Two recent lower bounds for interactive decision making

Yanjun Han (NYU Courant and CDS)

Joint work with:

Dylan Foster	Microsoft Research
Noah Golowich	MIT EECS
Jiantao Jiao	Berkeley EECS
Nived Rajaraman	Berkeley EECS
Kannan Ramchandran	Berkeley EECS

Math and Data Seminar
September 21, 2023

Interactive decision making

robotics

games

clinical systems

algorithm design

Examples:

- bandits
- reinforcement learning
- control
- online optimization
- dynamic pricing
- dynamic treatments

Aim of this talk

Characterize the optimal sample complexity/fundamental limits for interactive decision making problems.

The interactive model

Decision making with structured observations (DMSO)

At each round $t = 1, 2, \cdots, T$:

- learner chooses a decision $a_t \in \mathcal{A}$;
- nature reveals reward $r_t \in [0, 1]$ and observation $o_t \in \mathcal{O}$ (possibly empty).

The interactive model

Decision making with structured observations (DMSO)

At each round $t = 1, 2, \cdots, T$:

- learner chooses a decision $a_t \in \mathcal{A}$;
- nature reveals reward $r_t \in [0, 1]$ and observation $o_t \in \mathcal{O}$ (possibly empty).

Stochastic model:

- a given model class $\mathcal{M}$
- unknown true model $M^\star \in \mathcal{M}$
- $(r_t, o_t) \sim M^\star(a_t)$, with $\mathbb{E}[r_t \mid a_t = a] = r^{M^\star}(a)$
- for $M \in \mathcal{M}$, let $r_\star^M = \max_{a \in \mathcal{A}} r^M(a)$ be the maximum reward under $\mathcal{M}$
- learner's regret:

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(r_\star^{M^\star} - r^{M^\star}(a_t) \right)$$

DMSO examples

Multi-armed bandits:

- $\mathcal{A} = \{1, 2, \cdots, K\}$;
- $\mathcal{O} = \varnothing$;
- $\mathcal{M} =$ "all 1-subGaussian reward distributions"

DMSO examples

Multi-armed bandits:

- $\mathcal{A} = \{1, 2, \cdots, K\}$;
- $\mathcal{O} = \varnothing$;
- $\mathcal{M} =$ "all 1-subGaussian reward distributions"

Episodic reinforcement learning:

- $\mathcal{A} =$ a sequence of policies $(\pi_1, \cdots, \pi_H)$
- reward $r_t = \sum_{h=1}^{H} r_{t,h}$
- observation trajectory $o_t = \{(s_{t,1}, a_{t,1}, r_{t,1}), \cdots, (s_{t,H}, a_{t,H}, r_{t,H})\}$
- $\mathcal{M} =$ "a collection of transition and reward distributions"

Part I: Interactive two-point lower bound

Dylan Foster
Microsoft Research

Noah Golowich
MIT EECS

"Tight Guarantees for Interactive Decision Making with the Decision-Estimation Coefficient" (COLT 2023; arXiv: 2301.08215)

Decision-estimation coefficient (DEC)

DEC (Foster, Kakade, Qian, Rakhlin, 2021)

$$\mathrm{dec}_\gamma(\mathcal{M}, \overline{M}) = \inf_{p \in \Delta(\mathcal{A})} \sup_{M \in \mathcal{M}} \underbrace{\mathbb{E}_{a\sim p}[r_\star^M - r^M(a)]}_{\text{regret of decision}} - \gamma \underbrace{\mathbb{E}_{a\sim p}[H^2(M(a), \overline{M}(a))]}_{\text{information gain from obs.}}$$

- $\overline{M}$: a reference model
- $H^2(P, Q) = \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$ is the squared Hellinger distance
- $\gamma > 0$: a Lagrangian parameter

DEC result: an overview

Theorem (Foster, Kakade, Qian, Rakhlin, 2021)

For any model class $\mathcal{M}$:

- lower bound: for a worst case $M \in \mathcal{M}$, any algorithm must have

$$\mathbb{E}[\mathrm{Reg}(T)] \gtrsim \min_{\gamma>0} \left(\max_{\overline{M}\in\mathcal{M}} \mathrm{dec}_\gamma(\mathcal{M}_\gamma(\overline{M}), \overline{M}) \cdot T + \gamma \right)$$

  where $\mathcal{M}_\gamma \subseteq \mathcal{M}$ is a "localized set";

- upper bound: there is an algorithm that achieves

$$\mathbb{E}[\mathrm{Reg}(T)] \lesssim \min_{\gamma>0} \left(\max_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma(\mathcal{M}, \overline{M}) \cdot T + \gamma \cdot \mathrm{Est}(\mathcal{M}) \right),$$

  where $\mathrm{Est}(\mathcal{M}) \leq \log|\mathcal{M}|$ is the optimal rate for cond. density estimation for $\mathcal{M}$.

DEC result: an overview

Theorem (Foster, Kakade, Qian, Rakhlin, 2021)

For any model class $\mathcal{M}$:

- lower bound: for a worst case $M \in \mathcal{M}$, any algorithm must have

$$\mathbb{E}[\mathrm{Reg}(T)] \gtrsim \min_{\gamma>0} \left(\max_{\overline{M} \in \mathcal{M}} \mathrm{dec}_\gamma(\mathcal{M}_\gamma(\overline{M}), \overline{M}) \cdot T + \gamma \right)$$

  where $\mathcal{M}_\gamma \subseteq \mathcal{M}$ is a "localized set";

- upper bound: there is an algorithm that achieves

$$\mathbb{E}[\mathrm{Reg}(T)] \lesssim \min_{\gamma>0} \left(\max_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma(\mathcal{M}, \overline{M}) \cdot T + \gamma \cdot \mathrm{Est}(\mathcal{M}) \right),$$

  where $\mathrm{Est}(\mathcal{M}) \le \log |\mathcal{M}|$ is the optimal rate for cond. density estimation for $\mathcal{M}$.

Several gaps:

✗ UB has full class $\mathcal{M}$, LB has localized class $\mathcal{M}_\gamma(\overline{M})$

✗ UB takes $\overline{M} \in \mathrm{co}(\mathcal{M})$, LB takes $\overline{M} \in \mathcal{M}$

✗ UB has $\mathrm{Est}(\mathcal{M})$, LB does not

Constrained DEC

For $\varepsilon > 0$, the constrained decision-to-estimation coefficient (DEC) of a model class $\mathcal{M}$ is defined as

$$\mathrm{dec}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}} \inf_{p \in \Delta(\mathcal{A})} \sup_{M \in \mathcal{M} \cup \{\overline{M}\}} \left\{ \mathbb{E}_{a \sim p}[r_{\star}^{M} - r^{M}(a)] : \mathbb{E}_{a \sim p}[H^2(M(a), \overline{M}(a))] \leq \varepsilon^2 \right\}$$

Constrained DEC

Constrained DEC

For $\varepsilon > 0$, the constrained decision-to-estimation coefficient (DEC) of a model class $\mathcal{M}$ is defined as

$$\mathrm{dec}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}} \inf_{p \in \Delta(\mathcal{A})} \sup_{M \in \mathcal{M} \cup \{\overline{M}\}} \left\{ \mathbb{E}_{a \sim p}[r_{\star}^{M} - r^{M}(a)] : \mathbb{E}_{a \sim p}[H^2(M(a), \overline{M}(a))] \leq \varepsilon^2 \right\}$$

Features:

- hard constraint on the information gain
- connect with original DEC via Lagrangian:

$$\mathrm{dec}_{\varepsilon}(\mathcal{M}) \leq \inf_{\gamma > 0} \left\{ \sup_{\overline{M}} \mathrm{dec}_{\gamma}(\mathcal{M}, \overline{M}) + \gamma \varepsilon^2 \right\}$$

- converse does not hold (strong duality fails)

Connection to modulus of continuity in statistics

Hellinger modulus of continuity

$$\omega_\varepsilon(\mathcal{M}) = \sup_{M, M' \in \mathcal{M}} \left\{ \| T(M) - T(M') \| : H^2(M, M') \le \varepsilon^2 \right\}$$

- lower bound: Le Cam's two-point method ($\varepsilon \asymp T^{-1/2}$)
- simple upper bound: projection-based estimator ($\varepsilon \asymp \sqrt{\log |\mathcal{M}|/T}$)
- better upper bound: strong duality results ($\varepsilon \asymp T^{-1/2}$) when T is linear, e.g. [Donoho and Liu, 1987, 1991; Juditsky and Nemirovski, 2009; Polyanskiy and Wu, 2019]

Constrained DEC: main results

Theorem (Foster, Golowich, Han, 2023)

For any model class $\mathcal{M}$:

- lower bound: for a worst case $M \in \mathcal{M}$, any algorithm must have

$$\mathbb{E}[\mathrm{Reg}(T)] \gtrsim \mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot T,$$

  for $\underline{\varepsilon}(T) = \widetilde{\Theta}(\sqrt{1/T})$;

- upper bound: there is an algorithm that achieves

$$\mathbb{E}[\mathrm{Reg}(T)] \lesssim \mathrm{dec}_{\overline{\varepsilon}(T)}(\mathcal{M}) \cdot T,$$

  for $\overline{\varepsilon}(T) = \widetilde{\Theta}(\sqrt{\mathrm{Est}(\mathcal{M})/T}) = \widetilde{O}(\sqrt{\log|\mathcal{M}|/T})$.

Constrained DEC: main results

Theorem (Foster, Golowich, Han, 2023)

For any model class $\mathcal{M}$:

- lower bound: for a worst case $M \in \mathcal{M}$, any algorithm must have

$$\mathbb{E}[\mathrm{Reg}(T)] \gtrsim \mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot T,$$

  for $\underline{\varepsilon}(T) = \widetilde{\Theta}(\sqrt{1/T})$;

- upper bound: there is an algorithm that achieves

$$\mathbb{E}[\mathrm{Reg}(T)] \lesssim \mathrm{dec}_{\overline{\varepsilon}(T)}(\mathcal{M}) \cdot T,$$

  for $\overline{\varepsilon}(T) = \widetilde{\Theta}(\sqrt{\mathrm{Est}(\mathcal{M})/T}) = \widetilde{O}(\sqrt{\log|\mathcal{M}|/T})$.

Gaps revisited:

- ✓ no localization in both UB and LB
- ✓ no constraint on $\overline{M}$ in both UB and LB
- ✗ UB still has $\mathrm{Est}(\mathcal{M})$, LB does not (more in second part of the talk)
- ✓ uniformly improves over DEC results, with arbitrarily large separation

Constrained DEC: examples

setting	$\mathrm{dec}_\varepsilon(\mathcal{M})$	lower bound	LB tightness
Multi-Armed Bandit	$\varepsilon\sqrt{A}$	$\sqrt{AT}$	✓
Multi-Armed Bandit w/ gap	$\Delta \cdot 1(\varepsilon > \Delta/\sqrt{A})$	A/Δ	✓
Linear Bandit	$\varepsilon\sqrt{d}$	$\sqrt{dT}$	✗
Lipschitz Bandit	$\varepsilon^{1-\frac{d}{d+2}}$	$T^{\frac{d+1}{d+2}}$	✓
ReLU Bandit	$1(\varepsilon > 2^{-\Omega(d)})$	$2^{\Omega(d)}$	✓
Tabular RL	$\varepsilon\sqrt{HSA}$	$\sqrt{HSAT}$	✓
Linear MDP	$\varepsilon\sqrt{d}$	$\sqrt{dT}$	✗
RL w/ linear $Q^\star$	$1(\varepsilon \geq 2^{-\Omega(d)} \vee 2^{-\Omega(H)})$	$2^{\Omega(d)} \wedge 2^{\Omega(H)}$	✓
Deterministic RL w/ linear $Q^\star$	$1(\varepsilon \leq 1/\sqrt{d})$	d	✓

Proof of lower bound

$$\mathrm{dec}_\varepsilon(\mathcal{M}) = \sup_{\overline{M}} \inf_{p\in\Delta(\mathcal{A})} \sup_{M\in\mathcal{M}\cup\{\overline{M}\}} \left\{ \mathbb{E}_{a\sim p}[r_\star^M - r^M(a)] : \mathbb{E}_{a\sim p}[H^2(M(a), \overline{M}(a))] \leq \varepsilon^2 \right\}$$

Theorem (formal statement of lower bound)

Let $\underline{\varepsilon}(T) \asymp 1/\sqrt{T\log T}$, and assume that $\mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \geq C \cdot \underline{\varepsilon}(T)$ for a large constant C. Then for a worst case $M \in \mathcal{M}$, any algorithm must have

$$\mathbb{E}_M[\mathrm{Reg}(T)] \gtrsim \mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot T.$$

Proof of lower bound

$$\mathrm{dec}_{\varepsilon}(\mathcal{M}) = \sup_{\overline{M}} \inf_{p\in\Delta(\mathcal{A})} \sup_{M\in\mathcal{M}\cup\{\overline{M}\}} \left\{ \mathbb{E}_{a\sim p}[r_{\star}^{M} - r^{M}(a)] : \mathbb{E}_{a\sim p}[H^2(M(a), \overline{M}(a))] \leq \varepsilon^2 \right\}$$

Theorem (formal statement of lower bound)

Let $\underline{\varepsilon}(T) \asymp 1/\sqrt{T\log T}$, and assume that $\mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \geq C \cdot \underline{\varepsilon}(T)$ for a large constant C. Then for a worst case $M \in \mathcal{M}$, any algorithm must have

$$\mathbb{E}_M[\mathrm{Reg}(T)] \gtrsim \mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M}) \cdot T.$$

Preparations:

- $\overline{M} \in \mathcal{M}$: any fixed reference model
- $p_{\overline{M}} = \mathbb{E}_{\overline{M}}[T^{-1}\sum_{t=1}^{T} p_t(\cdot \mid \mathcal{H}_{t-1})]$: learner's average play under $\overline{M}$
- M: the inner maximizer under $p = p_{\overline{M}}$
- $p_M = \mathbb{E}_M[T^{-1}\sum_{t=1}^{T} p_t(\cdot \mid \mathcal{H}_{t-1})]$: learner's average play under M

Two-point argument

- Let $g^M(a) = r_\star^M - r^M(a)$ and $\Delta = \mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})$, it suffices to arrive at a contradiction based on the following inequalities:

$$\begin{aligned}
\mathbb{E}_{a\sim p_{\overline{M}}}[g^M(a)] &\geq \Delta, && \text{(defn. of constrained DEC - OBJ)}\\
\mathbb{E}_{a\sim p_{\overline{M}}}[H^2(M(a),\overline{M}(a))] &\leq \underline{\varepsilon}(T)^2, && \text{(constraints - C)}\\
\mathbb{E}_{a\sim p_M}[g^M(a)] &\leq c\Delta, && \text{(small regret under } M \text{ - } S_M)\\
\mathbb{E}_{a\sim p_{\overline{M}}}[g^{\overline{M}}(a)] &\leq c\Delta. && \text{(small regret under } \overline{M} \text{ - } S_{\overline{M}})
\end{aligned}$$

Two-point argument

- Let $g^M(a) = r_\star^M - r^M(a)$ and $\Delta = \mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})$, it suffices to arrive at a contradiction based on the following inequalities:

$$
\begin{aligned}
\mathbb{E}_{a\sim p_{\overline{M}}}[g^M(a)] &\geq \Delta, && \text{(defn. of constrained DEC - OBJ)}\\
\mathbb{E}_{a\sim p_{\overline{M}}}[H^2(M(a), \overline{M}(a))] &\leq \underline{\varepsilon}(T)^2, && \text{(constraints - C)}\\
\mathbb{E}_{a\sim p_M}[g^M(a)] &\leq c\Delta, && \text{(small regret under } M \text{ - } S_M\text{)}\\
\mathbb{E}_{a\sim p_{\overline{M}}}[g^{\overline{M}}(a)] &\leq c\Delta. && \text{(small regret under } \overline{M} \text{ - } S_{\overline{M}}\text{)}
\end{aligned}
$$

- Not hard to show that

$$(\mathrm{C}) \Rightarrow \mathrm{TV}(p_M, p_{\overline{M}}) \leq 0.1 \qquad \text{(indistinguishability - TV)}$$

Two-point argument

- Let $g^M(a) = r_\star^M - r^M(a)$ and $\Delta = \mathrm{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})$, it suffices to arrive at a contradiction based on the following inequalities:

$$\mathbb{E}_{a\sim p_{\overline{M}}}[g^M(a)] \geq \Delta, \qquad \text{(defn. of constrained DEC - OBJ)}$$

$$\mathbb{E}_{a\sim p_{\overline{M}}}[H^2(M(a), \overline{M}(a))] \leq \underline{\varepsilon}(T)^2, \qquad \text{(constraints - C)}$$

$$\mathbb{E}_{a\sim p_M}[g^M(a)] \leq c\Delta, \qquad \text{(small regret under } M \text{ - } S_M)$$

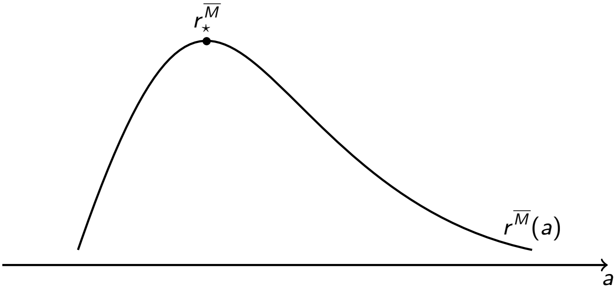
$$\mathbb{E}_{a\sim p_{\overline{M}}}[g^{\overline{M}}(a)] \leq c\Delta. \qquad \text{(small regret under } \overline{M} \text{ - } S_{\overline{M}})$$

- Not hard to show that

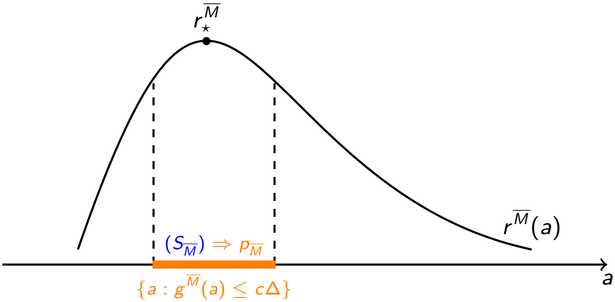
$$(C) \Rightarrow \mathrm{TV}(p_M, p_{\overline{M}}) \leq 0.1 \qquad \text{(indistinguishability - TV)}$$

- Problems with some attempts:
  - $\rightarrow$ $(S_{\overline{M}}) + (\mathrm{TV}) \Rightarrow \neg\ (S_M)$: $g^M(a) + g^{\overline{M}}(a)$ could be small
  - $\rightarrow$ $(\mathrm{OBJ}) + (\mathrm{TV}) \Rightarrow \neg\ (S_M)$: $g^M(a)$ might have a heavy tail under $a \sim p_{\overline{M}}$

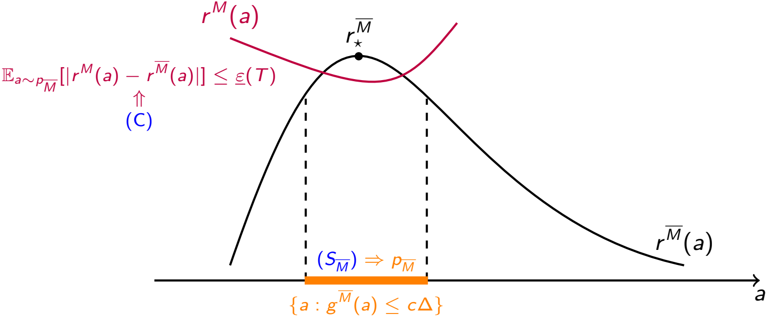
A pictorial proof



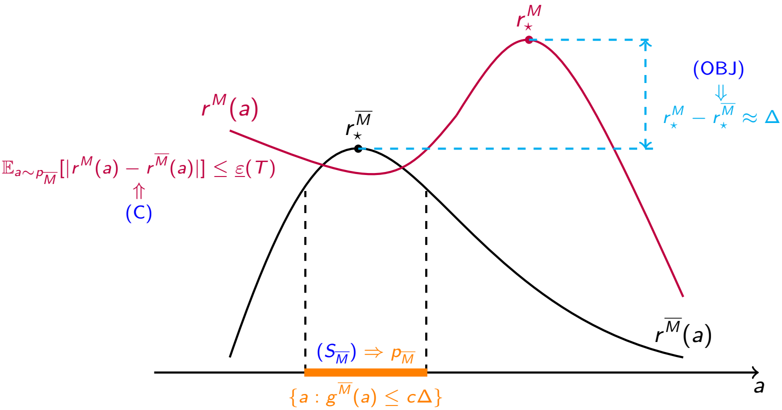
A pictorial proof



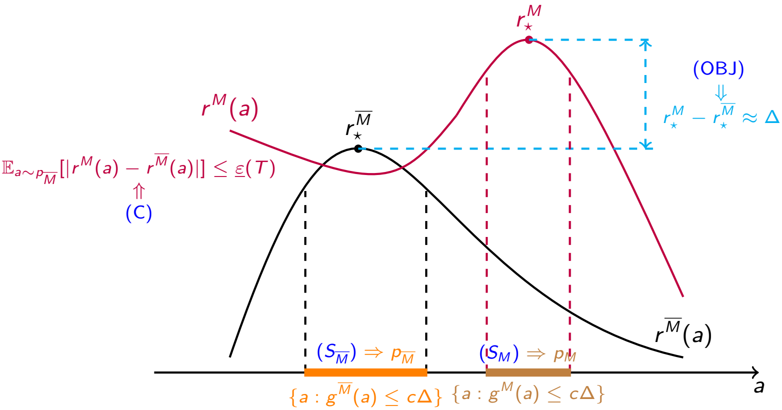
A pictorial proof



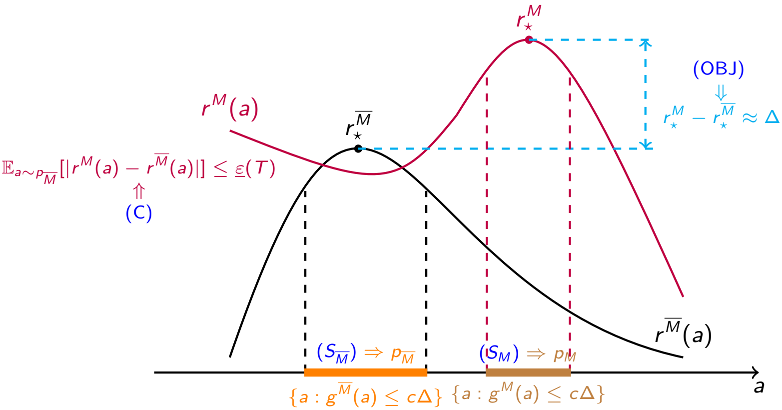
A pictorial proof



A pictorial proof



A pictorial proof



This is a contradiction to (TV)

Role of improper $\overline{M}$

Lower bound view:

- we use a reduction to deal with improper $\overline{M}$
- recently, [Glasgow and Rakhlin, 2023] showed that the condition $(S_{\overline{M}})$ could be replaced by $p_{\overline{M}}(g^{\overline{M}}(a) \in [b, b + c\Delta]) = \Omega(1)$ for any translation b

Role of improper $\overline{M}$

Lower bound view:

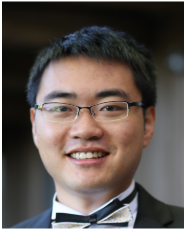
- we use a reduction to deal with improper $\overline{M}$
- recently, [Glasgow and Rakhlin, 2023] showed that the condition ($S_{\overline{M}}$) could be replaced by $p_{\overline{M}}(g^{\overline{M}}(a) \in [b, b + c\Delta]) = \Omega(1)$ for any translation b

Upper bound view:

- the learner could use an improper estimate $\widehat{M}_t$ for $M^\star$
- algorithmic idea: at time t, find an online estimatior $\widehat{M}_t$, then choose

$$a_t \sim p_t = \arg\min_p \left[\sup_{M \in \mathcal{M} \cup \{\widehat{M}_t\}} \left\{ \mathbb{E}_p[r_\star^M - r^M(a)] : \mathbb{E}_{a\sim p}[H^2(M(a), \widehat{M}_t(a))] \leq \frac{\mathrm{Est}(\mathcal{M})}{T} \right\} \right]$$

Part II: Interactive Fano-type lower bound



Jiantao Jiao
Berkeley EECS

Nived Rajaraman
Berkeley EECS

Kannan Ramchandran
Berkeley EECS

"Statistical Complexity and Optimal Algorithms for Non-linear Ridge Bandits"
(arXiv: 2302.06025)

Ridge bandits

Setting for ridge bandits:

- model class: $\mathcal{M} = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$
- action space: $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$
- mean reward: $r_\theta(a) = f(\langle \theta, a \rangle)$
- known link function: $f : [-1, 1] \to [-1, 1]$

Setting for ridge bandits:

- model class: $\mathcal{M} = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$
- action space: $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$
- mean reward: $r_\theta(a) = f(\langle \theta, a \rangle)$
- known link function: $f : [-1, 1] \to [-1, 1]$

Interactive version of generalized linear regression:

$$r_t = f(\langle \theta^\star, a_t \rangle) + \varepsilon_t, \quad t = 1, 2, \cdots, T.$$

Ridge bandits

Setting for ridge bandits:

- model class: $\mathcal{M} = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$
- action space: $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \le 1\}$
- mean reward: $r_\theta(a) = f(\langle \theta, a \rangle)$
- known link function: $f : [-1, 1] \to [-1, 1]$

Interactive version of generalized linear regression:

$$r_t = f(\langle \theta^\star, a_t \rangle) + \varepsilon_t, \quad t = 1, 2, \cdots, T.$$

Questions

- Does interactivity help?
- Does non-linearity of f make the problem more difficult/interesting?

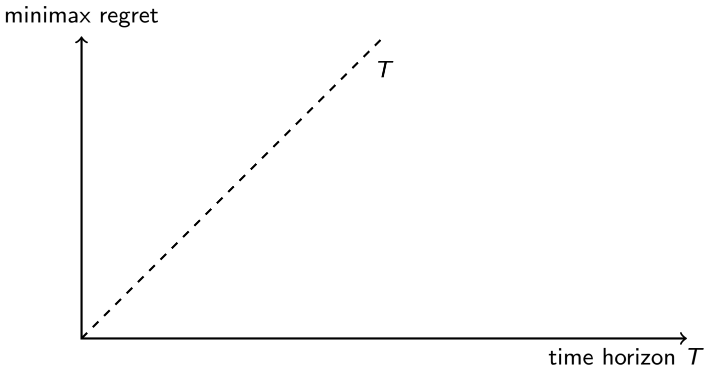
A non-linear bandit example

$$f(\langle\theta, a\rangle) = \langle\theta, a\rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$

A motivating example

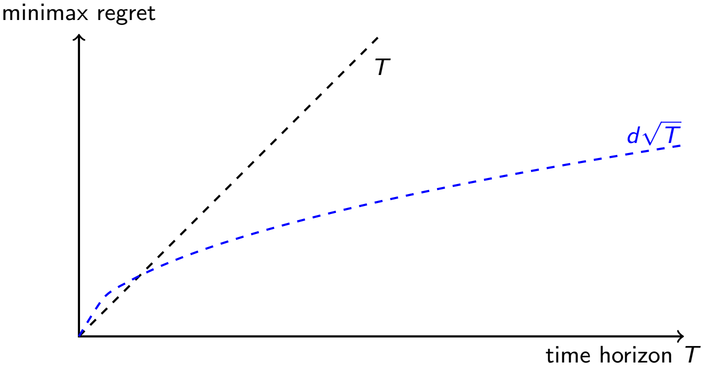
A non-linear bandit example

$$f(\langle \theta, a \rangle) = \langle \theta, a \rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



A non-linear bandit example

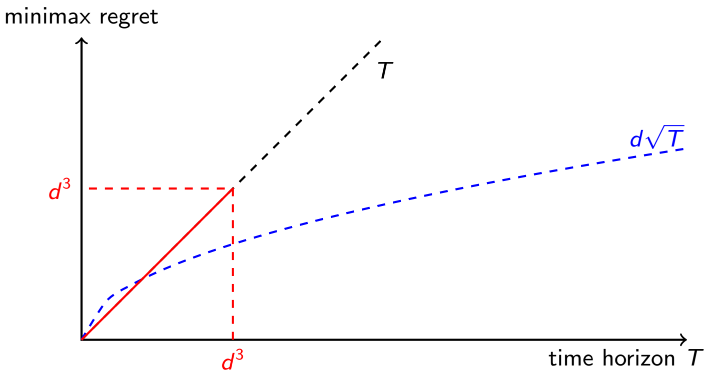
$$f(\langle\theta, a\rangle) = \langle\theta, a\rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



A motivating example

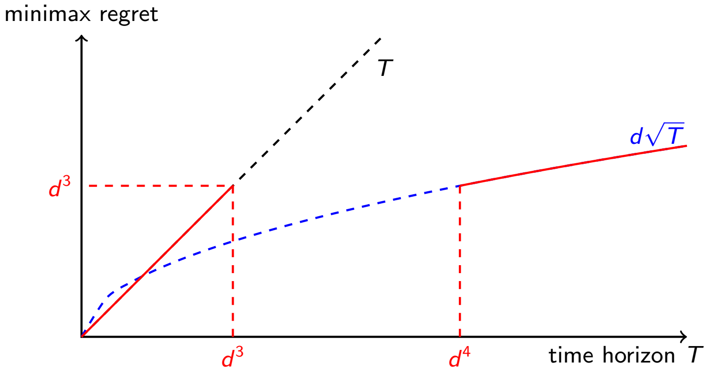
A non-linear bandit example

$$f(\langle\theta, a\rangle) = \langle\theta, a\rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



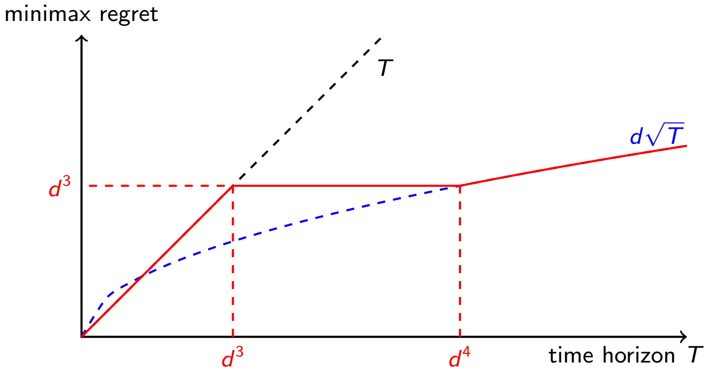
A non-linear bandit example

$$f(\langle \theta, a \rangle) = \langle \theta, a \rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



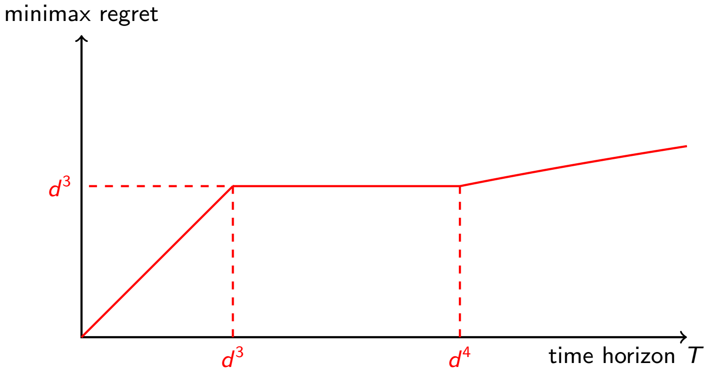
A non-linear bandit example

$$f(\langle\theta, a\rangle) = \langle\theta, a\rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



A non-linear bandit example

$$f(\langle\theta, a\rangle) = \langle\theta, a\rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$

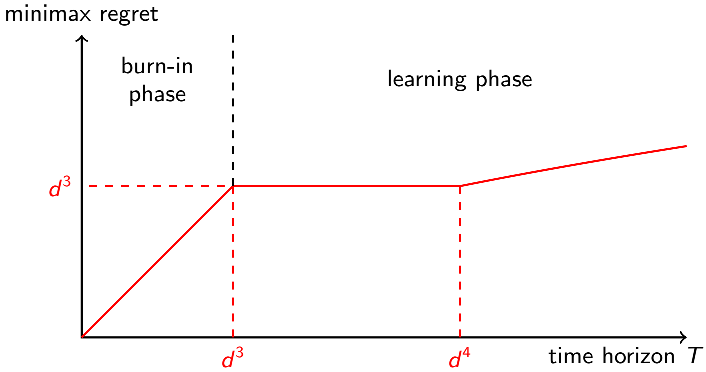


minimax regret $\asymp \min\{T, d^3 + d\sqrt{T}\}$.

A motivating example

A non-linear bandit example

$$f(\langle\theta, a\rangle) = \langle\theta, a\rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



minimax regret $\asymp \min\{T, d^3 + d\sqrt{T}\}$.

Curious phenomena

Curious phenomena in non-linear bandits:

- phase transition in the regret
- burn-in phase: regret grows linearly and results in a burn-in cost
  - $\rightarrow$ find a good "initial action" to start learning
- learning phase: regret grows sublinearly and looks like a linear bandit
  - $\rightarrow$ learning starts from the good initial action

Curious phenomena

Curious phenomena in non-linear bandits:

- phase transition in the regret
- burn-in phase: regret grows linearly and results in a burn-in cost
  $\rightarrow$ find a good “initial action” to start learning
- learning phase: regret grows sublinearly and looks like a linear bandit
  $\rightarrow$ learning starts from the good initial action

Questions

- what is the optimal burn-in cost?
- what algorithms should we use in different phases?

Ridge bandits:

- linear bandit $f(x) = x$: optimal regret $\widetilde{\Theta}(d\sqrt{T})$ [Dani et al. 2008, Chu et al. 2011, Abbasi-Yadkori et al. 2011]
- generalized linear bandit with $c_1 \leq |f'(x)| \leq c_2$: same as linear bandit [Filippi et al. 2010, Russo and Van Roy 2014]
- concave bandit (f is concave): same as linear bandit [Lattimore, 2021]
- bandit phase retrieval ($f(x) = x^2$): same as linear bandit [Lattimore and Hao, 2021]
- polynomial bandit ($f(x) = x^p, p \geq 2$): optimal regret $\widetilde{\Theta}(\sqrt{d^p T})$ assuming $\|\theta\|_2 \leq 1$ [Huang et al. 2021]

Ridge bandits:

- linear bandit $f(x) = x$: optimal regret $\widetilde{\Theta}(d\sqrt{T})$ [Dani et al. 2008, Chu et al. 2011, Abbasi-Yadkori et al. 2011]
- generalized linear bandit with $c_1 \le |f'(x)| \le c_2$: same as linear bandit [Filippi et al. 2010, Russo and Van Roy 2014]
- concave bandit (f is concave): same as linear bandit [Lattimore, 2021]
- bandit phase retrieval ($f(x) = x^2$): same as linear bandit [Lattimore and Hao, 2021]
- polynomial bandit ($f(x) = x^p, p \ge 2$): optimal regret $\widetilde{\Theta}(\sqrt{d^p T})$ assuming $\|\theta\|_2 \le 1$ [Huang et al. 2021]

General complexity measures for bandits:

- decision-estimation coefficient (DEC) [Foster et al. 2021, 2022]
- information ratio [Lattimore, 2022]
- often do not lead to tight regret dependence on d (the gap of $\mathrm{Est}(\mathcal{M})$)

Main result

Only assumption on f: f is increasing on $[-1, 1]$ with $f(0) = 0$

$\rightarrow$ aim to maximize the inner product $\langle \theta^\star, a_t \rangle$

Main result

Only assumption on f: f is increasing on $[-1,1]$ with $f(0)=0$

$\rightarrow$ aim to maximize the inner product $\langle \theta^\star, a_t \rangle$

Theorem (Rajaraman, Han, Jiao, Ramchandran, 2023)

The minimax sample complexity $T^\star(\varepsilon)$ of achieving $\langle \theta^\star, a_T \rangle \ge \varepsilon \in [1/\sqrt{d}, 1/2]$ satisfies (within poly-logarithmic factors)

$$T^\star(\varepsilon) \lesssim d^2 \cdot \int_{1/\sqrt{d}}^{\varepsilon} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \le y \le x} \min_{z \in [y/2, y]} f'(z)^2},$$

$$T^\star(\varepsilon) \gtrsim d \cdot \int_{1/\sqrt{d}}^{\varepsilon} \frac{\mathrm{d}(x^2)}{f(x)^2}.$$

Main result

Only assumption on f: f is increasing on $[-1, 1]$ with $f(0) = 0$

$\rightarrow$ aim to maximize the inner product $\langle \theta^\star, a_t \rangle$

Theorem (Rajaraman, Han, Jiao, Ramchandran, 2023)

The minimax sample complexity $T^\star(\varepsilon)$ of achieving $\langle \theta^\star, a_T \rangle \geq \varepsilon \in [1/\sqrt{d}, 1/2]$ satisfies (within poly-logarithmic factors)

$$T^\star(\varepsilon) \lesssim d^2 \cdot \int_{1/\sqrt{d}}^{\varepsilon} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \leq y \leq x} \min_{z \in [y/2, y]} f'(z)^2},$$

$$T^\star(\varepsilon) \gtrsim d \cdot \int_{1/\sqrt{d}}^{\varepsilon} \frac{\mathrm{d}(x^2)}{f(x)^2}.$$

- pointwise upper and lower bounds
- burn-in cost by choosing $\varepsilon = 1/2$
- learning trajectory via differential equations

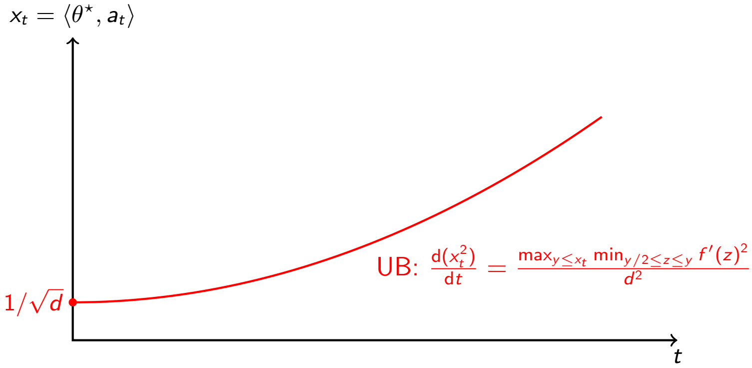
Learning trajectory

Theorem (learning trajectory)

Learning trajectory

Theorem (learning trajectory)

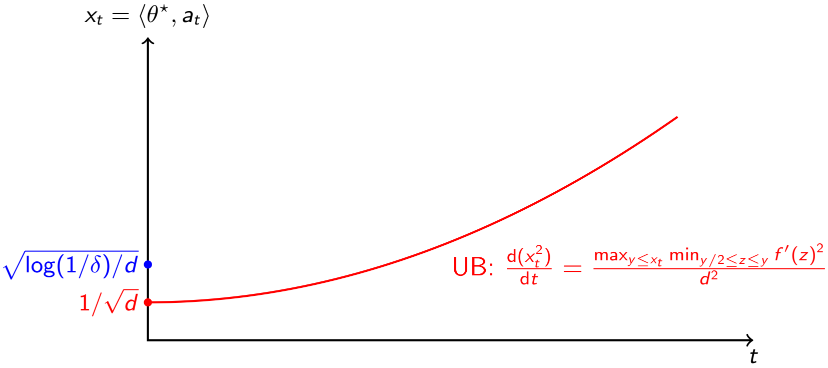
Learning trajectory



Theorem (learning trajectory)

- there is an algorithm attaining the UB learning curve

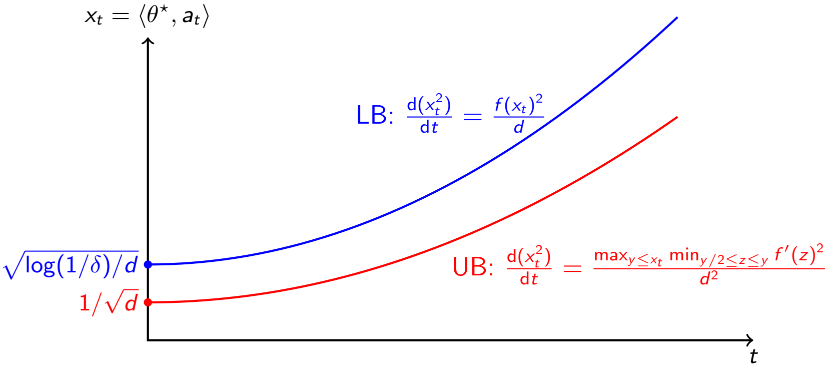
Learning trajectory



Theorem (learning trajectory)

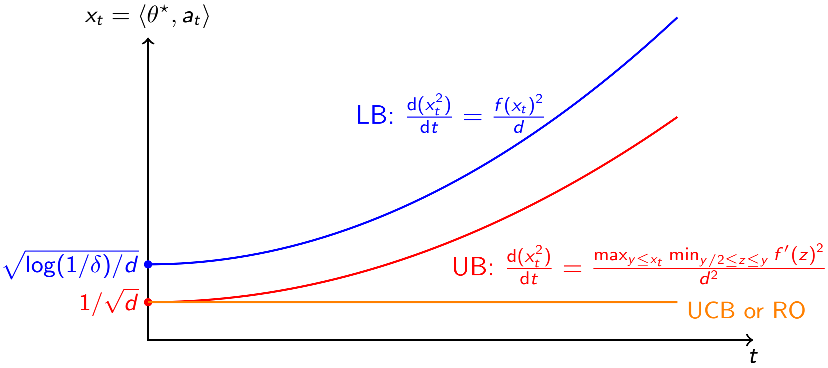
- there is an algorithm attaining the UB learning curve

Learning trajectory



Theorem (learning trajectory)

- there is an algorithm attaining the UB learning curve
- for any algorithm, its learning trajectory lies below the LB learning curve with probability at least $1 - T\delta$ under $\theta^\star \sim \mathrm{Unif}(\mathbb{S}^{d-1})$



Theorem (learning trajectory)

- there is an algorithm attaining the UB learning curve
- for any algorithm, its learning trajectory lies below the LB learning curve with probability at least $1 - T\delta$ under $\theta^\star \sim \mathrm{Unif}(\mathbb{S}^{d-1})$
- UCB or RO algorithms makes no progress whenever $t < d/f(1/\sqrt{d})^2$

Formal lower bound

Theorem (formal lower bound)

Let $\delta > 0$ be any parameter, and $c > 0$ be a large absolute constant. Define a sequence $\{\varepsilon_t\}_{t\geq 1}$ with

$$\varepsilon_1 = \sqrt{\frac{c\log(1/\delta)}{d}}, \quad \varepsilon_{t+1}^2 = \varepsilon_t^2 + \frac{c}{d} f(\varepsilon_t)^2, \quad t \geq 1.$$

Formal lower bound

Theorem (formal lower bound)

Let $\delta > 0$ be any parameter, and $c > 0$ be a large absolute constant. Define a sequence $\{\varepsilon_t\}_{t\geq 1}$ with

$$\varepsilon_1 = \sqrt{\frac{c\log(1/\delta)}{d}}, \quad \varepsilon_{t+1}^2 = \varepsilon_t^2 + \frac{c}{d} f(\varepsilon_t)^2, \quad t \geq 1.$$

Then if $\theta^\star \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, any learner $\{a_t\}_{t\geq 1}$ satisfies that

$$\mathbb{P}\left(\bigcap_{1\leq t\leq T} \{\langle \theta^\star, a_t\rangle \leq \varepsilon_t\}\right) \geq 1 - T\delta.$$

Theorem (formal lower bound)

Let $\delta > 0$ be any parameter, and $c > 0$ be a large absolute constant. Define a sequence $\{\varepsilon_t\}_{t \geq 1}$ with

$$\varepsilon_1 = \sqrt{\frac{c \log(1/\delta)}{d}}, \quad \varepsilon_{t+1}^2 = \varepsilon_t^2 + \frac{c}{d} f(\varepsilon_t)^2, \quad t \geq 1.$$

Then if $\theta^\star \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, any learner $\{a_t\}_{t \geq 1}$ satisfies that

$$\mathbb{P}\left(\bigcap_{1 \leq t \leq T} \{ \langle \theta^\star, a_t \rangle \leq \varepsilon_t \} \right) \geq 1 - T\delta.$$

- the continuous-time version of $\{\varepsilon_t\}$ gives the differential equation

Information-theoretic insights

Let $I_t = I(\theta^\star; \mathcal{H}_t)$ be the mutual information between the true parameter $\theta^\star$ and the history $\mathcal{H}_t$ up to time t, then

$$\begin{aligned}
I_{t+1} - I_t &= I(\theta^\star; r_{t+1} \mid a_{t+1}, \mathcal{H}_t) \\
&\leq \mathbb{E}\left[\frac{1}{2}\log\left(1 + \mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2]\right)\right] \\
&\leq \frac{1}{2}\mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2].
\end{aligned}$$

Information-theoretic insights

Let $I_t = I(\theta^\star; \mathcal{H}_t)$ be the mutual information between the true parameter $\theta^\star$ and the history $\mathcal{H}_t$ up to time t, then

$$\begin{aligned} I_{t+1} - I_t &= I(\theta^\star; r_{t+1} \mid a_{t+1}, \mathcal{H}_t) \\ &\le \mathbb{E}\left[\frac{1}{2}\log\left(1 + \mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2]\right)\right] \\ &\le \frac{1}{2}\mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2]. \end{aligned}$$

To argue that $\langle\theta^\star, a_{t+1}\rangle$ should not be large, note that

$$I(\theta^\star; a_{t+1}) \le I(\theta^\star; \mathcal{H}_t) = I_t.$$

Information-theoretic insights

Let $I_t = I(\theta^\star; \mathcal{H}_t)$ be the mutual information between the true parameter $\theta^\star$ and the history $\mathcal{H}_t$ up to time t, then

$$\begin{aligned} I_{t+1} - I_t &= I(\theta^\star; r_{t+1} \mid a_{t+1}, \mathcal{H}_t) \\ &\leq \mathbb{E}\left[\frac{1}{2}\log\left(1 + \mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2]\right)\right] \\ &\leq \frac{1}{2}\mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2]. \end{aligned}$$

To argue that $\langle\theta^\star, a_{t+1}\rangle$ should not be large, note that

$$I(\theta^\star; a_{t+1}) \leq I(\theta^\star; \mathcal{H}_t) = I_t.$$

Key insight

$I(\theta^\star; a) \leq I \Longrightarrow |\langle\theta^\star, a\rangle| \lesssim \sqrt{I/d}$ with high probability.

Information-theoretic insights

Let $I_t = I(\theta^\star; \mathcal{H}_t)$ be the mutual information between the true parameter $\theta^\star$ and the history $\mathcal{H}_t$ up to time t, then

$$\begin{aligned} I_{t+1} - I_t &= I(\theta^\star; r_{t+1} \mid a_{t+1}, \mathcal{H}_t) \\ &\leq \mathbb{E}\left[\frac{1}{2}\log\left(1 + \mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2]\right)\right] \\ &\leq \frac{1}{2}\mathbb{E}[f(\langle\theta^\star, a_{t+1}\rangle)^2]. \end{aligned}$$

To argue that $\langle\theta^\star, a_{t+1}\rangle$ should not be large, note that

$$I(\theta^\star; a_{t+1}) \leq I(\theta^\star; \mathcal{H}_t) = I_t.$$

Key insight

$$I(\theta^\star; a) \leq I \Longrightarrow |\langle\theta^\star, a\rangle| \lesssim \sqrt{I/d} \text{ with high probability.}$$

Applying the insight gives the desired recursion

$$\varepsilon_{t+1}^2 - \varepsilon_t^2 \lesssim \frac{1}{d} f(\varepsilon_t)^2.$$

More on the above insights

- reasoning behind the insight:

$$a \mid \theta^\star \sim \mathrm{Unif}(\{a \in \mathbb{S}^{d-1} : \langle a, \theta^\star \rangle \geq \varepsilon\}) \Longrightarrow I(a; \theta^\star) \asymp d\varepsilon^2$$

More on the above insights

- reasoning behind the insight:

$$a \mid \theta^\star \sim \mathsf{Unif}(\{a \in \mathbb{S}^{d-1} : \langle a, \theta^\star\rangle \geq \varepsilon\}) \Longrightarrow I(a; \theta^\star) \asymp d\varepsilon^2$$

- however, it does not hold with high probability: Fano's inequality only gives

$$\mathbb{P}(|\langle \theta^\star, a\rangle| \leq \varepsilon) \geq 1 - \frac{I(\theta^\star; a) + \log 2}{\Theta(d\varepsilon^2)},$$

  which is tight for the worst-case distribution of $(\theta^\star, a)$

More on the above insights

- reasoning behind the insight:

$$a \mid \theta^\star \sim \mathrm{Unif}(\{a \in \mathbb{S}^{d-1} : \langle a, \theta^\star \rangle \geq \varepsilon\}) \Longrightarrow I(a; \theta^\star) \asymp d\varepsilon^2$$

- however, it does not hold with high probability: Fano's inequality only gives

$$\mathbb{P}(|\langle \theta^\star, a \rangle| \leq \varepsilon) \geq 1 - \frac{I(\theta^\star; a) + \log 2}{\Theta(d\varepsilon^2)},$$

  which is tight for the worst-case distribution of $(\theta^\star, a)$
- our solution: use χ^2-informativity instead

Formal proof via χ^2-informativity

- χ^2-informativity between X and Y:

$$I_{\chi^2}(X;Y) = \inf_{Q_Y} \chi^2(P_{XY} \| P_X \times Q_Y),$$

where $\chi^2(P\|Q) = \int (\mathrm{d}P)^2/\mathrm{dQ} - 1$

Formal proof via χ^2-informativity

- χ^2-informativity between X and Y:

$$I_{\chi^2}(X;Y) = \inf_{Q_Y} \chi^2(P_{XY} \| P_X \times Q_Y),$$

  where $\chi^2(P\|Q) = \int (\mathrm{d}P)^2/\mathrm{dQ} - 1$

- error probability lower bound using χ^2-informativity:

$$\mathbb{P}(|\langle \theta^\star, a\rangle| \le \varepsilon) \ge 1 - e^{-\Theta(d\varepsilon^2)} \cdot \sqrt{I_{\chi^2}(\theta^\star; a) + 1}.$$

Formal proof via χ^2-informativity

- χ^2-informativity between X and Y:

$$I_{\chi^2}(X;Y) = \inf_{Q_Y} \chi^2(P_{XY} \| P_X \times Q_Y),$$

  where $\chi^2(P\|Q) = \int (\mathrm{d}P)^2/\mathrm{dQ} - 1$
- error probability lower bound using χ^2-informativity:

$$\mathbb{P}(|\langle \theta^\star, a\rangle| \le \varepsilon) \ge 1 - e^{-\Theta(d\varepsilon^2)} \cdot \sqrt{I_{\chi^2}(\theta^\star; a) + 1}.$$

- suffices to upper bound $I_{\chi^2}(\theta^\star; a_{t+1}) \le I_{\chi^2}(\theta^\star; \mathcal{H}_t)$ for each t

Formal proof via χ^2-informativity

- χ^2-informativity between X and Y:

$$I_{\chi^2}(X;Y) = \inf_{Q_Y} \chi^2(P_{XY} \| P_X \times Q_Y),$$

  where $\chi^2(P\|Q) = \int (\mathrm{d}P)^2/\mathrm{d}Q - 1$
- error probability lower bound using χ^2-informativity:

$$\mathbb{P}(|\langle \theta^\star, a\rangle| \le \varepsilon) \ge 1 - e^{-\Theta(d\varepsilon^2)} \cdot \sqrt{I_{\chi^2}(\theta^\star; a) + 1}.$$

- suffices to upper bound $I_{\chi^2}(\theta^\star; a_{t+1}) \le I_{\chi^2}(\theta^\star; \mathcal{H}_t)$ for each t
- issue: χ^2-informativity does not satisfy the chain rule or subadditivity

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t} \{ |\langle \theta^\star, a_s \rangle| \le \varepsilon_s \}$ be the error event

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t} \{ |\langle \theta^\star, a_s \rangle| \le \varepsilon_s \}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t} \{ |\langle \theta^\star, a_s \rangle| \le \varepsilon_s \}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \min_{\mathbb{Q}_{t-1}} \int \frac{\overbrace{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)} \pi(\theta^\star) \prod_{s \le t} \varphi(r_s - f(\langle \theta^\star, a_s \rangle)) \right]^2}^{\mathbb{P}(\theta^\star, \mathcal{H}_t | \mathcal{E}_t)^2}}{\underbrace{\pi(\theta^\star) \mathbb{Q}_{t-1}(r^{t-1}) \cdot \varphi(r_t)}_{\pi(\theta^\star) \mathbb{Q}_t(\mathcal{H}_t)}} \mathrm{d}\theta^\star \mathrm{d}r^t$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t}\{|\langle \theta^\star, a_s \rangle| \le \varepsilon_s\}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \min_{\mathbb{Q}_{t-1}} \int \frac{\overbrace{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)} \pi(\theta^\star) \prod_{s \le t} \varphi(r_s - f(\langle \theta^\star, a_s \rangle))\right]^2}^{\mathbb{P}(\theta^\star, \mathcal{H}_t | \mathcal{E}_t)^2}}{\underbrace{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1}) \cdot \varphi(r_t)}_{\pi(\theta^\star)\mathbb{Q}_t(\mathcal{H}_t)}} \mathrm{d}\theta^\star \mathrm{d}r^t$$

$$= \min_{\mathbb{Q}_{t-1}} \int \frac{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)} \pi(\theta^\star) \prod_{s \le t-1} \varphi(r_s - f(\langle \theta^\star, a_s \rangle))\right]^2}{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1})} \cdot \exp(f(\langle \theta^\star, a_t \rangle)^2) \mathrm{d}\theta^\star \mathrm{d}r^{t-1}$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t}\{|\langle \theta^\star, a_s\rangle| \le \varepsilon_s\}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \min_{\mathbb{Q}_{t-1}} \int \frac{\overbrace{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)}\pi(\theta^\star)\prod_{s\le t}\varphi(r_s - f(\langle\theta^\star, a_s\rangle))\right]^2}^{\mathbb{P}(\theta^\star, \mathcal{H}_t | \mathcal{E}_t)^2}}{\underbrace{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1}) \cdot \varphi(r_t)}_{\pi(\theta^\star)\mathbb{Q}_t(\mathcal{H}_t)}} \mathrm{d}\theta^\star \mathrm{d}r^t$$

$$= \min_{\mathbb{Q}_{t-1}} \int \frac{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)}\pi(\theta^\star)\prod_{s\le t-1}\varphi(r_s - f(\langle\theta^\star, a_s\rangle))\right]^2}{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1})} \cdot \exp(f(\langle\theta^\star, a_t\rangle)^2)\mathrm{d}\theta^\star \mathrm{d}r^{t-1}$$

$$\le \exp(f(\varepsilon_t)^2) \cdot \min_{\mathbb{Q}_{t-1}} \int \frac{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)}\pi(\theta^\star)\prod_{s\le t-1}\varphi(r_s - f(\langle\theta^\star, a_s\rangle))\right]^2}{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1})} \mathrm{d}r^{t-1}$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t}\{|\langle \theta^\star, a_s\rangle| \le \varepsilon_s\}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \min_{\mathbb{Q}_{t-1}} \int \frac{\overbrace{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)} \pi(\theta^\star) \prod_{s \le t} \varphi(r_s - f(\langle \theta^\star, a_s\rangle))\right]^2}^{\mathbb{P}(\theta^\star, \mathcal{H}_t | \mathcal{E}_t)^2}}{\underbrace{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1}) \cdot \varphi(r_t)}_{\pi(\theta^\star)\mathbb{Q}_t(\mathcal{H}_t)}} \mathrm{d}\theta^\star \mathrm{d}r^t$$

$$= \min_{\mathbb{Q}_{t-1}} \int \frac{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)} \pi(\theta^\star) \prod_{s \le t-1} \varphi(r_s - f(\langle \theta^\star, a_s\rangle))\right]^2}{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1})} \cdot \exp(f(\langle \theta^\star, a_t\rangle)^2) \mathrm{d}\theta^\star \mathrm{d}r^{t-1}$$

$$\le \exp(f(\varepsilon_t)^2) \cdot \min_{\mathbb{Q}_{t-1}} \int \frac{\left[\frac{\mathbb{1}(\mathcal{E}_t)}{\mathbb{P}(\mathcal{E}_t)} \pi(\theta^\star) \prod_{s \le t-1} \varphi(r_s - f(\langle \theta^\star, a_s\rangle))\right]^2}{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1})} \mathrm{d}r^{t-1}$$

$$\le \frac{\exp(f(\varepsilon_t)^2)}{\mathbb{P}(\mathcal{E}_t \mid \mathcal{E}_{t-1})^2} \cdot \min_{\mathbb{Q}_{t-1}} \int \frac{\left[\frac{\mathbb{1}(\mathcal{E}_{t-1})}{\mathbb{P}(\mathcal{E}_{t-1})} \pi(\theta^\star) \prod_{s \le t-1} \varphi(r_s - f(\langle \theta^\star, a_s\rangle))\right]^2}{\pi(\theta^\star)\mathbb{Q}_{t-1}(r^{t-1})} \mathrm{d}r^{t-1}$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t} \{ |\langle \theta^\star, a_s \rangle| \le \varepsilon_s \}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(f(\varepsilon_t)^2)}{\mathbb{P}(\mathcal{E}_t \mid \mathcal{E}_{t-1})^2} \left(I_{\chi^2}(\theta^\star; \mathcal{H}_{t-1} \mid \mathcal{E}_{t-1}) + 1 \right).$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s\le t}\{|\langle \theta^\star, a_s\rangle| \le \varepsilon_s\}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(f(\varepsilon_t)^2)}{\mathbb{P}(\mathcal{E}_t \mid \mathcal{E}_{t-1})^2} \left(I_{\chi^2}(\theta^\star; \mathcal{H}_{t-1} \mid \mathcal{E}_{t-1}) + 1\right).$$

- continuing this process gives

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(\sum_{s\le t} f(\varepsilon_s)^2)}{\mathbb{P}(\mathcal{E}_t)^2}.$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t}\{|\langle \theta^\star, a_s \rangle| \le \varepsilon_s\}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(f(\varepsilon_t)^2)}{\mathbb{P}(\mathcal{E}_t \mid \mathcal{E}_{t-1})^2} \left(I_{\chi^2}(\theta^\star; \mathcal{H}_{t-1} \mid \mathcal{E}_{t-1}) + 1\right).$$

- continuing this process gives

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(\sum_{s \le t} f(\varepsilon_s)^2)}{\mathbb{P}(\mathcal{E}_t)^2}.$$

- recursion of error probability:

$$\mathbb{P}(\mathcal{E}_{t+1}) = \mathbb{P}(\mathcal{E}_t) \cdot \mathbb{P}(|\langle \theta^\star, a_{t+1} \rangle| \le \varepsilon_{t+1} \mid \mathcal{E}_t)$$

Conditioning technique

- let $\mathcal{E}_t = \cap_{s\le t}\{|\langle\theta^\star, a_s\rangle| \le \varepsilon_s\}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star;\mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(f(\varepsilon_t)^2)}{\mathbb{P}(\mathcal{E}_t \mid \mathcal{E}_{t-1})^2}\left(I_{\chi^2}(\theta^\star;\mathcal{H}_{t-1} \mid \mathcal{E}_{t-1}) + 1\right).$$

- continuing this process gives

$$I_{\chi^2}(\theta^\star;\mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(\sum_{s\le t} f(\varepsilon_s)^2)}{\mathbb{P}(\mathcal{E}_t)^2}.$$

- recursion of error probability:

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{t+1}) &= \mathbb{P}(\mathcal{E}_t)\cdot\mathbb{P}(|\langle\theta^\star, a_{t+1}\rangle| \le \varepsilon_{t+1} \mid \mathcal{E}_t)\\ &\ge \mathbb{P}(\mathcal{E}_t)\left(1 - e^{-\Theta(d\varepsilon_{t+1}^2)}\sqrt{I_{\chi^2}(\theta^\star;\mathcal{H}_t \mid \mathcal{E}_t) + 1}\right)\end{aligned}$$

.

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \le t}\{|\langle \theta^\star, a_s\rangle| \le \varepsilon_s\}$ be the error event
- upper bound of conditioned χ^2-informativity:

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(f(\varepsilon_t)^2)}{\mathbb{P}(\mathcal{E}_t \mid \mathcal{E}_{t-1})^2} \left(I_{\chi^2}(\theta^\star; \mathcal{H}_{t-1} \mid \mathcal{E}_{t-1}) + 1\right).$$

- continuing this process gives

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1 \le \frac{\exp(\sum_{s \le t} f(\varepsilon_s)^2)}{\mathbb{P}(\mathcal{E}_t)^2}.$$

- recursion of error probability:

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{t+1}) &= \mathbb{P}(\mathcal{E}_t) \cdot \mathbb{P}(|\langle \theta^\star, a_{t+1}\rangle| \le \varepsilon_{t+1} \mid \mathcal{E}_t) \\ &\ge \mathbb{P}(\mathcal{E}_t)\left(1 - e^{-\Theta(d\varepsilon_{t+1}^2)}\sqrt{I_{\chi^2}(\theta^\star; \mathcal{H}_t \mid \mathcal{E}_t) + 1}\right) \\ &\ge \mathbb{P}(\mathcal{E}_t) - \underbrace{e^{-\Theta(d\varepsilon_{t+1}^2) + \frac{1}{2}\sum_{s \le t} f(\varepsilon_s)^2}}_{=\delta}.\end{aligned}$$

- fill in the gap between upper and lower bounds

$$I_t - I_{t-1} \leq \mathrm{Var}(f(\langle \theta^\star, a_t \rangle) \mid a_t, \mathcal{H}_{t-1}) \overset{?}{\lesssim} \max_{y \leq \varepsilon_t} \frac{f'(y)^2}{d}$$

- unclear if the above holds with high probability
- for linear f, posterior concentration holds using Brascamp-Lieb theory

Concluding remarks

- interactive lower bounds are more challenging to establish, while we still have the counterparts of two-point and Fano
- when the rewards are observable, via a two-point argument, constrained DEC gives the right complexity up to a factor of $\mathsf{Est}(\mathcal{M})$
- the Fano-type argument could derive a complicated interactive learning trajectory, suggesting the difficulty of closing the gap of $\mathsf{Est}(\mathcal{M})$ in general

Concluding remarks

- interactive lower bounds are more challenging to establish, while we still have the counterparts of two-point and Fano
- when the rewards are observable, via a two-point argument, constrained DEC gives the right complexity up to a factor of $\mathrm{Est}(\mathcal{M})$
- the Fano-type argument could derive a complicated interactive learning trajectory, suggesting the difficulty of closing the gap of $\mathrm{Est}(\mathcal{M})$ in general

Thank You!