Beyond UCB: statistical complexity and optimal algorithm for non-linear ridge bandits

Yanjun Han (MIT IDSS)

Joint work with:

Jiantao Jiao	Berkeley EECS
Nived Rajaraman	Berkeley EECS
Kannan Ramchandran	Berkeley EECS

AMCS Colloquium, UPenn
October 21, 2022

General setting of stochastic bandit

Input parameters:

- parameter set Θ
- action space $\mathcal{A}$
- reward function class $\mathcal{F} = (f_\theta)_{\theta \in \Theta}$
- time horizon T

General setting of stochastic bandit

Input parameters:

- parameter set Θ
- action space $\mathcal{A}$
- reward function class $\mathcal{F} = (f_\theta)_{\theta \in \Theta}$
- time horizon T

Stochastic bandit environment:

- nature chooses $\theta^\star \in \Theta$, fixed across time and unknown to the learner
- at time $t = 1, \cdots, T$, learner chooses action $a_t \in \mathcal{A}$ and observes a random reward r_t with $\mathbb{E}[r_t \mid a_t = a] = f_{\theta^\star}(a)$
- learner aims to minimize the worst-case (pseudo) regret

$$\text{MinmaxReg}(\Theta, \mathcal{A}, \mathcal{F}, T) = \inf_{a^T} \sup_{\theta^\star \in \Theta} \mathbb{E}_{\theta^\star} \left[T \cdot \max_{a \in \mathcal{A}} f_{\theta^\star}(a) - \sum_{t=1}^{T} f_{\theta^\star}(a_t) \right].$$

General setting of stochastic bandit

Input parameters:

- parameter set Θ
- action space $\mathcal{A}$
- reward function class $\mathcal{F} = (f_\theta)_{\theta \in \Theta}$
- time horizon T

Stochastic bandit environment:

- nature chooses $\theta^\star \in \Theta$, fixed across time and unknown to the learner
- at time $t = 1, \cdots, T$, learner chooses action $a_t \in \mathcal{A}$ and observes a random reward r_t with $\mathbb{E}[r_t \mid a_t = a] = f_{\theta^\star}(a)$
- learner aims to minimize the worst-case (pseudo) regret

$$\text{MinmaxReg}(\Theta, \mathcal{A}, \mathcal{F}, T) = \inf_{a^T} \sup_{\theta^\star \in \Theta} \mathbb{E}_{\theta^\star} \left[T \cdot \max_{a \in \mathcal{A}} f_{\theta^\star}(a) - \sum_{t=1}^{T} f_{\theta^\star}(a_t) \right].$$

Linear bandit

$f_\theta(a) = \langle \theta, \phi(a) \rangle$ with a known feature map $\phi : \mathcal{A} \to \mathbb{R}^d$

Non-linear bandit: a motivating example

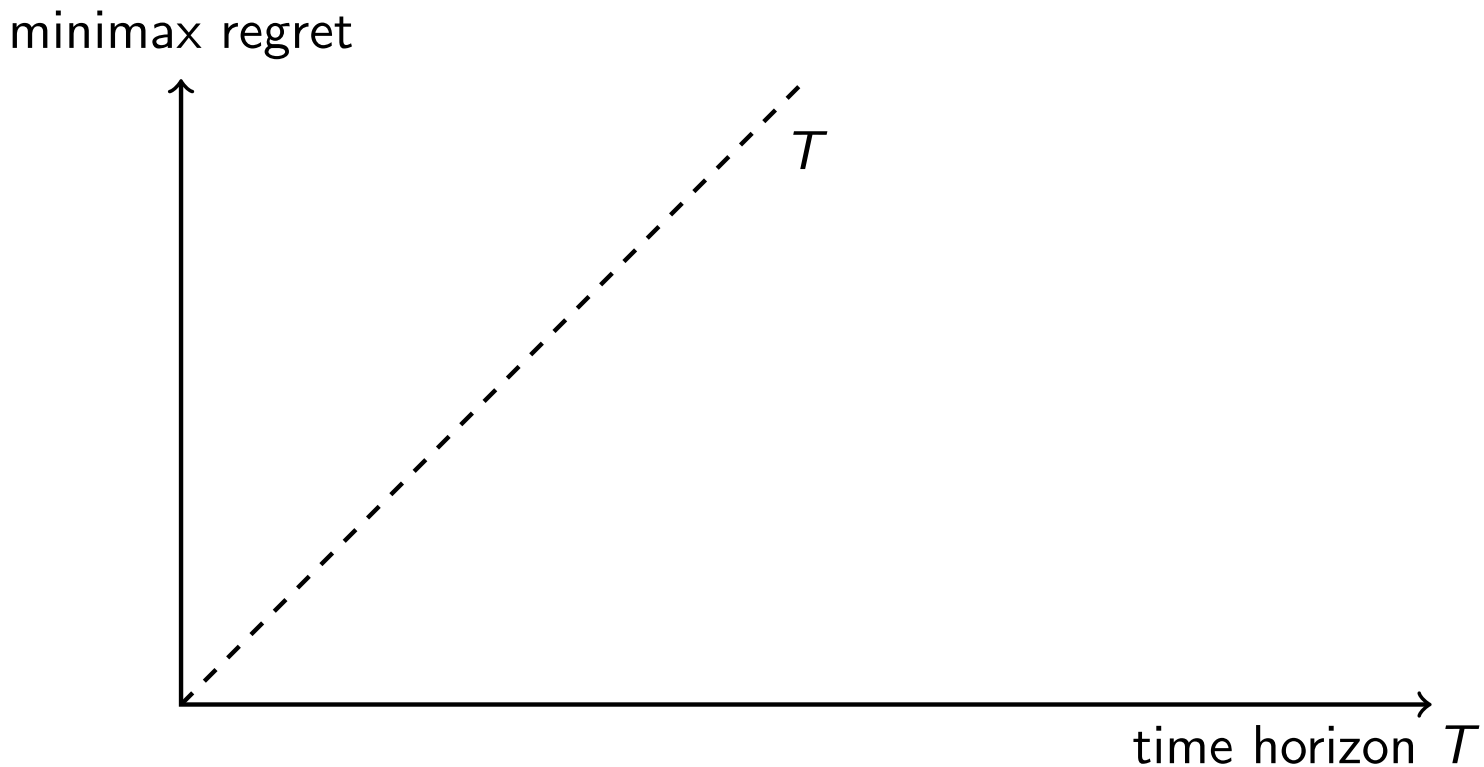
A non-linear bandit example

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

Non-linear bandit: a motivating example

A non-linear bandit example

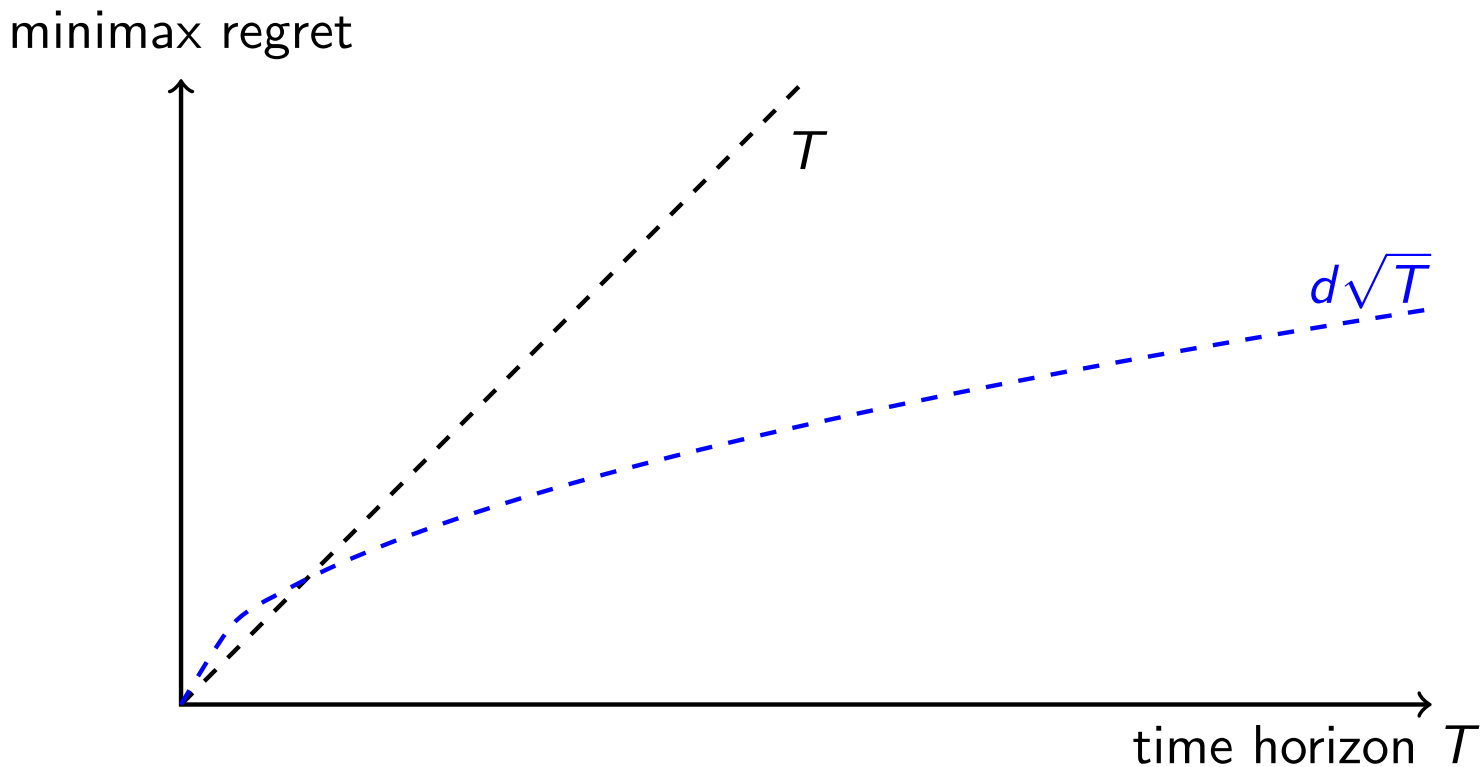
$$f_\theta(a) = \langle \theta, a \rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



Non-linear bandit: a motivating example

A non-linear bandit example

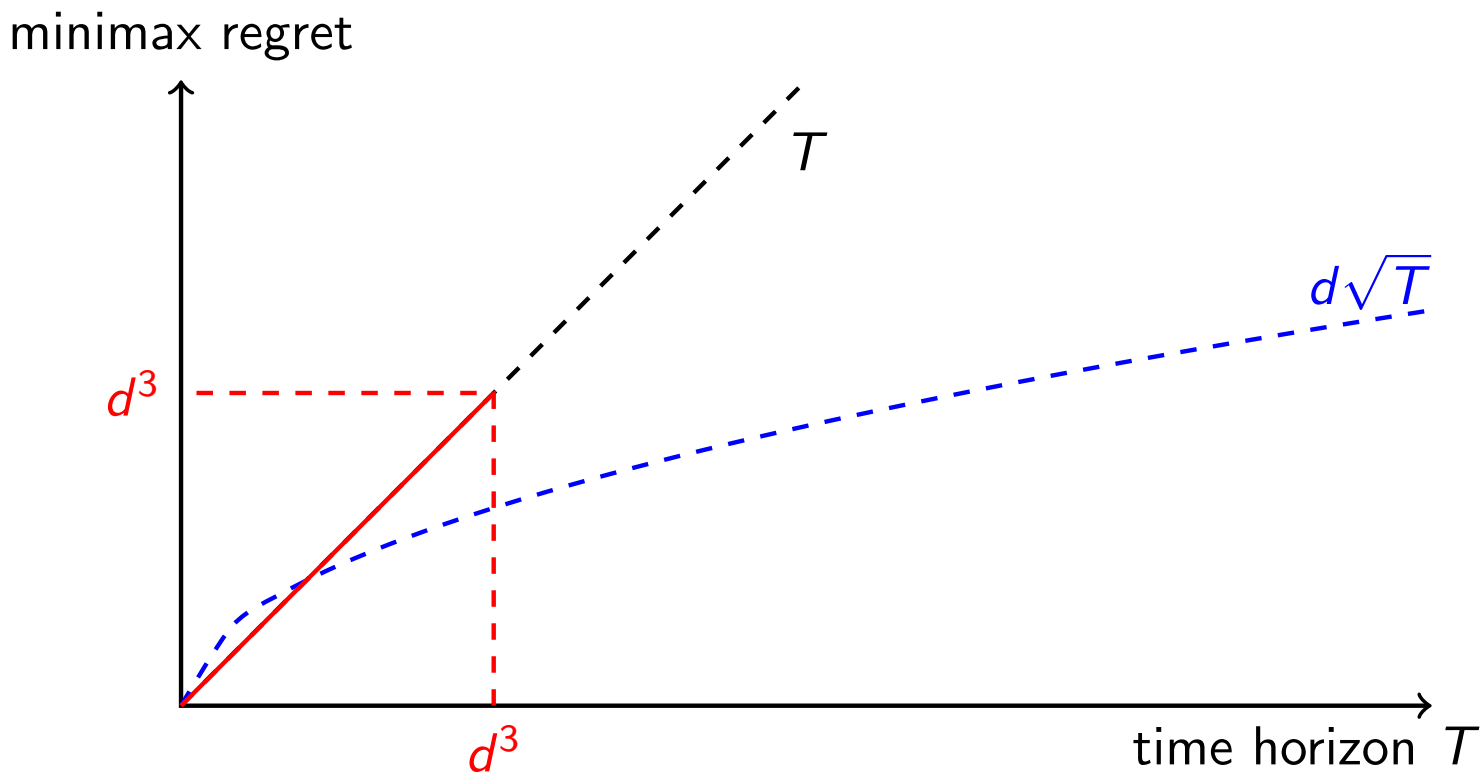
$$f_\theta(a) = \langle \theta, a \rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



Non-linear bandit: a motivating example

A non-linear bandit example

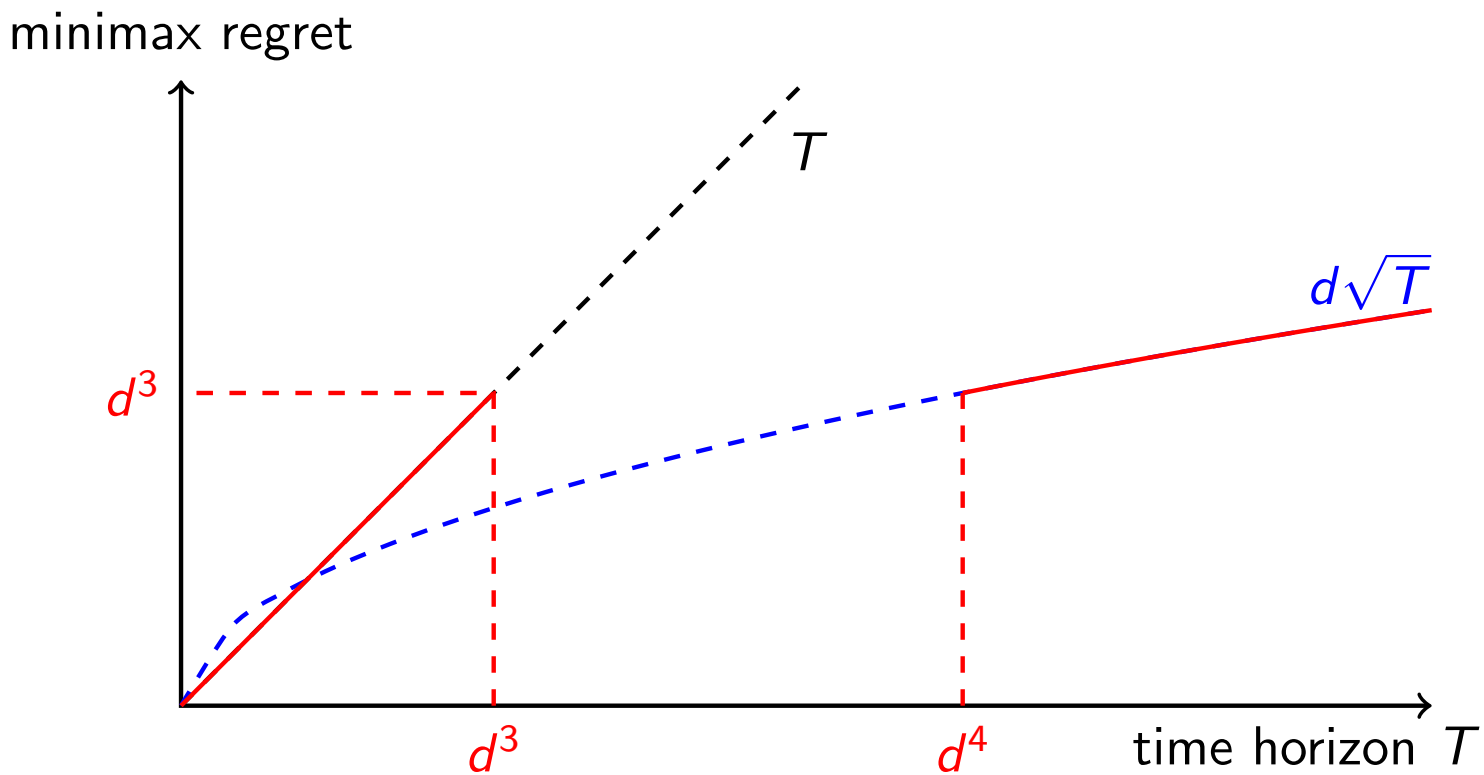
$$f_\theta(a) = \langle \theta, a \rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



Non-linear bandit: a motivating example

A non-linear bandit example

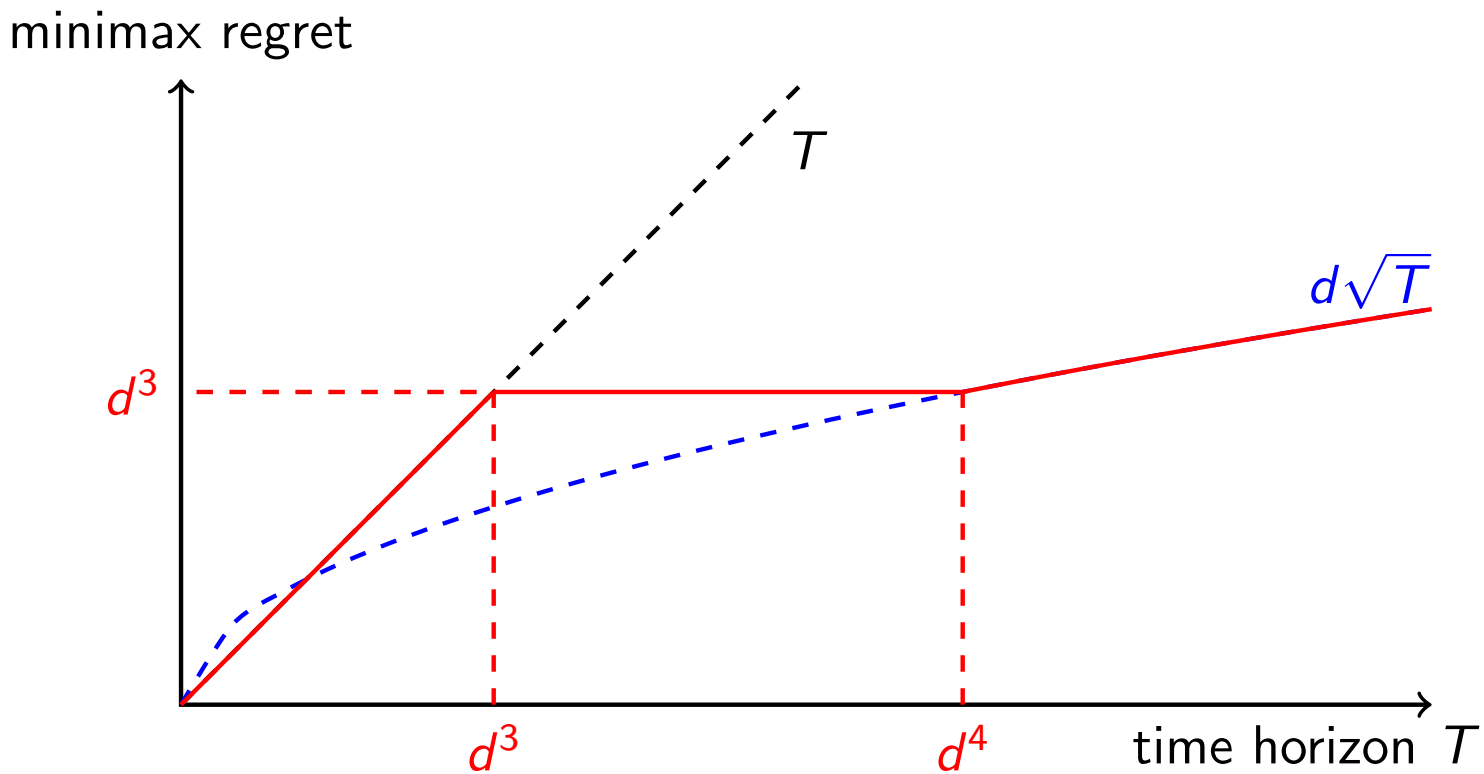
$$f_\theta(a) = \langle \theta, a \rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



Non-linear bandit: a motivating example

A non-linear bandit example

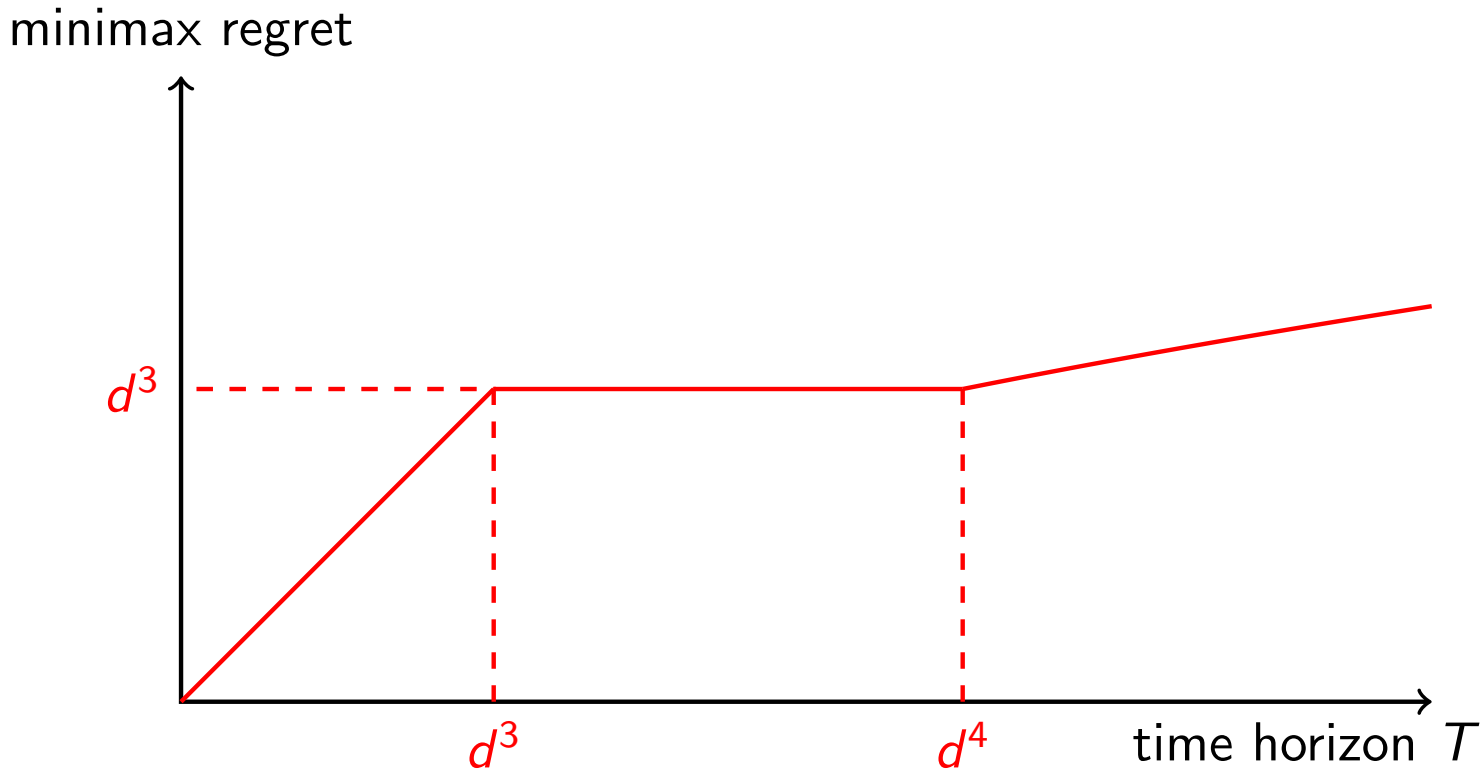
$$f_\theta(a) = \langle \theta, a \rangle^3 : \qquad \theta \in \mathbb{S}^{d-1}, \quad a \in \mathbb{B}^d.$$



Non-linear bandit: a motivating example

A non-linear bandit example

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

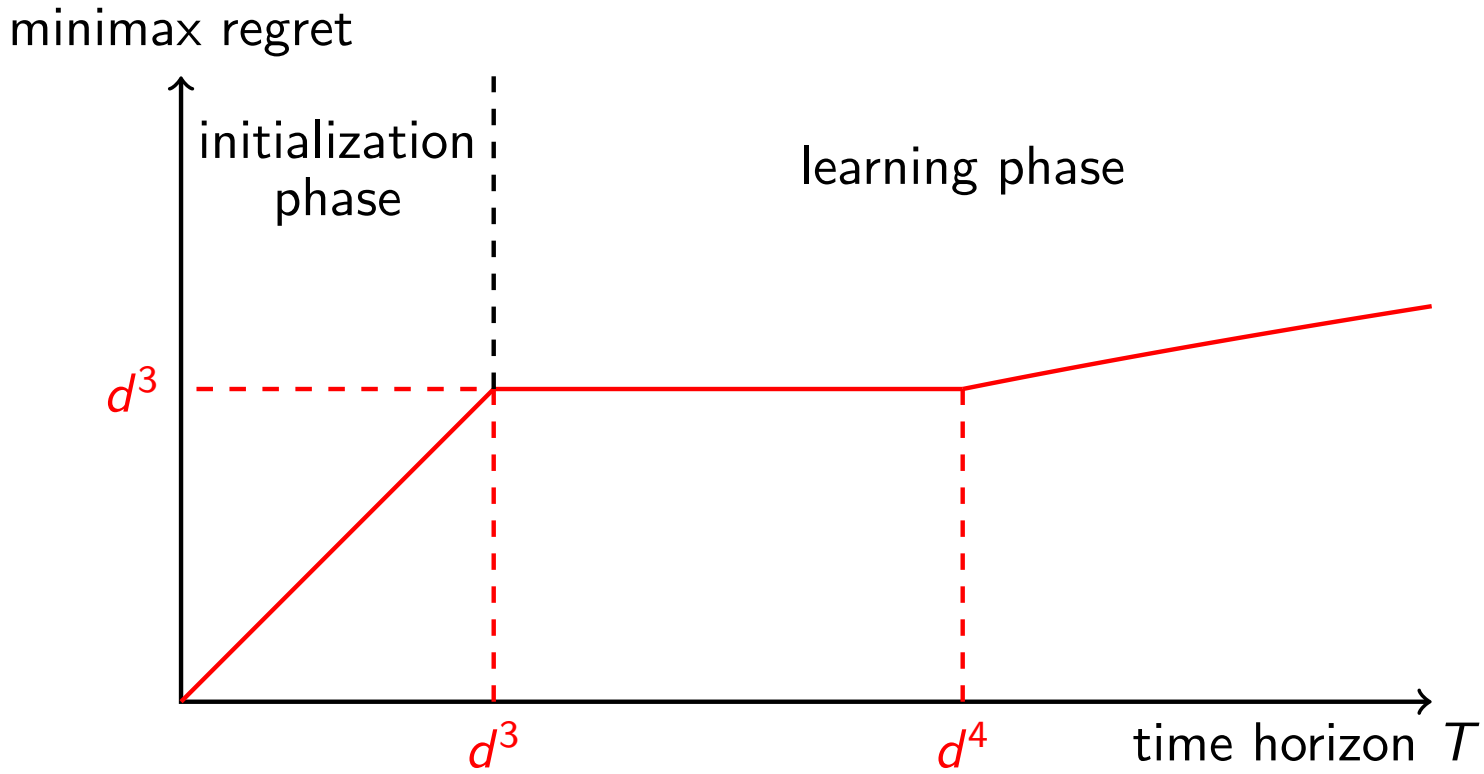


$$\text{MinmaxReg}(T, d) \asymp \min\{T, d^3 + d\sqrt{T}\}.$$

Non-linear bandit: a motivating example

A non-linear bandit example

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



$$\mathrm{MinmaxReg}(T, d) \asymp \min\{T, d^3 + d\sqrt{T}\}.$$

Curious phenomena

Curious phenomena in non-linear bandits:

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

Curious phenomena

Curious phenomena in non-linear bandits:

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

Aim of this talk:

Questions

Curious phenomena

Curious phenomena in non-linear bandits:

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

Aim of this talk:

Questions

- what is the optimal fixed cost in the initialization phase?

Curious phenomena

Curious phenomena in non-linear bandits:

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

Aim of this talk:

Questions

- what is the optimal fixed cost in the initialization phase?
- what algorithms should we use in different phases?

Curious phenomena

Curious phenomena in non-linear bandits:

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

Aim of this talk:

Questions

- what is the optimal fixed cost in the initialization phase?
- what algorithms should we use in different phases?
- how to explore when learner has not started learning?

Plan of this talk

- setting and main results
- proof of upper bound
- proof of lower bound
- discussions and extensions

Setting: non-linear ridge bandits

- parameter space $\Theta = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$
- action space $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$
- reward function $f_\theta(a) = f(\langle \theta, a \rangle)$ with a known link function f

Setting: non-linear ridge bandits

- parameter space $\Theta = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$
- action space $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$
- reward function $f_\theta(a) = f(\langle \theta, a \rangle)$ with a known link function f

Assumptions

Setting: non-linear ridge bandits

- parameter space $\Theta = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$
- action space $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$
- reward function $f_\theta(a) = f(\langle \theta, a \rangle)$ with a known link function f

Assumptions

- monotonicity: $f : [-1, 1] \to [-1, 1]$ is increasing (or $f(-x) = f(x)$ and f is increasing on $[0, 1]$) with $f(0) = 0, f(1) \asymp 1$

  $\to$ best action is $a = \theta^\star$

Setting: non-linear ridge bandits

- parameter space $\Theta = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$
- action space $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$
- reward function $f_\theta(a) = f(\langle \theta, a \rangle)$ with a known link function f

Assumptions

- monotonicity: $f : [-1, 1] \to [-1, 1]$ is increasing (or $f(-x) = f(x)$ and f is increasing on $[0, 1]$) with $f(0) = 0, f(1) \asymp 1$
  $\to$ best action is $a = \theta^\star$
- local linearity near 1: $\max_{x \in [0.1,1]} f'(x) / \min_{x \in [0.1,1]} f'(x) \leq c < \infty$
  $\to$ essentially linear reward when $\langle \theta^\star, a \rangle$ becomes large

Literature review

Literature review

Ridge bandit $f_\theta(a) = f(\langle \theta, a \rangle)$:

- 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 achieved by noisy gradient method [Huang et al. 2021]

Literature review

Ridge bandit $f_\theta(a) = f(\langle \theta, a \rangle)$:

- 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 achieved by noisy gradient method [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

Main Results

Main results: minimax regret

Main results: minimax regret

Theorem (main upper bound, informal)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \le y \le x} f'(y)^2} + d\sqrt{T}, T \right\}.$$

Main results: minimax regret

Theorem (main upper bound, informal)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \le y \le x} f'(y)^2} + d\sqrt{T}, T \right\}.$$

- a useful corollary:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{f'(x)^2} + d\sqrt{T}, T \right\}.$$

Main results: minimax regret

Theorem (main upper bound, informal)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \le y \le x} f'(y)^2} + d\sqrt{T}, T \right\}.$$

- a useful corollary:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{f'(x)^2} + d\sqrt{T}, T \right\}.$$

- the formal version:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \sum_{m=1}^{d/4} \frac{1}{\max\limits_{0 \le y \le \sqrt{m/d}} \min\limits_{z \in [y, 2y]} (f(z + 1/\sqrt{d}) - f(z))^2} + d\sqrt{T}.$$

Main results: minimax regret

Theorem (main upper bound, informal)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \le y \le x} f'(y)^2} + d\sqrt{T}, T \right\}.$$

Theorem (main lower bound)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \gtrsim \min\left\{ d \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{f(x)^2} + d\sqrt{T}, T \right\}.$$

Main results: minimax regret

Theorem (main upper bound, informal)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \le y \le x} f'(y)^2} + d\sqrt{T}, T \right\}.$$

Theorem (main lower bound)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \gtrsim \min\left\{ d \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{f(x)^2} + d\sqrt{T}, T \right\}.$$

- both results within poly-logarithmic factors
- pointwise upper and lower bounds
- fixed cost depends on the entire function f

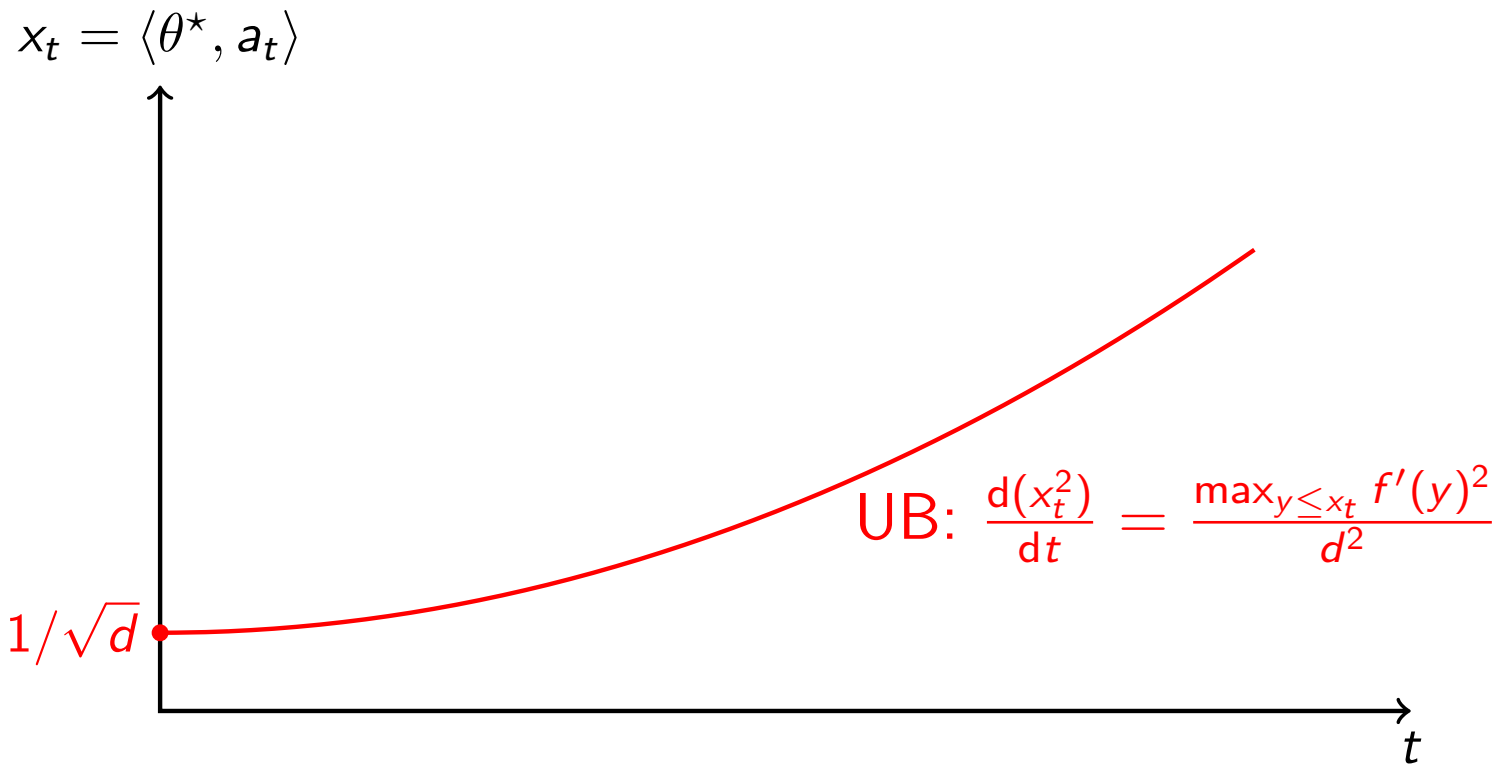
Main results: learning trajectory in the initialization phase

Theorem (learning trajectory)

Main results: learning trajectory in the initialization phase

Theorem (learning trajectory)

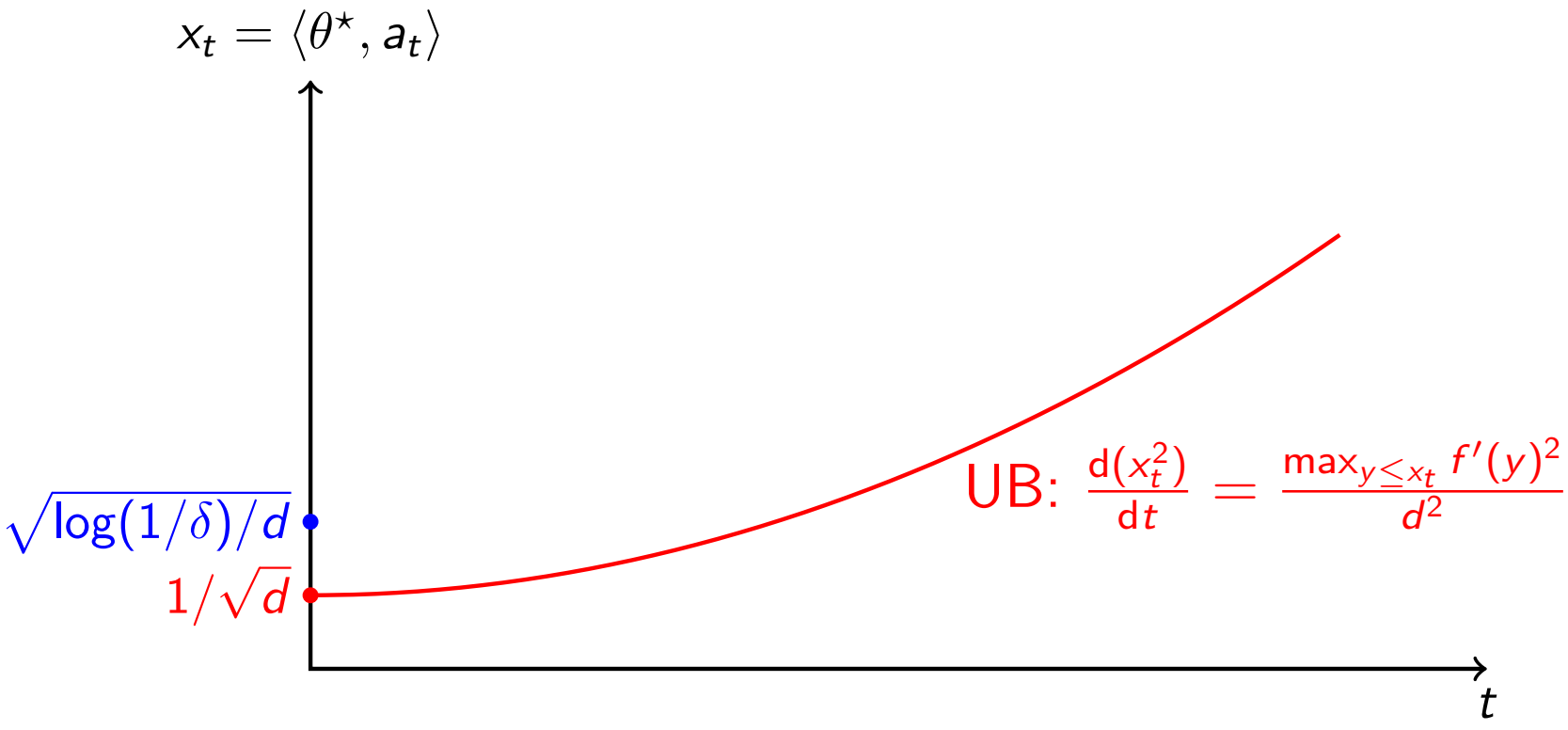
Main results: learning trajectory in the initialization phase



Theorem (learning trajectory)

- there is an algorithm attaining the UB learning curve

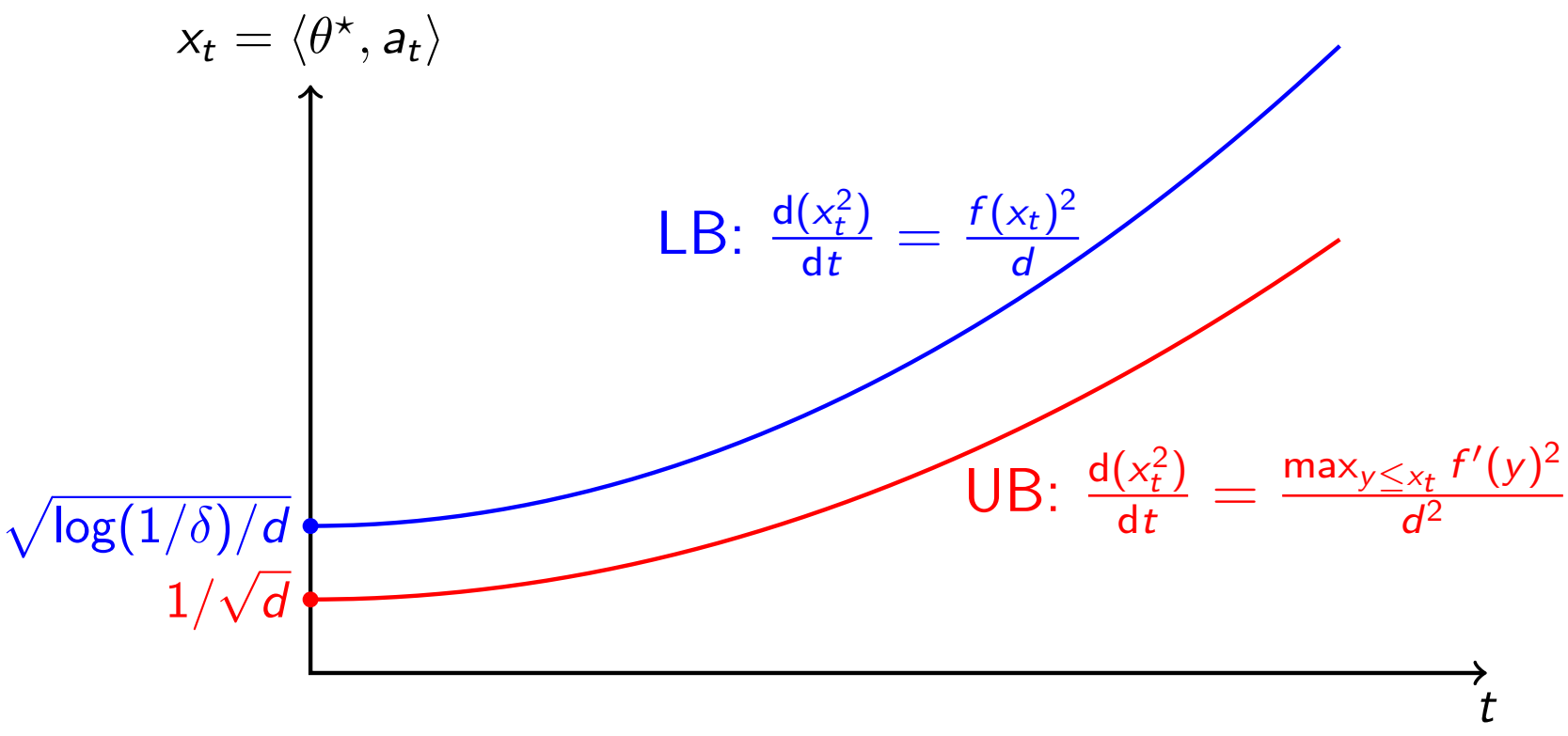
Main results: learning trajectory in the initialization phase



Theorem (learning trajectory)

- there is an algorithm attaining the UB learning curve

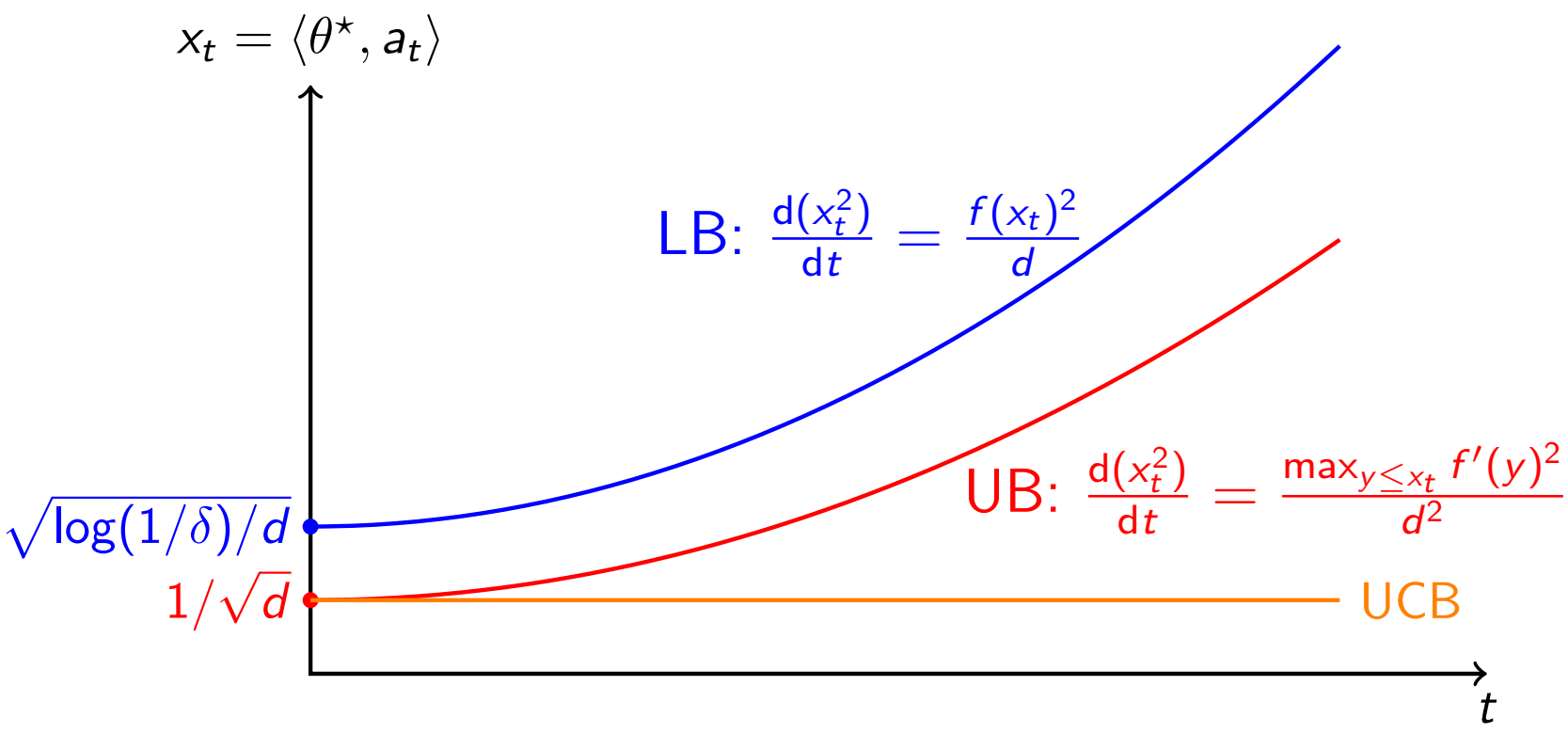
Main results: learning trajectory in the initialization phase



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 $\theta^\star \sim \mathrm{Unif}(\mathbb{S}^{d-1})$

Main results: learning trajectory in the initialization phase



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 algorithm makes no progress whenever $t < d/f(1/\sqrt{d})^2$!

Examples

- polynomial bandit $f(x) = x^p$:

$$\mathrm{MinmaxReg} \asymp \begin{cases} \min\{d\sqrt{T}, T\} & \text{if } 0 < p \leq 2, \\ \min\{d\sqrt{T} + d^p, T\} & \text{if } p > 2. \end{cases}$$

  $\rightarrow$ both Eluder-UCB and information-directed sampling give an additional $O(d^{p+1})$ term when $p > 1$

Examples

- polynomial bandit $f(x) = x^p$:

$$\mathrm{MinmaxReg} \asymp \begin{cases} \min\{d\sqrt{T}, T\} & \text{if } 0 < p \leq 2, \\ \min\{d\sqrt{T} + d^p, T\} & \text{if } p > 2. \end{cases}$$

  $\rightarrow$ both Eluder-UCB and information-directed sampling give an additional $O(d^{p+1})$ term when $p > 1$

- ReLU bandit $f(x) = (x - 0.1)_+$: $T = e^{\Omega(d)}$ is necessary for sublinear regret

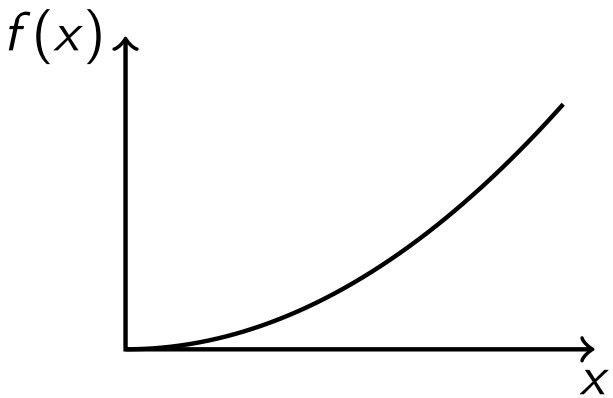
Examples

- polynomial bandit $f(x) = x^p$:

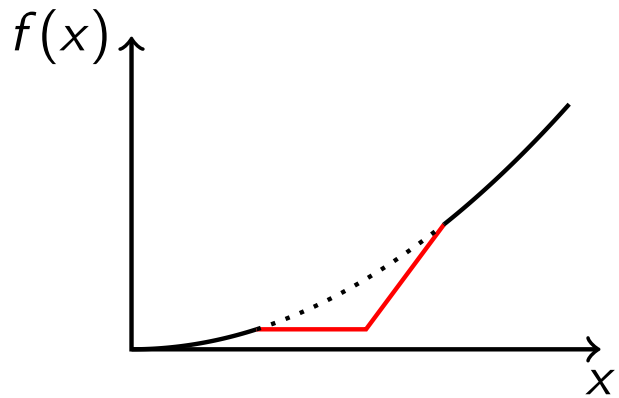
$$\mathrm{MinmaxReg} \asymp \begin{cases} \min\{d\sqrt{T}, T\} & \text{if } 0 < p \le 2, \\ \min\{d\sqrt{T} + d^p, T\} & \text{if } p > 2. \end{cases}$$

  $\rightarrow$ both Eluder-UCB and information-directed sampling give an additional $O(d^{p+1})$ term when $p > 1$

- ReLU bandit $f(x) = (x - 0.1)_+$: $T = e^{\Omega(d)}$ is necessary for sublinear regret
- importance of f at every point:



fixed cost $\asymp d^2$



fixed cost $\asymp d^3$

Upper Bounds

Upper bound: learning phase

Key feature in the learning phase

The learner has found a good "initial action" a_0 such that $\langle a_0, \theta^\star \rangle \geq \text{const}$.

Upper bound: learning phase

Key feature in the learning phase

The learner has found a good "initial action" a_0 such that $\langle a_0, \theta^\star \rangle \geq \text{const}$.

A simple explore-then-commit (ETC) algorithm:

Upper bound: learning phase

Key feature in the learning phase

The learner has found a good "initial action" a_0 such that $\langle a_0, \theta^\star \rangle \geq \text{const}$.

A simple explore-then-commit (ETC) algorithm:

- for the first m rounds, uniformly explore the following $2d$ directions:

$$\left\{ \lambda a_0 \pm \sqrt{1-\lambda^2} e_1, \cdots, \lambda a_0 \pm \sqrt{1-\lambda^2} e_d \right\}, \quad \lambda = \lambda(\text{const});$$

Upper bound: learning phase

Key feature in the learning phase

The learner has found a good "initial action" a_0 such that $\langle a_0, \theta^\star \rangle \geq \text{const}$.

A simple explore-then-commit (ETC) algorithm:

- for the first m rounds, uniformly explore the following $2d$ directions:

$$\left\{ \lambda a_0 \pm \sqrt{1-\lambda^2} e_1, \cdots, \lambda a_0 \pm \sqrt{1-\lambda^2} e_d \right\}, \quad \lambda = \lambda(\text{const});$$

- find the least squares estimate of $\theta^\star$:

$$\widehat{\theta} = \arg \min_{\theta : \langle \theta, a_0 \rangle \geq \text{const}} \frac{1}{2} \sum_{t=1}^{m} (r_m - f(\langle \theta, a_t \rangle))^2;$$

Upper bound: learning phase

Key feature in the learning phase

The learner has found a good "initial action" a_0 such that $\langle a_0, \theta^\star \rangle \geq \text{const}$.

A simple explore-then-commit (ETC) algorithm:

- for the first m rounds, uniformly explore the following $2d$ directions:

$$\left\{ \lambda a_0 \pm \sqrt{1-\lambda^2} e_1, \cdots, \lambda a_0 \pm \sqrt{1-\lambda^2} e_d \right\}, \quad \lambda = \lambda(\text{const});$$

- find the least squares estimate of $\theta^\star$:

$$\widehat{\theta} = \arg \min_{\theta : \langle \theta, a_0 \rangle \geq \text{const}} \frac{1}{2} \sum_{t=1}^{m} (r_m - f(\langle \theta, a_t \rangle))^2;$$

- for the remaining rounds, greedily play $a_t = \widehat{\theta}$.

Analysis of learning phase

- standard least squares analysis gives w.h.p.

$$\sum_{t=1}^{m}(f(\langle\widehat{\theta}, a_t\rangle) - f(\langle\theta^\star, a_t\rangle))^2 = \widetilde{O}(d);$$

Analysis of learning phase

- standard least squares analysis gives w.h.p.

$$\sum_{t=1}^{m}(f(\langle\widehat{\theta}, a_t\rangle) - f(\langle\theta^\star, a_t\rangle))^2 = \widetilde{O}(d);$$

- local linearity of f near 1 implies that

$$\|\widehat{\theta} - \theta^\star\|_2^2 = \widetilde{O}\left(\frac{d^2}{m \cdot f'(1)^2}\right);$$

Analysis of learning phase

- standard least squares analysis gives w.h.p.

$$\sum_{t=1}^{m}(f(\langle\widehat{\theta}, a_t\rangle) - f(\langle\theta^\star, a_t\rangle))^2 = \widetilde{O}(d);$$

- local linearity of f near 1 implies that

$$\|\widehat{\theta} - \theta^\star\|_2^2 = \widetilde{O}\left(\frac{d^2}{m \cdot f'(1)^2}\right);$$

- instantaneous regret when greedily plays $\widehat{\theta}$:

$$f(1) - f(\langle\theta^\star, \widehat{\theta}\rangle) \lesssim f'(1)(1 - \langle\theta^\star, \widehat{\theta}\rangle) \lesssim \frac{d^2}{m \cdot f'(1)};$$

Analysis of learning phase

- standard least squares analysis gives w.h.p.

$$\sum_{t=1}^{m}(f(\langle\widehat{\theta}, a_t\rangle) - f(\langle\theta^\star, a_t\rangle))^2 = \widetilde{O}(d);$$

- local linearity of f near 1 implies that

$$\|\widehat{\theta} - \theta^\star\|_2^2 = \widetilde{O}\left(\frac{d^2}{m \cdot f'(1)^2}\right);$$

- instantaneous regret when greedily plays $\widehat{\theta}$:

$$f(1) - f(\langle\theta^\star, \widehat{\theta}\rangle) \lesssim f'(1)(1 - \langle\theta^\star, \widehat{\theta}\rangle) \lesssim \frac{d^2}{m \cdot f'(1)};$$

- total regret in the learning phase:

$$m \cdot f'(1) + (T - m) \cdot \frac{d^2}{m \cdot f'(1)} \overset{m \asymp d\sqrt{T}/f'(1)}{\asymp} d\sqrt{T}.$$

Upper bound: initialization phase

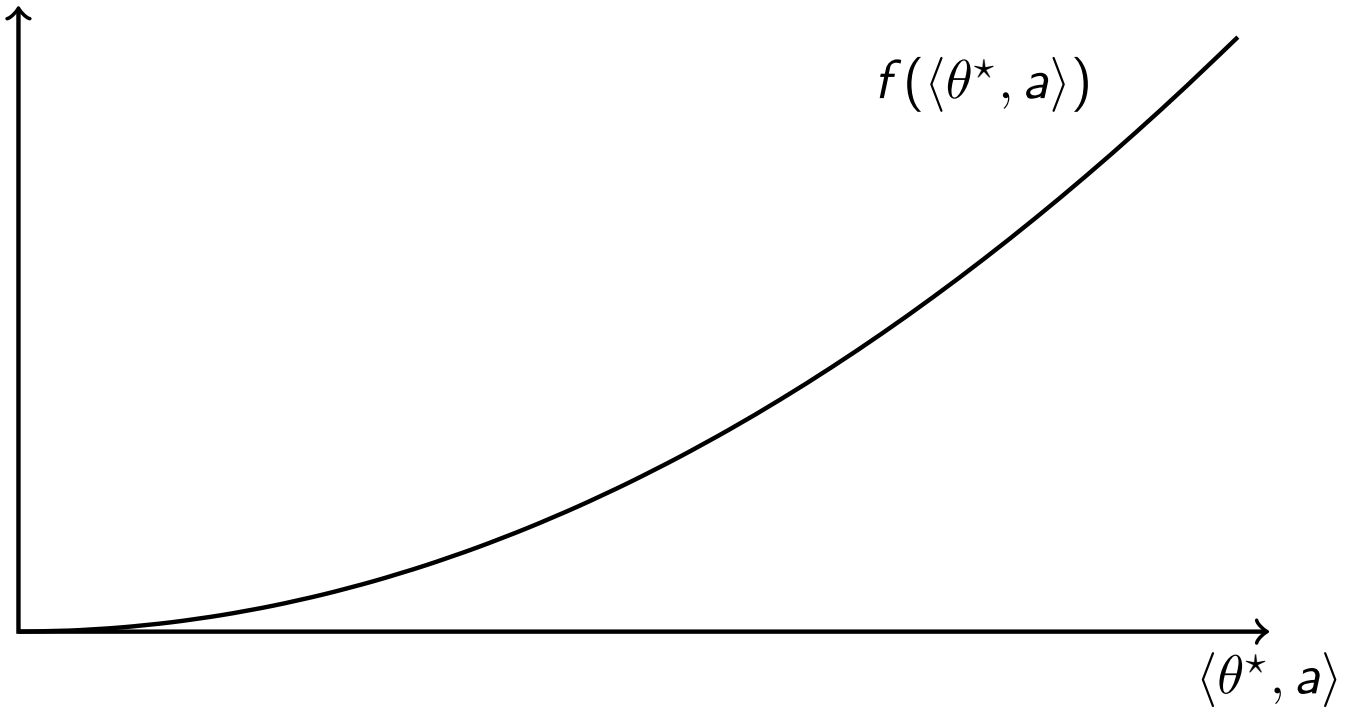
Target in the initialization phase

Find a good "initial action" a_0 with a large inner product $\langle a_0, \theta^\star \rangle \geq \text{const}$.

Upper bound: initialization phase

Target in the initialization phase

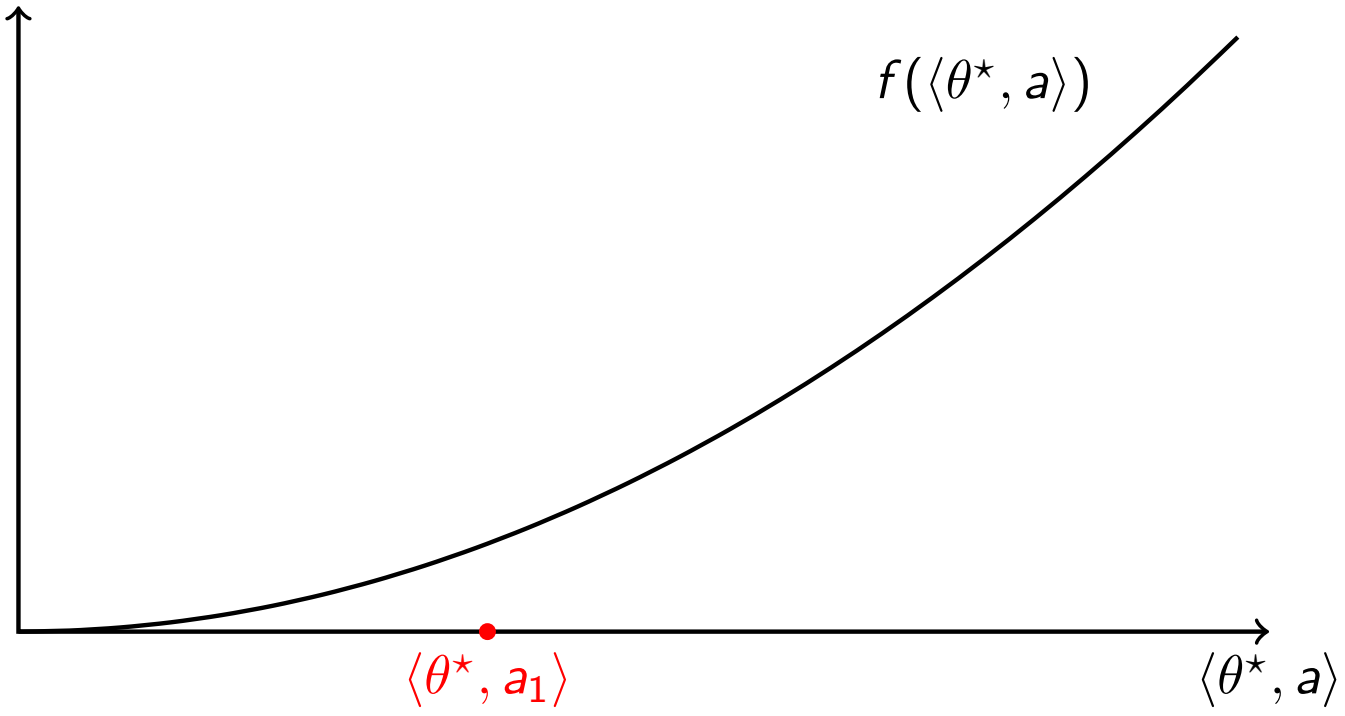
Find a good "initial action" a_0 with a large inner product $\langle a_0, \theta^\star \rangle \geq \text{const}$.



Upper bound: initialization phase

Target in the initialization phase

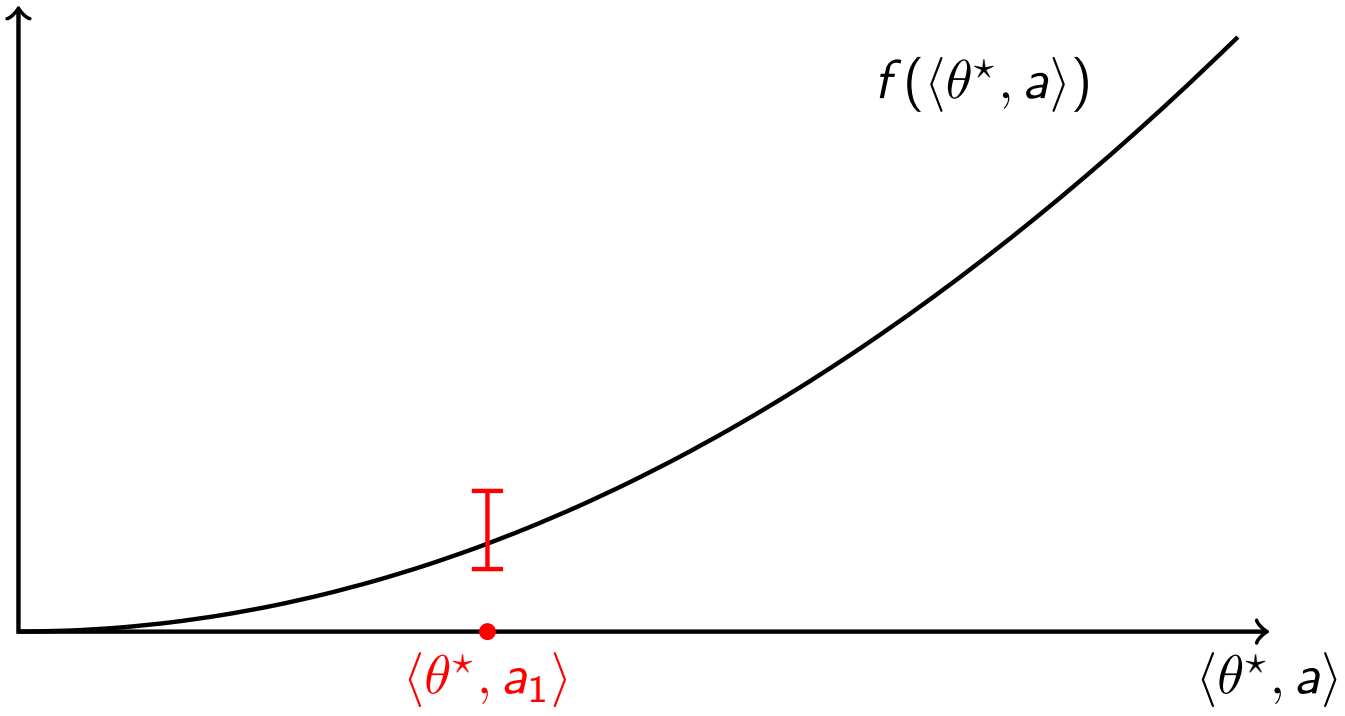
Find a good "initial action" a_0 with a large inner product $\langle a_0, \theta^\star \rangle \geq \text{const}$.



Upper bound: initialization phase

Target in the initialization phase

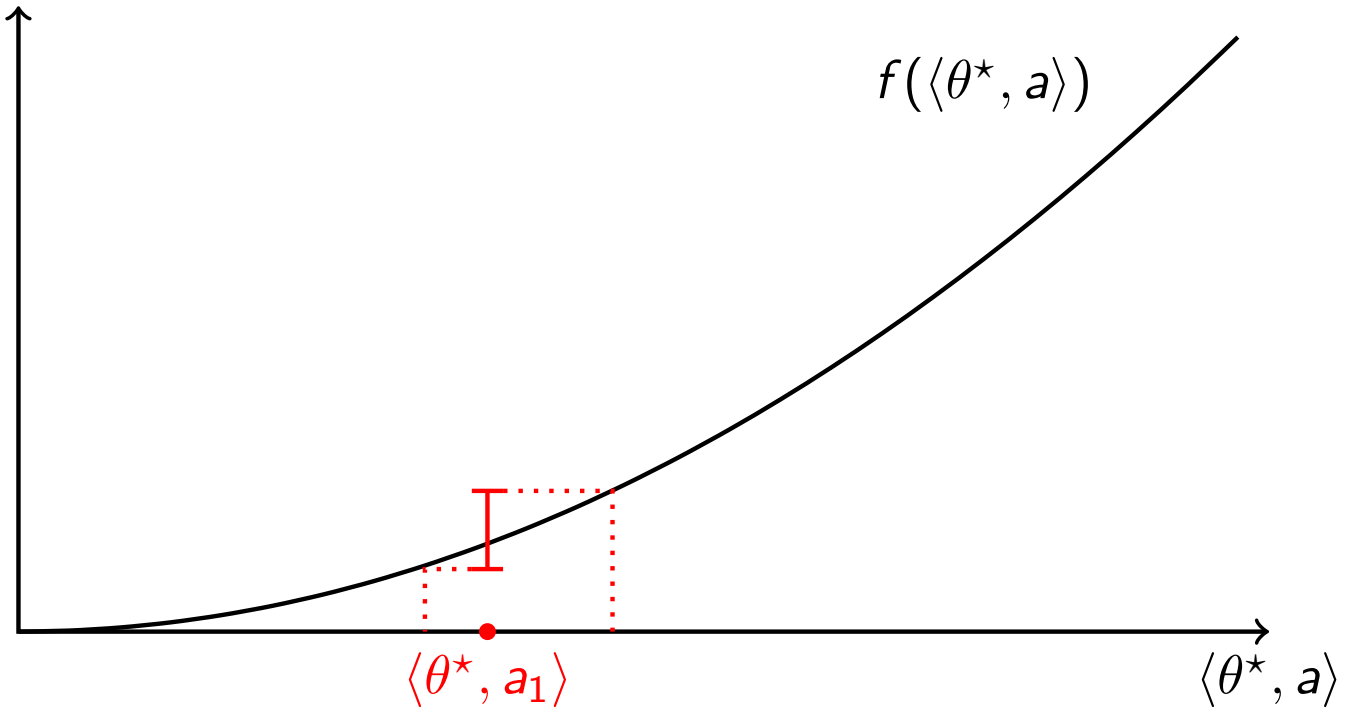
Find a good "initial action" a_0 with a large inner product $\langle a_0, \theta^\star \rangle \geq \text{const}$.



Upper bound: initialization phase

Target in the initialization phase

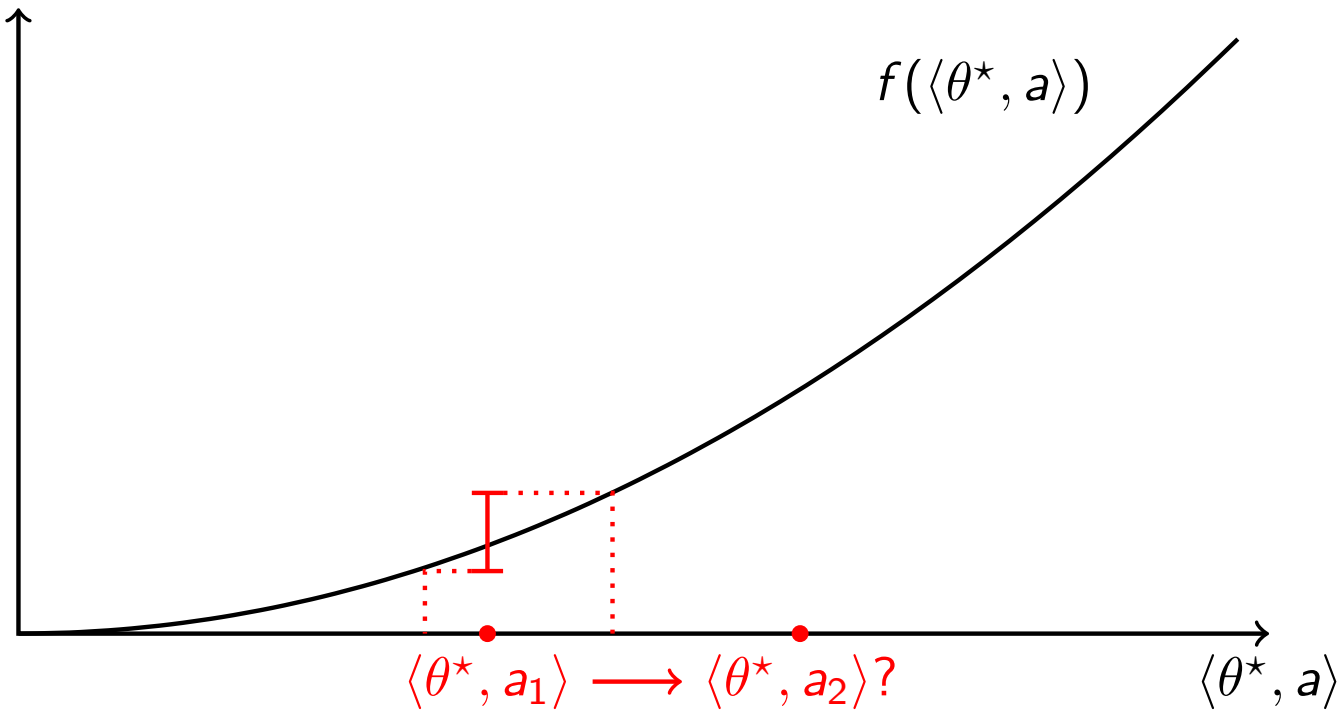
Find a good "initial action" a_0 with a large inner product $\langle a_0, \theta^\star \rangle \geq \text{const}$.



Upper bound: initialization phase

Target in the initialization phase

Find a good "initial action" a_0 with a large inner product $\langle a_0, \theta^\star \rangle \geq \text{const}$.



Certify that $\langle \theta^\star, a \rangle \in [r - \delta, r + \delta]$ can be done with $\widetilde{O}(1/[\delta f'(r)]^2)$ samples

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$
- for proper λ, if $\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]$, then $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} = \sqrt{x_{\text{pre}}^2 + 1/d}$

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$
- for proper λ, if $\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]$, then $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} = \sqrt{x_{\text{pre}}^2 + 1/d}$
- exploration: easy, as $\mathbb{P}(\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]) = \Omega(1)$ for uniform a

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$
- for proper λ, if $\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]$, then $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} = \sqrt{x_{\text{pre}}^2 + 1/d}$
- exploration: easy, as $\mathbb{P}(\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]) = \Omega(1)$ for uniform a
- certification: should make use of a_{pre}!

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$
- for proper λ, if $\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]$, then $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} = \sqrt{x_{\text{pre}}^2 + 1/d}$
- exploration: easy, as $\mathbb{P}(\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]) = \Omega(1)$ for uniform a
- certification: should make use of a_{pre}!

$$\frac{\langle \theta^\star, a \rangle}{\sqrt{2}} = \left\langle \theta^\star, \frac{a + a_{\text{pre}}}{\sqrt{2}} \right\rangle - \left\langle \theta^\star, \frac{a_{\text{pre}}}{\sqrt{2}} \right\rangle$$

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$
- for proper λ, if $\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]$, then $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} = \sqrt{x_{\text{pre}}^2 + 1/d}$
- exploration: easy, as $\mathbb{P}(\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]) = \Omega(1)$ for uniform a
- certification: should make use of a_{pre}!

$$\frac{\langle \theta^\star, a \rangle}{\sqrt{2}} = \left\langle \theta^\star, \frac{a + a_{\text{pre}}}{\sqrt{2}} \right\rangle - \left\langle \theta^\star, \frac{a_{\text{pre}}}{\sqrt{2}} \right\rangle$$

$\rightarrow$ each terms uses $\widetilde{O}(d/f'(x_{\text{pre}})^2)$ samples for certification

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$
- for proper λ, if $\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]$, then $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} = \sqrt{x_{\text{pre}}^2 + 1/d}$
- exploration: easy, as $\mathbb{P}(\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]) = \Omega(1)$ for uniform a
- certification: should make use of a_{pre}!

$$\frac{\langle \theta^\star, a \rangle}{\sqrt{2}} = \left\langle \theta^\star, \frac{a + a_{\text{pre}}}{\sqrt{2}} \right\rangle - \left\langle \theta^\star, \frac{a_{\text{pre}}}{\sqrt{2}} \right\rangle$$

$\rightarrow$ each terms uses $\widetilde{O}(d/f'(x_{\text{pre}})^2)$ samples for certification

$\rightarrow$ total sample complexity is roughly $d^2 \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}x^2}{f'(x)^2}$

Exploration and certification

Recursive step

Given an action a_{pre} with $\langle \theta^\star, a_{\text{pre}} \rangle \in [x_{\text{pre}}, 2x_{\text{pre}}]$, where x_{pre} is known, how to find a_{now} and certify that $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} > x_{\text{pre}}$?

- idea: find $a \perp a_{\text{pre}}$ with $\langle \theta^\star, a \rangle \asymp 1/\sqrt{d}$ and play $a_{\text{now}} = \lambda a_{\text{pre}} + \sqrt{1-\lambda^2}\, a$
- for proper λ, if $\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]$, then $\langle \theta^\star, a_{\text{now}} \rangle \in [x_{\text{now}}, 2x_{\text{now}}]$ with $x_{\text{now}} = \sqrt{x_{\text{pre}}^2 + 1/d}$
- exploration: easy, as $\mathbb{P}(\langle \theta^\star, a \rangle \in [1/\sqrt{d}, 2/\sqrt{d}]) = \Omega(1)$ for uniform a
- certification: should make use of a_{pre}!

$$\frac{\langle \theta^\star, a \rangle}{\sqrt{2}} = \left\langle \theta^\star, \frac{a + \mu a_{\text{pre}}}{\sqrt{2}} \right\rangle - \left\langle \theta^\star, \frac{\mu a_{\text{pre}}}{\sqrt{2}} \right\rangle$$

$\rightarrow$ each terms uses $\widetilde{O}(d/\max_{y \le x} f'(y)^2)$ samples for certification;

$\rightarrow$ total sample complexity is roughly $d^2 \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}x^2}{\max_{y \le x} f'(y)^2}$.

Certification details

Target of certification

Given actions a and $a + b$ with $\langle \theta^\star, a \rangle \in [x, 2x]$, find a test which

- outputs "failure" w.h.p. if $\langle \theta^\star, b \rangle \notin [z, 2z]$;
- outputs "success" w.h.p. if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$.

Certification details

Target of certification

Given actions a and $a+b$ with $\langle \theta^\star, a \rangle \in [x, 2x]$, find a test which

- outputs "failure" w.h.p. if $\langle \theta^\star, b \rangle \notin [z, 2z]$;
- outputs "success" w.h.p. if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$.

- pull both actions $\widetilde{O}(1/\delta^2)$ times to obtain

$$|\widehat{f}_1 - f(\langle \theta^\star, a \rangle)| \leq \delta, \qquad |\widehat{f}_2 - f(\langle \theta^\star, a+b \rangle)| \leq \delta;$$

Certification details

Target of certification

Given actions a and $a+b$ with $\langle \theta^\star, a \rangle \in [x, 2x]$, find a test which

- outputs "failure" w.h.p. if $\langle \theta^\star, b \rangle \notin [z, 2z]$;
- outputs "success" w.h.p. if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$.

- pull both actions $\widetilde{O}(1/\delta^2)$ times to obtain

$$|\widehat{f}_1 - f(\langle \theta^\star, a \rangle)| \le \delta, \qquad |\widehat{f}_2 - f(\langle \theta^\star, a+b \rangle)| \le \delta;$$

- test returns "success" iff $\exists u \in [x, 2x], v \in [1.2z, 1.8z]$ such that $|\widehat{f}_1 - f(u)| \le \delta$ and $|\widehat{f}_2 - f(u+v)| \le \delta$:

Certification details

Target of certification

Given actions a and $a+b$ with $\langle \theta^\star, a \rangle \in [x, 2x]$, find a test which

- outputs "failure" w.h.p. if $\langle \theta^\star, b \rangle \notin [z, 2z]$;
- outputs "success" w.h.p. if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$.

- pull both actions $\widetilde{O}(1/\delta^2)$ times to obtain

$$|\widehat{f}_1 - f(\langle \theta^\star, a \rangle)| \le \delta, \qquad |\widehat{f}_2 - f(\langle \theta^\star, a+b \rangle)| \le \delta;$$

- test returns "success" iff $\exists u \in [x, 2x], v \in [1.2z, 1.8z]$ such that $|\widehat{f}_1 - f(u)| \le \delta$ and $|\widehat{f}_2 - f(u+v)| \le \delta$:
  $\rightarrow$ if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$, then $(u, v) = (\langle \theta^\star, a \rangle, \langle \theta^\star, b \rangle)$ passes the test;

Certification details

Target of certification

Given actions a and $a+b$ with $\langle \theta^\star, a \rangle \in [x, 2x]$, find a test which

- outputs "failure" w.h.p. if $\langle \theta^\star, b \rangle \notin [z, 2z]$;
- outputs "success" w.h.p. if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$.

- pull both actions $\widetilde{O}(1/\delta^2)$ times to obtain

$$|\widehat{f}_1 - f(\langle \theta^\star, a \rangle)| \le \delta, \qquad |\widehat{f}_2 - f(\langle \theta^\star, a+b \rangle)| \le \delta;$$

- test returns "success" iff $\exists u \in [x, 2x], v \in [1.2z, 1.8z]$ such that $|\widehat{f}_1 - f(u)| \le \delta$ and $|\widehat{f}_2 - f(u+v)| \le \delta$:
  - $\rightarrow$ if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$, then $(u, v) = (\langle \theta^\star, a \rangle, \langle \theta^\star, b \rangle)$ passes the test;
  - $\rightarrow$ if $\langle \theta^\star, b \rangle \notin [z, 2z]$, then existence of (u, v) implies

$$|u - \langle \theta^\star, a \rangle| \ge 0.2z, \quad \text{or} \quad |u + v - \langle \theta^\star, a+b \rangle| \ge 0.2z;$$

Certification details

Target of certification

Given actions a and $a+b$ with $\langle \theta^\star, a \rangle \in [x, 2x]$, find a test which

- outputs "failure" w.h.p. if $\langle \theta^\star, b \rangle \notin [z, 2z]$;
- outputs "success" w.h.p. if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$.

- pull both actions $\widetilde{O}(1/\delta^2)$ times to obtain

$$|\widehat{f}_1 - f(\langle \theta^\star, a \rangle)| \le \delta, \qquad |\widehat{f}_2 - f(\langle \theta^\star, a+b \rangle)| \le \delta;$$

- test returns "success" iff $\exists u \in [x, 2x], v \in [1.2z, 1.8z]$ such that $|\widehat{f}_1 - f(u)| \le \delta$ and $|\widehat{f}_2 - f(u+v)| \le \delta$:
  - $\rightarrow$ if $\langle \theta^\star, b \rangle \in [1.2z, 1.8z]$, then $(u, v) = (\langle \theta^\star, a \rangle, \langle \theta^\star, b \rangle)$ passes the test;
  - $\rightarrow$ if $\langle \theta^\star, b \rangle \notin [z, 2z]$, then existence of (u, v) implies

$$|u - \langle \theta^\star, a \rangle| \ge 0.2z, \quad \text{or} \quad |u + v - \langle \theta^\star, a+b \rangle| \ge 0.2z;$$

  - $\rightarrow$ test works if $\delta < \min_{y \in [x, 2x+2z]} [f(y + 0.2z) - f(y)]/2$.

Lower Bounds

Formal statement

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 statement

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

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

Formal statement

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

Formal statement

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
- hard(?) to prove via usual arguments of hypothesis testing

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

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

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

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

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

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

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

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

Applying the insight gives the desired recursion

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

More on the above insights

More on the above insights

- reasoning behind the insight:

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

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:

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

Formal proof via χ^2-informativity

- χ^2-informativity between X and Y:

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

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

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

Formal proof via χ^2-informativity

- χ^2-informativity between X and Y:

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

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

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

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

Formal proof via χ^2-informativity

- χ^2-informativity between X and Y:

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

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

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

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

Conditioning technique

- let $\mathcal{E}_t = \cap_{s \leq t} \{ \langle \theta^\star, a_s \rangle \leq \varepsilon_s \}$ 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:

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

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:

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

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

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:

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

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

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

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:

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

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

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

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

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:

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

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:

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

- continuing this process gives

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

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:

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

- continuing this process gives

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

- 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:

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

- continuing this process gives

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

- 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:

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

- continuing this process gives

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

- recursion of error probability:

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

Discussions and Further Questions

Suboptimality of UCB

Suboptimality of UCB

- Eluder-UCB algorithm [Russo and Van Roy 2014]:

Suboptimality of UCB

- Eluder-UCB algorithm [Russo and Van Roy 2014]: at each time t,
  - $\rightarrow$ form the least squares estimator $\widehat{\theta}_t = \arg\min_\theta \sum_{s<t}(r_s - f(\langle \theta, a_s \rangle))^2$;

Suboptimality of UCB

- Eluder-UCB algorithm [Russo and Van Roy 2014]: at each time t,
  - $\rightarrow$ form the least squares estimator $\widehat{\theta}_t = \arg\min_\theta \sum_{s<t}(r_s - f(\langle \theta, a_s \rangle))^2$;
  - $\rightarrow$ construct the confidence set of $\theta^\star$:

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{S}^{d-1} : \sum_{s<t} (f(\langle \theta, a_s \rangle) - f(\langle \widehat{\theta}_t, a_s \rangle))^2 = \widetilde{O}(d) \right\};$$

Suboptimality of UCB

- Eluder-UCB algorithm [Russo and Van Roy 2014]: at each time t,
  - $\rightarrow$ form the least squares estimator $\widehat{\theta}_t = \arg\min_\theta \sum_{s<t}(r_s - f(\langle \theta, a_s \rangle))^2$;
  - $\rightarrow$ construct the confidence set of $\theta^\star$:

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{S}^{d-1} : \sum_{s<t} (f(\langle \theta, a_s \rangle) - f(\langle \widehat{\theta}_t, a_s \rangle))^2 = \widetilde{O}(d) \right\};$$

  - $\rightarrow$ play action $a_t = \arg\max_a \max_{\theta \in \mathcal{C}_t} f(\langle \theta, a \rangle)$.

Suboptimality of UCB

- Eluder-UCB algorithm [Russo and Van Roy 2014]: at each time t,
  - $\rightarrow$ form the least squares estimator $\widehat{\theta}_t = \arg\min_\theta \sum_{s<t}(r_s - f(\langle \theta, a_s \rangle))^2$;
  - $\rightarrow$ construct the confidence set of $\theta^\star$:

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{S}^{d-1} : \sum_{s<t} (f(\langle \theta, a_s \rangle) - f(\langle \widehat{\theta}_t, a_s \rangle))^2 = \widetilde{O}(d) \right\};$$

  - $\rightarrow$ play action $a_t = \arg\max_a \max_{\theta \in \mathcal{C}_t} f(\langle \theta, a \rangle)$.

Theorem (lower bound for Eluder-UCB)

For every f, there exist a bandit instance such that for (a certain tie-breaking rule of) Eluder-UCB, achieving a sublinear regret requires

$$T \gtrsim \max_K \min \left\{ K, \frac{d}{f(\sqrt{(\log K)/d})^2} \right\}.$$

Suboptimality of UCB

- Eluder-UCB algorithm [Russo and Van Roy 2014]: at each time t,
  - $\rightarrow$ form the least squares estimator $\widehat{\theta}_t = \arg\min_\theta \sum_{s<t}(r_s - f(\langle\theta, a_s\rangle))^2$;
  - $\rightarrow$ construct the confidence set of $\theta^\star$:

$$\mathcal{C}_t = \left\{\theta \in \mathbb{S}^{d-1} : \sum_{s<t}(f(\langle\theta, a_s\rangle) - f(\langle\widehat{\theta}_t, a_s\rangle))^2 = \widetilde{O}(d)\right\};$$

  - $\rightarrow$ play action $a_t = \arg\max_a \max_{\theta\in\mathcal{C}_t} f(\langle\theta, a\rangle)$.

Theorem (lower bound for Eluder-UCB)

For every f, there exist a bandit instance such that for (a certain tie-breaking rule of) Eluder-UCB, achieving a sublinear regret requires

$$T \gtrsim \max_K \min\left\{K, \frac{d}{f(\sqrt{(\log K)/d})^2}\right\}.$$

- for $f(x) = x^3$, Eluder-UCB requires $T \gtrsim d^4$, but optimal is $T \gtrsim d^3$

Suboptimality of online regression oracle

Suboptimality of online regression oracle

- Online regression oracle model [Foster et al. 2020]: for any adversarial sequence $\{a_t\}$, oracle outputs $\{\widehat{\theta}_t\}$ such that

$$\sum_{t=1}^{T}(f(\langle\theta^\star, a_t\rangle) - f(\langle\widehat{\theta}_t, a_t\rangle))^2 \leq \mathrm{Reg}_{\mathrm{Sq}}(T)$$

Suboptimality of online regression oracle

- Online regression oracle model [Foster et al. 2020]: for any adversarial sequence $\{a_t\}$, oracle outputs $\{\widehat{\theta}_t\}$ such that

$$\sum_{t=1}^{T}(f(\langle\theta^\star, a_t\rangle) - f(\langle\widehat{\theta}_t, a_t\rangle))^2 \leq \mathrm{Reg}_{\mathrm{Sq}}(T)$$

  $\rightarrow$ learner only observes $\widehat{\theta}_t$ in the oracle model, but not r_t;

Suboptimality of online regression oracle

- Online regression oracle model [Foster et al. 2020]: for any adversarial sequence $\{a_t\}$, oracle outputs $\{\widehat{\theta}_t\}$ such that

$$\sum_{t=1}^{T}(f(\langle\theta^\star, a_t\rangle) - f(\langle\widehat{\theta}_t, a_t\rangle))^2 \leq \mathrm{Reg}_{\mathrm{Sq}}(T)$$

  $\rightarrow$ learner only observes $\widehat{\theta}_t$ in the oracle model, but not r_t;
  $\rightarrow$ a natural choice of $\mathrm{Reg}_{\mathrm{Sq}}(T)$ is $\widetilde{O}(d)$.

Suboptimality of online regression oracle

- Online regression oracle model [Foster et al. 2020]: for any adversarial sequence $\{a_t\}$, oracle outputs $\{\widehat{\theta}_t\}$ such that

$$\sum_{t=1}^{T}(f(\langle\theta^\star, a_t\rangle) - f(\langle\widehat{\theta}_t, a_t\rangle))^2 \le \mathrm{Reg}_{\mathrm{Sq}}(T)$$

  $\rightarrow$ learner only observes $\widehat{\theta}_t$ in the oracle model, but not r_t;
  $\rightarrow$ a natural choice of $\mathrm{Reg}_{\mathrm{Sq}}(T)$ is $\widetilde{O}(d)$.

Theorem (lower bound for the oracle model)

For every f, there exist a bandit instance under the oracle model such that for every algorithm, achieving a sublinear regret requires

$$T \gtrsim \max_K \min\left\{K, \frac{d}{f(\sqrt{(\log K)/d})^2}\right\}.$$

Suboptimality of online regression oracle

- Online regression oracle model [Foster et al. 2020]: for any adversarial sequence $\{a_t\}$, oracle outputs $\{\widehat{\theta}_t\}$ such that

$$\sum_{t=1}^{T}(f(\langle\theta^\star, a_t\rangle) - f(\langle\widehat{\theta}_t, a_t\rangle))^2 \le \mathrm{Reg}_{\mathrm{Sq}}(T)$$

  $\rightarrow$ learner only observes $\widehat{\theta}_t$ in the oracle model, but not r_t;
  $\rightarrow$ a natural choice of $\mathrm{Reg}_{\mathrm{Sq}}(T)$ is $\widetilde{O}(d)$.

Theorem (lower bound for the oracle model)

For every f, there exist a bandit instance under the oracle model such that for every algorithm, achieving a sublinear regret requires

$$T \gtrsim \max_K \min\left\{K, \frac{d}{f(\sqrt{(\log K)/d})^2}\right\}.$$

- Key modeling difference: in oracle model, choosing repeated action may not reduce the estimation error

Infinite vs finite actions

- in linear bandit, fewer actions lead to smaller regret

Infinite vs finite actions

- in linear bandit, fewer actions lead to smaller regret
  $\rightarrow$ minimax regret decreases from $\Theta(d\sqrt{T})$ to $\Theta(\sqrt{dT \log K})$ with K actions

Infinite vs finite actions

- in linear bandit, fewer actions lead to smaller regret
  - $\rightarrow$ minimax regret decreases from $\Theta(d\sqrt{T})$ to $\Theta(\sqrt{dT \log K})$ with K actions
  - $\rightarrow$ intuition: UCB needs to construct fewer confidence intervals

Infinite vs finite actions

- in linear bandit, fewer actions lead to smaller regret
  - $\rightarrow$ minimax regret decreases from $\Theta(d\sqrt{T})$ to $\Theta(\sqrt{dT \log K})$ with K actions
  - $\rightarrow$ intuition: UCB needs to construct fewer confidence intervals
- does similar phenomenon hold for non-linear bandits?

Infinite vs finite actions

- in linear bandit, fewer actions lead to smaller regret
  - $\rightarrow$ minimax regret decreases from $\Theta(d\sqrt{T})$ to $\Theta(\sqrt{dT \log K})$ with K actions
  - $\rightarrow$ intuition: UCB needs to construct fewer confidence intervals
- does similar phenomenon hold for non-linear bandits?

Theorem (lower bound for finite actions)

For every link function f and $K = \text{poly}(d)$, there exist an K-armed ridge bandit instance such that achieving a sublinear regret requires

$$T \gtrsim \min\left\{K, \frac{1}{f(1/\sqrt{d})^2}\right\}.$$

Infinite vs finite actions

- in linear bandit, fewer actions lead to smaller regret
  - $\rightarrow$ minimax regret decreases from $\Theta(d\sqrt{T})$ to $\Theta(\sqrt{dT \log K})$ with K actions
  - $\rightarrow$ intuition: UCB needs to construct fewer confidence intervals
- does similar phenomenon hold for non-linear bandits?

Theorem (lower bound for finite actions)

For every link function f and $K = \text{poly}(d)$, there exist an K-armed ridge bandit instance such that achieving a sublinear regret requires

$$T \gtrsim \min\left\{K, \frac{1}{f(1/\sqrt{d})^2}\right\}.$$

- implication: for $f(x) = x^3$, the fixed cost for the finite-action problem is already d^3, same as the infinite-action problem

Infinite vs finite actions

- in linear bandit, fewer actions lead to smaller regret
  - $\rightarrow$ minimax regret decreases from $\Theta(d\sqrt{T})$ to $\Theta(\sqrt{dT \log K})$ with K actions
  - $\rightarrow$ intuition: UCB needs to construct fewer confidence intervals
- does similar phenomenon hold for non-linear bandits?

Theorem (lower bound for finite actions)

For every link function f and $K = \text{poly}(d)$, there exist an K-armed ridge bandit instance such that achieving a sublinear regret requires

$$T \gtrsim \min\left\{K, \frac{1}{f(1/\sqrt{d})^2}\right\}.$$

- implication: for $f(x) = x^3$, the fixed cost for the finite-action problem is already d^3, same as the infinite-action problem
- reason: the learner cannot explore every direction in the initialization phase

Unit sphere vs unit ball

What happens if we assume that $\theta^\star \in \mathbb{B}^d$ instead of $\theta^\star \in \mathbb{S}^{d-1}$?

Unit sphere vs unit ball

What happens if we assume that $\theta^\star \in \mathbb{B}^d$ instead of $\theta^\star \in \mathbb{S}^{d-1}$?

Theorem (modified upper bound)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \max_{r \in [0,1]} \min \left\{ d^2 \frac{f(r)}{r^4} \int_{r/\sqrt{d}}^{r/2} \frac{\mathrm{d}(x^2)}{\max\limits_{r/\sqrt{d} \le y \le x} f'(y)^2} + d\sqrt{T}, Tf(r) \right\}.$$

Theorem (modified lower bound)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \gtrsim \max_{r \in [0,1]} \min \left\{ d \frac{f(r)}{r^2} \int_{r/\sqrt{d}}^{r/2} \frac{\mathrm{d}(x^2)}{f(x)^2} + d\sqrt{T}, Tf(r) \right\}.$$

Unit sphere vs unit ball

What happens if we assume that $\theta^\star \in \mathbb{B}^d$ instead of $\theta^\star \in \mathbb{S}^{d-1}$?

Theorem (modified upper bound)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \lesssim \max_{r \in [0,1]} \min \left\{ d^2 \frac{f(r)}{r^4} \int_{r/\sqrt{d}}^{r/2} \frac{\mathrm{d}(x^2)}{\max\limits_{r/\sqrt{d} \le y \le x} f'(y)^2} + d\sqrt{T}, Tf(r) \right\}.$$

Theorem (modified lower bound)

Under monotonicity and local linearity of f:

$$\mathrm{MinmaxReg}(T, d, f) \gtrsim \max_{r \in [0,1]} \min \left\{ d \frac{f(r)}{r^2} \int_{r/\sqrt{d}}^{r/2} \frac{\mathrm{d}(x^2)}{f(x)^2} + d\sqrt{T}, Tf(r) \right\}.$$

- minimax regret often exhibits only one elbow instead of two

Further questions

Further questions

- fill in the gap between upper and lower bounds

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

Further questions

- fill in the gap between upper and lower bounds

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

- analyze the learning trajectory of information-directed sampling $a_t = \arg\max_a I(\theta^\star; r_t \mid \mathcal{H}_{t-1}, a_t = a)$

Further questions

- fill in the gap between upper and lower bounds

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

- analyze the learning trajectory of information-directed sampling $a_t = \arg\max_a I(\theta^\star; r_t \mid \mathcal{H}_{t-1}, a_t = a)$
- more general reward functions such as $f_\theta(a) = \sum_{i=1}^m f_i(\langle \theta_i, a \rangle)$

Further questions

- fill in the gap between upper and lower bounds

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

- analyze the learning trajectory of information-directed sampling $a_t = \arg\max_a I(\theta^\star; r_t \mid \mathcal{H}_{t-1}, a_t = a)$
- more general reward functions such as $f_\theta(a) = \sum_{i=1}^m f_i(\langle \theta_i, a \rangle)$
- more systematic methods for exploration in the initialization phase

Further questions

- fill in the gap between upper and lower bounds

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

- analyze the learning trajectory of information-directed sampling $a_t = \arg\max_a I(\theta^\star; r_t \mid \mathcal{H}_{t-1}, a_t = a)$
- more general reward functions such as $f_\theta(a) = \sum_{i=1}^m f_i(\langle \theta_i, a \rangle)$
- more systematic methods for exploration in the initialization phase
- more complicated settings such as contextual bandits and RL

Concluding remarks

Take-home message:

Concluding remarks

Take-home message:

- there could be a phase transition in the regret of non-linear bandits

Concluding remarks

Take-home message:

- there could be a phase transition in the regret of non-linear bandits
- in the initialization phase, the learner needs algorithms beyond UCB to explore a good initial action, which incurs a fixed cost

Concluding remarks

Take-home message:

- there could be a phase transition in the regret of non-linear bandits
- in the initialization phase, the learner needs algorithms beyond UCB to explore a good initial action, which incurs a fixed cost
- in the learning phase, the learner can employ UCB-type algorithm around the good initial action

Concluding remarks

Take-home message:

- there could be a phase transition in the regret of non-linear bandits
- in the initialization phase, the learner needs algorithms beyond UCB to explore a good initial action, which incurs a fixed cost
- in the learning phase, the learner can employ UCB-type algorithm around the good initial action
- learning trajectory of the initialization phase could be characterized by proper differential equations

Concluding remarks

Take-home message:

- there could be a phase transition in the regret of non-linear bandits
- in the initialization phase, the learner needs algorithms beyond UCB to explore a good initial action, which incurs a fixed cost
- in the learning phase, the learner can employ UCB-type algorithm around the good initial action
- learning trajectory of the initialization phase could be characterized by proper differential equations
- traditional learning algorithms may fail to obtain the optimal initialization cost

Concluding remarks

Take-home message:

- there could be a phase transition in the regret of non-linear bandits
- in the initialization phase, the learner needs algorithms beyond UCB to explore a good initial action, which incurs a fixed cost
- in the learning phase, the learner can employ UCB-type algorithm around the good initial action
- learning trajectory of the initialization phase could be characterized by proper differential equations
- traditional learning algorithms may fail to obtain the optimal initialization cost

Thank You!