \nabla^2 \psi(w_t)$
- construct unbiased loss estimator $\widehat{\ell}_t$

Recall: SCRIBLE

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$
  - ψ: ν-self-concordant barrier over Ω
  - D_ψ: Bregman divergence with respect to ψ
- choose $\widetilde{w}_t$ from Dikin ellipsoid $\|\widetilde{w}_t - w_t\|_{H_t} = 1$ and observe $\langle \widetilde{w}_t, \ell_t \rangle$
  - $H_t = \nabla^2 \psi(w_t)$
- construct unbiased loss estimator $\widehat{\ell}_t$

Key challenge in obtaining h.p. bound:

control the variance of $\left\langle w_t - u, \widehat{\ell}_t \right\rangle$

Recall: SCRiBLe

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \mathrm{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$
  - ψ: ν-self-concordant barrier over Ω
  - D_ψ: Bregman divergence with respect to ψ
- choose $\widetilde{w}_t$ from Dikin ellipsoid $\|\widetilde{w}_t - w_t\|_{H_t} = 1$ and observe $\langle \widetilde{w}_t, \ell_t \rangle$
  - $H_t = \nabla^2 \psi(w_t)$
- construct unbiased loss estimator $\widehat{\ell}_t$

Key challenge in obtaining h.p. bound:

control the variance of $\left\langle w_t - u, \widehat{\ell}_t \right\rangle \Longrightarrow$ control $\|u\|_{H_t}$ and $\|w_t\|_{H_t}$

Recall: SCRIBLE

For each round $t = 1, 2, \ldots, T$

- compute $w_t = \operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_{t-1} \right\rangle + D_\psi(w, w_{t-1}) \right\}$
  - ψ: ν-self-concordant barrier over Ω
  - D_ψ: Bregman divergence with respect to ψ
- choose $\widetilde{w}_t$ from Dikin ellipsoid $\|\widetilde{w}_t - w_t\|_{H_t} = 1$ and observe $\langle \widetilde{w}_t, \ell_t \rangle$
  - $H_t = \nabla^2 \psi(w_t)$
- construct unbiased loss estimator $\widehat{\ell}_t$

Key challenge in obtaining h.p. bound:

$$\text{control the variance of } \left\langle w_t - u, \widehat{\ell}_t \right\rangle \Longrightarrow \text{control } \|u\|_{H_t} \text{ and } \|w_t\|_{H_t}$$

A strengthened Freedman's inequality is needed as classic Freedman's inequality depends on the *fixed* upper bound for $\left\langle w_t - u, \widehat{\ell}_t \right\rangle$

First Challenge: Control $\max_{t\in[T]} \|w_t\|_{H_t}$

First Challenge: Control $\max_{t\in[T]} \|w_t\|_{H_t}$

- θ-normal barriers ψ on a proper cone K:

First Challenge: Control $\max_{t\in[T]} \|w_t\|_{H_t}$

- θ-normal barriers ψ on a proper cone K:
  - self-concordant with domain int K

First Challenge: Control $\max_{t\in[T]} \|w_t\|_{H_t}$

- θ-normal barriers ψ on a proper cone K:
  - self-concordant with domain int K
  - $\psi(tx) = \psi(x) - \theta \ln(t), \forall x \in \text{int } K, t > 0$

First Challenge: Control $\max_{t \in [T]} \|w_t\|_{H_t}$

- θ-normal barriers ψ on a proper cone K:
  - self-concordant with domain int K
  - $\psi(tx) = \psi(x) - \theta \ln(t), \forall x \in$ int $K, t > 0$
- if ψ is also a θ-normal barrier:

$$\|w_t\|_{H_t} \leq \sqrt{\theta}$$

First Challenge: Control $\max_{t\in[T]} \|w_t\|_{H_t}$

- θ-normal barriers ψ on a proper cone K:
  - self-concordant with domain int K
  - $\psi(tx) = \psi(x) - \theta \ln(t), \forall x \in$ int $K, t > 0$
- if ψ is also a θ-normal barrier:
$$\|w_t\|_{H_t} \le \sqrt{\theta}$$
- however, normal barriers are only defined on cones instead of general convex bodies

First Challenge: Control $\max_{t\in[T]} \|w_t\|_{H_t}$

- θ-normal barriers ψ on a proper cone K:
  - self-concordant with domain int K
  - $\psi(tx) = \psi(x) - \theta \ln(t), \forall x \in \text{int } K, t > 0$
- if ψ is also a θ-normal barrier:
$$\|w_t\|_{H_t} \leq \sqrt{\theta}$$
- however, normal barriers are only defined on cones instead of general convex bodies
- solution: lifting the problem from $\mathbb{R}^d$ to $\mathbb{R}^{d+1}$!

Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$

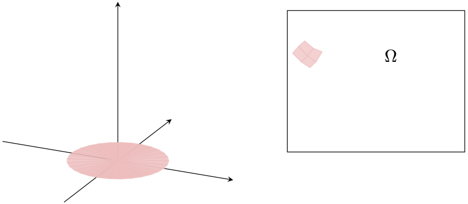


Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$
  $\Rightarrow$ lifted to $\mathbb{R}^{d+1}$: $\boldsymbol{\Omega} = (\Omega, 1)$

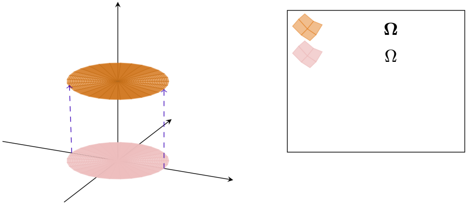


Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$
  $\Rightarrow$ lifted to $\mathbb{R}^{d+1}$: $\boldsymbol{\Omega} = (\Omega, 1)$
- construct the conic hull of $\boldsymbol{\Omega}$

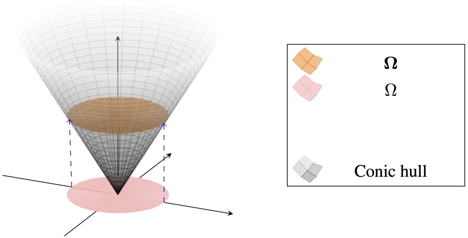


Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$
  $\Rightarrow$ lifted to $\mathbb{R}^{d+1}$: $\boldsymbol{\Omega} = (\Omega, 1)$
- construct the conic hull of $\boldsymbol{\Omega}$
- lift the point $w \in \Omega$ to $\boldsymbol{w} = (w, 1) \in \boldsymbol{\Omega}$

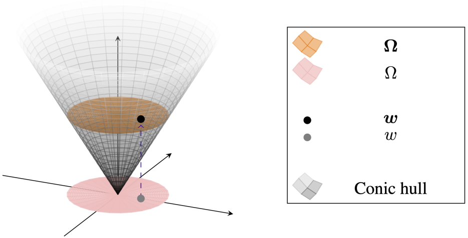


Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$
  $\Rightarrow$ lifted to $\mathbb{R}^{d+1}$: $\boldsymbol{\Omega} = (\Omega, 1)$
- construct the conic hull of $\boldsymbol{\Omega}$
- lift the point $w \in \Omega$ to $\boldsymbol{w} = (w, 1) \in \boldsymbol{\Omega}$
- construct the Dikin ellipsoid with respect to $\boldsymbol{w}$ according to a normal barrier Ψ

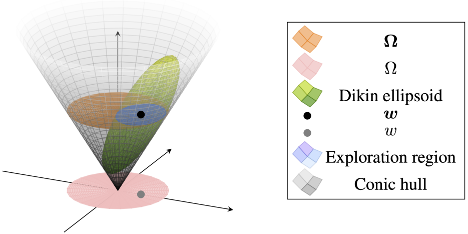


Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$
  $\Rightarrow$ lifted to $\mathbb{R}^{d+1}$: $\boldsymbol{\Omega} = (\Omega, 1)$
- construct the conic hull of $\boldsymbol{\Omega}$
- lift the point $w \in \Omega$ to $\boldsymbol{w} = (w, 1) \in \boldsymbol{\Omega}$
- construct the Dikin ellipsoid with respect to $\boldsymbol{w}$ according to a normal barrier Ψ
  - any normal barrier Ψ is applicable here

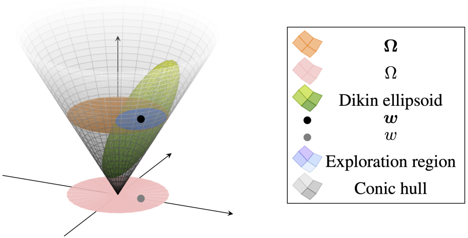


Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$
  $\Rightarrow$ lifted to $\mathbb{R}^{d+1}$: $\boldsymbol{\Omega} = (\Omega, 1)$
- construct the conic hull of $\boldsymbol{\Omega}$
- lift the point $w \in \Omega$ to $\boldsymbol{w} = (w, 1) \in \boldsymbol{\Omega}$
- construct the Dikin ellipsoid with respect to $\boldsymbol{w}$ according to a normal barrier Ψ
  - any normal barrier Ψ is applicable here
  - a natural construction of Ψ from a self-concordant barrier ψ of Ω: $\Psi(w, b) = 400(\psi(\frac{w}{b}) - 2\nu \ln b)$

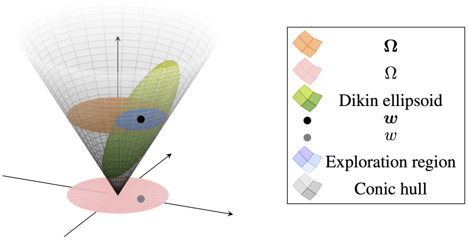
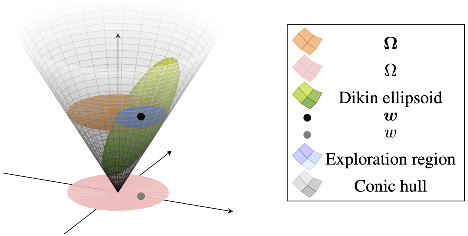


Illustration of lifting

- feasible set $\Omega \subseteq \mathbb{R}^d$
  $\Rightarrow$ lifted to $\mathbb{R}^{d+1}$: $\boldsymbol{\Omega} = (\Omega, 1)$
- construct the conic hull of $\boldsymbol{\Omega}$
- lift the point $w \in \Omega$ to $\boldsymbol{w} = (w, 1) \in \boldsymbol{\Omega}$
- construct the Dikin ellipsoid with respect to $\boldsymbol{w}$ according to a normal barrier Ψ
  - any normal barrier Ψ is applicable here
  - a natural construction of Ψ from a self-concordant barrier ψ of Ω: $\Psi(w, b) = 400(\psi(\frac{w}{b}) - 2\nu \ln b)$
- sample from the boundary of the intersection of the Dikin ellipsoid and $\boldsymbol{\Omega}$



Comparison with SCRIBLE

Comparison with SCRIBLE

- SCRIBLE update: $\operatorname{argmin}_{w\in\Omega}\left\{\left\langle w, \widehat{\ell}_t\right\rangle + D_\psi(w, w_t)\right\}$

Comparison with SCRIBLE

- SCRIBLE update: $\mathrm{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_t \right\rangle + D_\psi(w, w_t) \right\}$
- lifted problem: $\mathrm{argmin}_{\boldsymbol{w} \in \boldsymbol{\Omega}} \left\{ \left\langle \boldsymbol{w}, \widehat{\boldsymbol{\ell}}_t \right\rangle + D_\Psi(\boldsymbol{w}, \boldsymbol{w}_t) \right\}$

Comparison with SCRIBLE

- SCRIBLE update: $\operatorname{argmin}_{w \in \Omega} \left\{ \left\langle w, \widehat{\ell}_t \right\rangle + D_\psi(w, w_t) \right\}$
- lifted problem: $\operatorname{argmin}_{\boldsymbol{w} \in \boldsymbol{\Omega}} \left\{ \left\langle \boldsymbol{w}, \widehat{\boldsymbol{\ell}}_t \right\rangle + D_\Psi(\boldsymbol{w}, \boldsymbol{w}_t) \right\}$
- Observe that $\Psi(w, b) = 400(\psi(\frac{w}{b}) - 2\nu \ln b)$

$$\Psi(\boldsymbol{w}) = \Psi(w, 1) = 400\psi(w), \boldsymbol{w} \in \boldsymbol{\Omega}$$

Comparison with SCRIBLE

- SCRIBLE update: $\operatorname{argmin}_{w\in\Omega}\left\{\left\langle w, \widehat{\ell}_t\right\rangle + D_\psi(w, w_t)\right\}$
- lifted problem: $\operatorname{argmin}_{\boldsymbol{w}\in\boldsymbol{\Omega}}\left\{\left\langle \boldsymbol{w}, \widehat{\boldsymbol{\ell}}_t\right\rangle + D_\Psi(\boldsymbol{w}, \boldsymbol{w}_t)\right\}$
- Observe that $\Psi(w, b) = 400(\psi(\frac{w}{b}) - 2\nu \ln b)$

$$\Psi(\boldsymbol{w}) = \Psi(w, 1) = 400\psi(w), \boldsymbol{w} \in \boldsymbol{\Omega}$$

- SCRIBLE with a new sampling scheme!

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]
- **The effect of increasing learning rate at time** t:

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]
- **The effect of increasing learning rate at time t:**

  increase $\eta \rightarrow (1+\epsilon)\eta$

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]
- **The effect of increasing learning rate at time t**:

$$\text{increase } \eta \to (1+\epsilon)\eta \quad \Longrightarrow \quad \text{create negative regret } \frac{-\epsilon}{(1+\epsilon)\eta} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t)$$

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]
- **The effect of increasing learning rate at time t:**

$$\text{increase } \eta \rightarrow (1+\epsilon)\eta \quad \Longrightarrow \quad \text{create negative regret } \tfrac{-\epsilon}{(1+\epsilon)\eta} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t)$$

$$\text{with normal barriers:} \quad \tfrac{\epsilon}{(1+\epsilon)} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t) \gtrsim \|\boldsymbol{u}\|_{\boldsymbol{H}_t} \qquad \text{(cancelling variance!)}$$

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]
- **The effect of increasing learning rate at time t**:

$$\text{increase } \eta \to (1+\epsilon)\eta \quad \Longrightarrow \quad \text{create negative regret } \tfrac{-\epsilon}{(1+\epsilon)\eta} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t)$$

$$\text{with normal barriers:} \quad \tfrac{\epsilon}{(1+\epsilon)} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t) \gtrsim \|\boldsymbol{u}\|_{\boldsymbol{H}_t} \qquad \text{(cancelling variance!)}$$

- when to increase learning rate?

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]
- **The effect of increasing learning rate at time** t:

$$\text{increase } \eta \to (1+\epsilon)\eta \quad \Longrightarrow \quad \text{create negative regret } \tfrac{-\epsilon}{(1+\epsilon)\eta} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t)$$

$$\text{with normal barriers:} \quad \tfrac{\epsilon}{(1+\epsilon)} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t) \gtrsim \|\boldsymbol{u}\|_{\boldsymbol{H}_t} \qquad \text{(cancelling variance!)}$$

- when to increase learning rate? when $\boldsymbol{H}_t$ is "large"

Second Challenge: Control $\max_{t\in[T]} \|u\|_{H_t}$

- **Idea**: increasing learning rate (which creates negative regret!). e.g.,
  - combining algorithms with different regret bounds under different settings [ALNE17]
  - deriving small-loss and other data-dependent bounds [WL18,LLZ20]
- **The effect of increasing learning rate at time t:**

$$\text{increase } \eta \to (1+\epsilon)\eta \quad \Longrightarrow \quad \text{create negative regret } \tfrac{-\epsilon}{(1+\epsilon)\eta} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t)$$

$$\text{with normal barriers:} \quad \tfrac{\epsilon}{(1+\epsilon)} D_\Psi(\boldsymbol{u}, \boldsymbol{w}_t) \gtrsim \|\boldsymbol{u}\|_{\boldsymbol{H}_t} \qquad \text{(cancelling variance!)}$$

- when to increase learning rate? when $\boldsymbol{H}_t$ is "large"

  $\lambda_{\max}\left(\boldsymbol{H}_t - \sum_{\tau\in\mathcal{S}} \boldsymbol{H}_\tau\right) > 0$, where $\mathcal{S}$ is the set of previous time steps at which we increase learning rate

Regret Bounds

With probability at least $1-\delta$

$$\mathrm{Reg} = \begin{cases} \tilde{\mathcal{O}}\left(d^2\nu\sqrt{T\ln\frac{1}{\delta}} + d^2\nu\ln\frac{1}{\delta}\right), & \text{against an oblivious adversary;} \\ \tilde{\mathcal{O}}\left(d^2\nu\sqrt{dT\ln\frac{1}{\delta}} + d^3\nu\ln\frac{1}{\delta}\right), & \text{against an adaptive adversary} \end{cases}$$

Regret Bounds

With probability at least $1-\delta$

$$\text{Reg} = \begin{cases} \tilde{\mathcal{O}}\left(d^2\nu\sqrt{T\ln\frac{1}{\delta}} + d^2\nu\ln\frac{1}{\delta}\right), & \text{against an oblivious adversary;} \\ \tilde{\mathcal{O}}\left(d^2\nu\sqrt{dT\ln\frac{1}{\delta}} + d^3\nu\ln\frac{1}{\delta}\right), & \text{against an adaptive adversary} \end{cases}$$

if $\langle w, \ell_t\rangle \geq 0$ for all $w \in \Omega$, $t \in [T]$, then T can be replaced by $L^\star = \min_{u\in\Omega}\sum_{t=1}^T \langle u, \ell_t\rangle$, or other data-dependent values with optimistic estimators

High Probability Small-Loss Bound for Markov Decision Process

Achieving High Probability Small-Loss Bound

With the help of increasing learning rate, we obtain the first high probability small-loss regret bound for adversarial MDP, improving the result of [JJLSY19]

Achieving High Probability Small-Loss Bound

With the help of increasing learning rate, we obtain the first high probability small-loss regret bound for adversarial MDP, improving the result of [JJLSY19]

With high probability, $\mathrm{Reg} = \tilde{\mathcal{O}}\left(\sqrt{L^\star}\right)$, for both oblivious and adaptive adversaries

Achieving High Probability Small-Loss Bound

With the help of increasing learning rate, we obtain the first high probability small-loss regret bound for adversarial MDP, improving the result of [JJLSY19]

> With high probability, $\mathrm{Reg} = \tilde{\mathcal{O}}\left(\sqrt{L^\star}\right)$, for both oblivious and adaptive adversaries

- clipping technique and implicit exploration may not be directly applicable here to obtain small-loss bound

Achieving High Probability Small-Loss Bound

With the help of increasing learning rate, we obtain the first high probability small-loss regret bound for adversarial MDP, improving the result of [JJLSY19]

With high probability, $\mathrm{Reg} = \tilde{\mathcal{O}}\left(\sqrt{L^\star}\right)$, for both oblivious and adaptive adversaries

- clipping technique and implicit exploration may not be directly applicable here to obtain small-loss bound
- not clear how to obtain other data-dependent bounds as there are several terms in the regret that are naturally only related to $L^\star$

Summary

This work:

Summary

This work:

- **Linear bandits**: first efficient algorithm with high probability data-dependent bound for general feasible sets.
  techniques:
  - lifting
  - normal barrier
  - increasing learning rate

Summary

This work:

- **Linear bandits**: first efficient algorithm with high probability data-dependent bound for general feasible sets.
  techniques:
  - lifting
  - normal barrier
  - increasing learning rate
- **Adversarial MDP**: high probability small-loss regret bounds with bandit feedback and unknown transition

Summary

This work:

- **Linear bandits**: first efficient algorithm with high probability data-dependent bound for general feasible sets.
  techniques:
  - lifting
  - normal barrier
  - increasing learning rate
- **Adversarial MDP**: high probability small-loss regret bounds with bandit feedback and unknown transition

Open problems:

Summary

This work:

- **Linear bandits**: first efficient algorithm with high probability data-dependent bound for general feasible sets.
  techniques:
  - lifting
  - normal barrier
  - increasing learning rate
- **Adversarial MDP**: high probability small-loss regret bounds with bandit feedback and unknown transition

Open problems:

- **Linear bandits**: improving the dependence on d

Summary

This work:

- **Linear bandits**: first efficient algorithm with high probability data-dependent bound for general feasible sets.
  techniques:
  - lifting
  - normal barrier
  - increasing learning rate
- **Adversarial MDP**: high probability small-loss regret bounds with bandit feedback and unknown transition

Open problems:

- **Linear bandits**: improving the dependence on d
- **MDP**: other types of data-dependent bounds