Toward Better Use of Data in Contextual and Linear Bandits

Nima Hamidi and Mohsen Bayati

Stanford University

October 2, 2020

References: arXiv 2002.05152 & arXiv 2006.06790

Overview

How to Test New Medical Interventions?

- A hospital wants to reduce post-discharge complications:
  - Use one of two newly designed telehealth (A or B)
- Should select one of A or B per patient
- A/B test or Randomized Control Trial (RCT) have high opportunity cost
  - In healthcare,experimentation is costly or unethical[1]

A

B

[1] Sibbald, Bonnie. 1998. Understanding controlled trials: Why are randomized controlled trials important?, British Medical Journal (Clinical Research Ed.) 316(201).

Beyond Healthcare

"Today, Microsoft and several other leading companies, including Amazon, Booking.com, Facebook, and Google, each conduct more than 10,000 online controlled experiments annually, with many tests engaging millions of users."

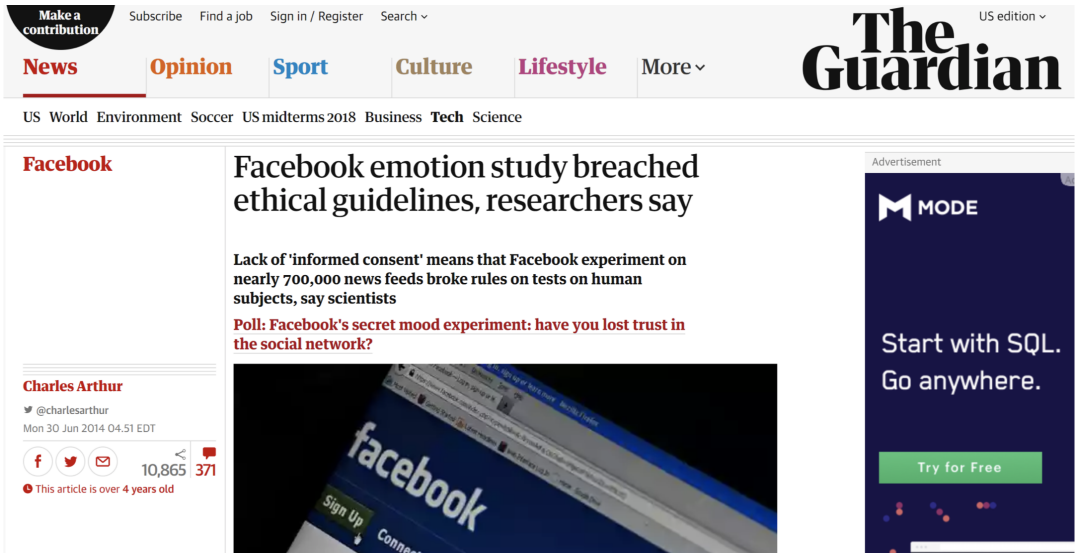
Kohavi and Thompke, Harvard Business Review, 2017

Beyond Healthcare

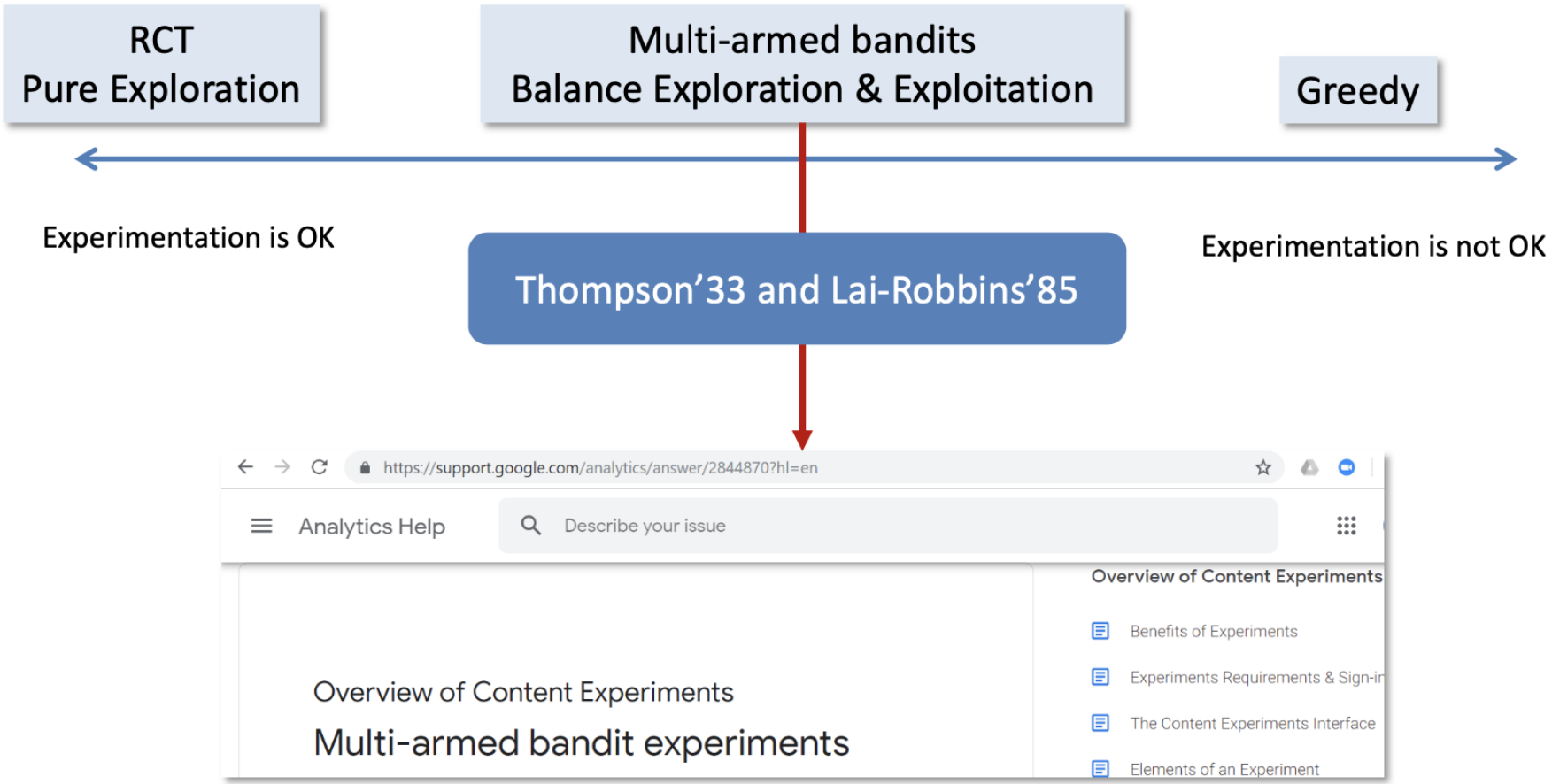
"Today, Microsoft and several other leading companies, including Amazon, Booking.com, Facebook, and Google, each conduct more than 10,000 online controlled experiments annually, with many tests engaging millions of users."

Kohavi and Thompke, Harvard Business Review, 2017

Also,



Multi-armed Bandit Experiments



Example (Google Analytics)[2]

- A/B testing
  - Website configurations A and B with conversion rates 4% and 5% respectively

- Using **Thompson Sampling**, instead of A/B testing, can run experiment with 78.5% less data $\rightarrow$ 97.5 conversions saved (on avg.)

[2] **Source:** Google Analytics Support Page

Stochastic Linear Bandit Problem

- Let $\Theta^\star \in \mathbb{R}^d$ be fixed (and unknown).
- At time t, the **action set** $\mathcal{A}_t \subseteq \mathbb{R}^d$ is revealed to a **policy** π.
- The policy chooses $\widetilde{A}_t \in \mathcal{A}_t$.
- It observes a **reward** $r_t = \langle \Theta^\star, \widetilde{A}_t \rangle + \varepsilon_t$.
- Conditional on the history, ε_t has zero mean.

Evaluation Metric

- The objective is to **improve using past experiences**.

- The **cumulative regret** is defined as

$$\text{Regret}(T, \Theta^\star, \pi) := \mathbb{E}\left[\sum_{i=1}^{T} \sup_{A \in \mathcal{A}_t} \langle \Theta^\star, A \rangle - \langle \Theta^\star, \widetilde{A}_t \rangle \,\middle|\, \Theta^\star\right].$$

Evaluation Metric

- The objective is to **improve using past experiences**.
- The **cumulative regret** is defined as

$$\text{Regret}(T, \Theta^\star, \pi) := \mathbb{E}\left[\sum_{i=1}^{T} \sup_{A \in \mathcal{A}_t} \langle \Theta^\star, A \rangle - \langle \Theta^\star, \widetilde{A}_t \rangle \,\middle|\, \Theta^\star\right].$$

- In the Bayesian setting, the **Bayesian regret** is given by

$$\text{BayesRegret}(T, \pi) := \mathbb{E}_{\Theta^\star \sim \mathcal{P}}[\text{Regret}(T, \Theta^\star, \pi)].$$

Special Cases

- Standard multi-armed bandit problem

Special Cases

- Standard multi-armed bandit problem
- k-armed contextual bandit problem

Special Cases

- Standard multi-armed bandit problem
- k-armed contextual bandit problem
- Dynamic-pricing with demand covariates

$$\text{Expected Demand} = \alpha + \beta\, p + \langle \Gamma, X \rangle$$
$$\text{Expected Revenue} = \alpha p + \beta\, p^2 + \langle \Gamma, X \rangle p$$

Special Cases

- Standard multi-armed bandit problem
- k-armed contextual bandit problem
- Dynamic-pricing with demand covariates

$$\text{Expected Demand} = \alpha + \beta\, p + \langle \Gamma, X \rangle$$
$$\text{Expected Revenue} = \alpha p + \beta\, p^2 + \langle \Gamma, X \rangle p$$

can be mapped to a linear bandit by setting

$$\mathcal{A} = \{(p, p^2, pX) | p \in [p_{\min}, p_{\max}]\} \text{ and } \Theta^\star = \begin{bmatrix} \alpha \\ \beta \\ \Gamma \end{bmatrix}$$

Related Literature

- **UCB/OFUL**: Auer, Cesa-Bianchi, and Fischer 2002; Dani, Hayes, and Kakade 2008; Rusmevichientong and Tsitsiklis 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011
- **Thompson sampling:** Agrawal and Goyal 2013; Russo and Van Roy 2014, 2016; Abeille and Lazaric 2017
- ϵ-**Greedy and variants:** Langford and Zhang 2008; Goldenshluger and Zeevi 2013

Related Literature

- **UCB/OFUL**: Auer, Cesa-Bianchi, and Fischer 2002; Dani, Hayes, and Kakade 2008; Rusmevichientong and Tsitsiklis 2010; Abbasi-Yadkori, Pál, and Szepesvári 2011
- **Thompson sampling:** Agrawal and Goyal 2013; Russo and Van Roy 2014, 2016; Abeille and Lazaric 2017
- ϵ-**Greedy and variants:** Langford and Zhang 2008; Goldenshluger and Zeevi 2013

Learning and earning in operations: Carvalho and Puterman05, Araman and Caldentey09, Besbes and Zeevi0911, Harrison et al.12, den Boer and Zwart14-16, Keskin and Zeevi14-16, Gur et al.'14, Johnson et al.15, Chen et al.15, Cohen et al 16, Bayati and Bastani'15, Kallus and Udell16, Javanmard and Nazerzadeh16, Javanmard17, Elmachtoub et. al.'17, Ban and Keskin17, Cheung et al '18, Bastani et al.'19, and many more!

Algorithms

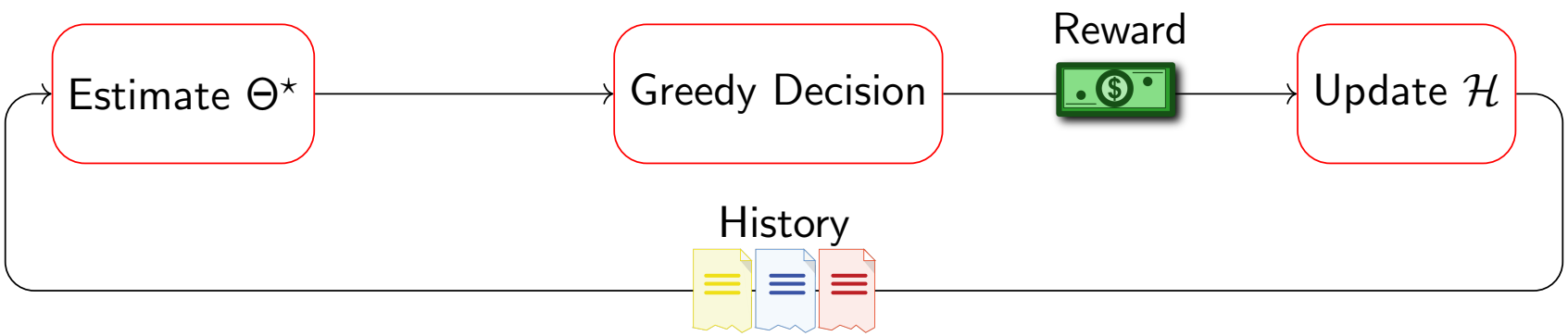
Greedy

At time $t = 1, 2, \cdots, T$:

- Using the set of observations

$$\mathcal{H}_{t-1} := \{(\widetilde{A}_1, r_1), \cdots, (\widetilde{A}_{t-1}, r_{t-1})\},$$

- Construct an **estimate** $\widehat{\Theta}_{t-1}$ for $\Theta^\star$,
- Choose the action $A \in \mathcal{A}_t$ with **largest** $\langle A, \widehat{\Theta}_{t-1} \rangle$.



Greedy

The **ridge estimator** is used to obtain $\widehat{\Theta}_t$ (for a fixed λ):

$$\mathbf{V}_t := \lambda \mathbf{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top \in \mathbb{R}^{d \times d}, \tag{1}$$

and

$$\widehat{\Theta}_t := \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right) \in \mathbb{R}^d. \tag{2}$$

Greedy

Algorithm 1 Greedy algorithm

1: **for** $t = 1$ to T **do**
2: Pull $\widetilde{A}_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widehat{\Theta}_{t-1} \rangle$
3: Observe the reward r_t
4: Compute $\mathbf{V}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
5: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$
6: **end for**

Greedy

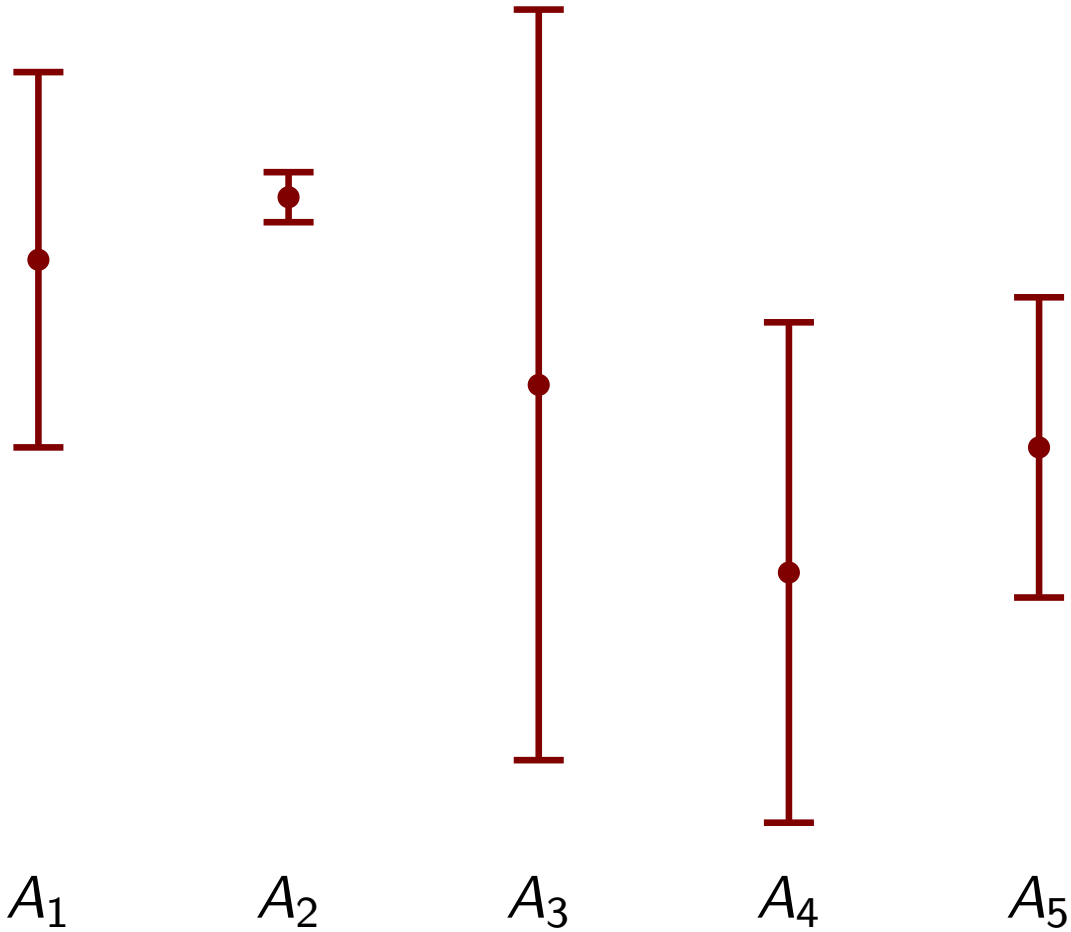
Algorithm 2 Greedy algorithm

1: **for** $t = 1$ to T **do**
2: Pull $\widetilde{A}_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widehat{\Theta}_{t-1} \rangle$
3: Observe the reward r_t
4: Compute $\mathbf{V}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
5: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$
6: **end for**

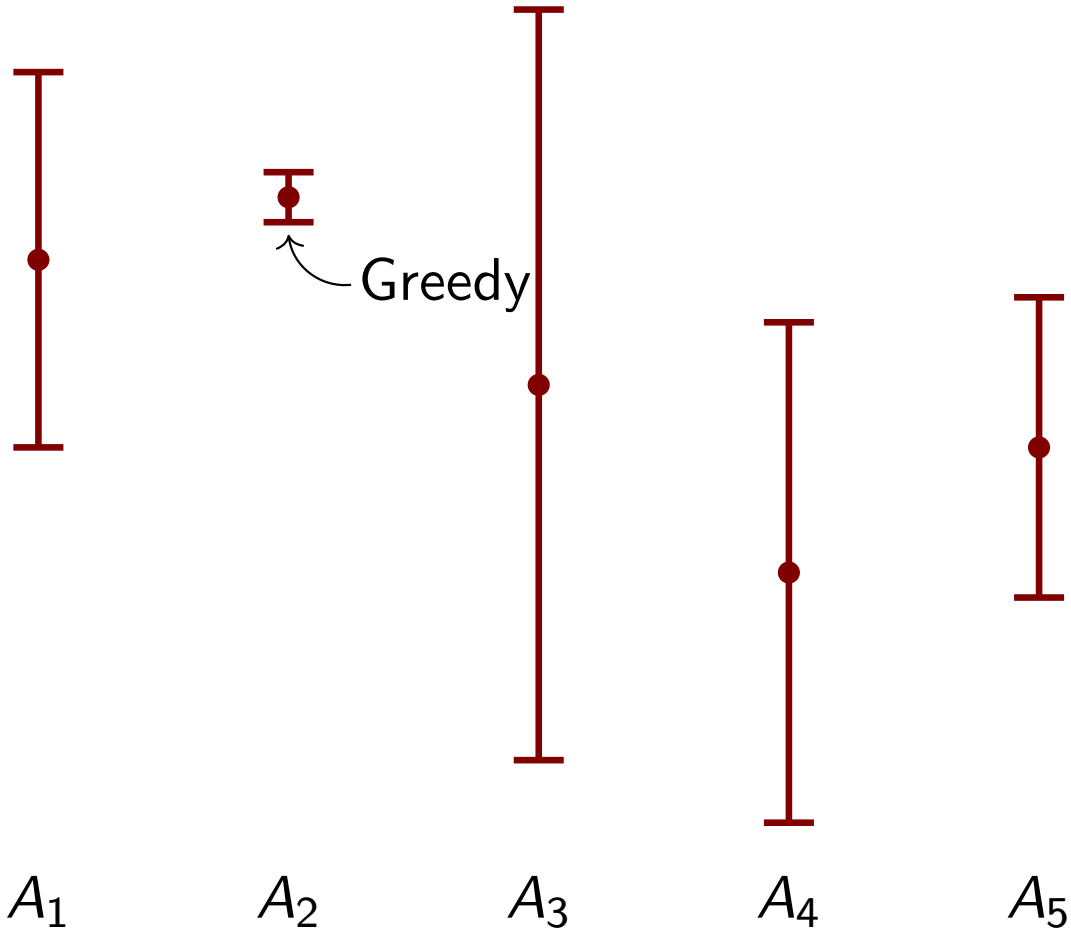
Greedy makes wrong decisions due to **over-** or **under-estimating** the true rewards.

- The over-estimation is **automatically** corrected.
- The under-estimation can cause **linear regret**.

Greedy



Greedy



Optimism in Face of Uncertainty (OFU) Algorithm

- Key idea: **be optimistic** when estimating the reward of actions.

Optimism in Face of Uncertainty (OFU) Algorithm

- Key idea: **be optimistic** when estimating the reward of actions.
- For $\rho > 0$, define the **confidence set** $\mathcal{C}_{t-1}(\rho)$ to be

$$\mathcal{C}_{t-1}(\rho) := \{\Theta \mid \|\Theta - \widehat{\Theta}_{t-1}\|_{\mathbf{V}_{t-1}} \leq \rho\},$$

where

$$\|X\|^2_{\mathbf{V}_{t-1}} = X^\top \mathbf{V}_{t-1} X \in \mathbb{R}^+.$$

Optimism in Face of Uncertainty (OFU) Algorithm

- Key idea: **be optimistic** when estimating the reward of actions.
- For $\rho > 0$, define the **confidence set** $\mathcal{C}_{t-1}(\rho)$ to be

$$\mathcal{C}_{t-1}(\rho) := \{\Theta \mid \|\Theta - \widehat{\Theta}_{t-1}\|_{\mathbf{V}_{t-1}} \leq \rho\},$$

where

$$\|X\|^2_{\mathbf{V}_{t-1}} = X^\top \mathbf{V}_{t-1} X \in \mathbb{R}^+.$$

Theorem (Informal, Abbasi-Yadkori, Pál, and Szepesvári 2011)

Letting $\rho := \widetilde{\mathcal{O}}(\sqrt{d})$, we have $\Theta^\star \in \mathcal{C}_{t-1}(\rho)$ with high probability.

Optimism in Face of Uncertainty (OFU) Algorithm

Algorithm 3 OFUL algorithm

1: **for** $t = 1$ to T **do**
2: Pull $\widetilde{A}_t := \arg\max_{A \in \mathcal{A}_t} \sup_{\Theta \in \mathcal{C}_{t-1}(\rho)} \langle A, \Theta \rangle$
3: Observe the reward r_t
4: Compute $\mathbf{V}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
5: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$
6: **end for**

Optimism in Face of Uncertainty (OFU) Algorithm

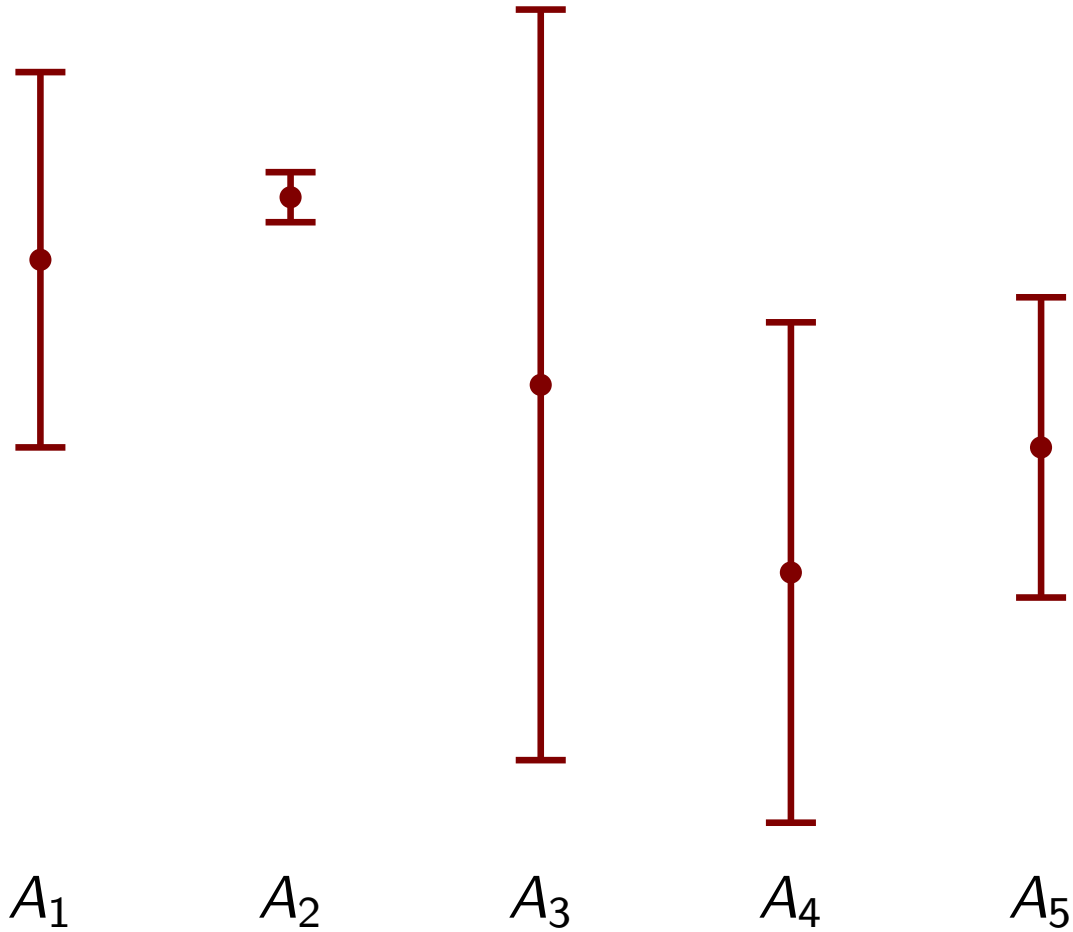
Algorithm 3 OFUL algorithm

1: **for** $t = 1$ to T **do**
2: Pull $\widetilde{A}_t := \arg\max_{A \in \mathcal{A}_t} \sup_{\Theta \in \mathcal{C}_{t-1}(\rho)} \langle A, \Theta \rangle$
3: Observe the reward r_t
4: Compute $\mathbf{V}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
5: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$
6: **end for**

It can be shown that

$$\sup_{\Theta \in \mathcal{C}_{t-1}(\rho)} \langle A, \Theta \rangle = \langle A, \widehat{\Theta}_{t-1} \rangle + \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}.$$

Optimism in Face of Uncertainty (OFU) Algorithm



Linear Thompson Sampling (LinTS) Algorithm

- Key idea: use **randomization** to address under-estimation.

Linear Thompson Sampling (LinTS) Algorithm

- Key idea: use **randomization** to address under-estimation.
- LinTS samples from the **posterior** distribution of $\Theta^\star$.

Algorithm 4 LinTS algorithm

1: **for** $t = 1$ to T **do**
2: Sample $\widetilde{\Theta}_{t-1} \sim \mathbb{P}(\Theta^\star \mid \mathcal{H}_{t-1})$
3: Pull $A_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widetilde{\Theta}_{t-1} \rangle$
4: Observe the reward r_t
5: Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \bigcup \{(A_t, r_t)\}$
6: **end for**

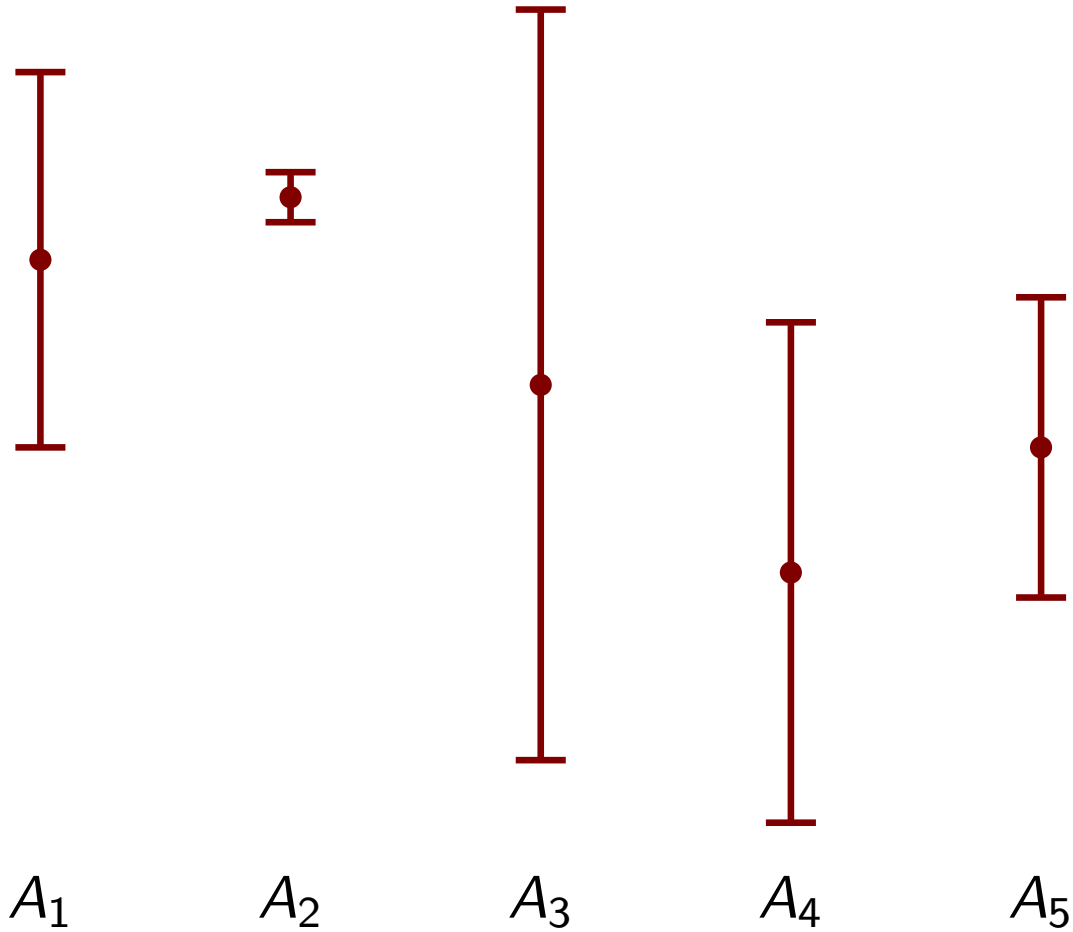
Linear Thompson Sampling (LinTS) Algorithm

- Under **normality**, LinTS becomes:

Algorithm 5 LinTS algorithm under normality

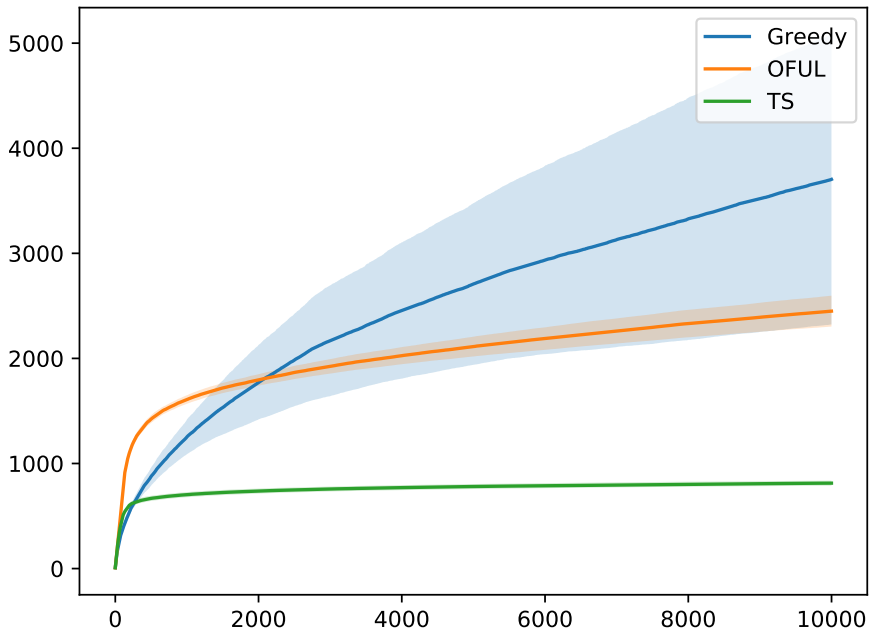
1: **for** $t = 1$ to T **do**
2: Sample $\widetilde{\Theta}_{t-1} \sim \mathcal{N}(\widehat{\Theta}_{t-1}, \mathbf{V}_{t-1}^{-1})$
3: Pull $A_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widetilde{\Theta}_{t-1} \rangle$
4: Observe the reward r_t
5: Compute $\mathbf{V}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
6: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$
7: **end for**

Linear Thompson Sampling (LinTS) Algorithm



Why Is LinTS Popular?

- **Empirical superiority**:
  - $d = 120$, $\Theta^\star \sim \mathcal{N}(0, \mathbf{I}_d)$,
  - $k = 10$, $X \sim \mathcal{N}(0, \mathbf{I}_{12})$,
  - Each A_t contains X as a block[3].



[3] This is the 10-armed contextual bandit with 12 dimensional covariates.

Comparison of Regret Bounds

Theorem (Abbasi-Yadkori, Pál, and Szepesvári 2011)

Under some conditions, the regret of OFUL is bounded by

$$\text{Regret}(T, \Theta^\star, \pi^{OFUL}) \leq \widetilde{\mathcal{O}}(d\sqrt{T}).$$

Comparison of Regret Bounds

Theorem (Abbasi-Yadkori, Pál, and Szepesvári 2011)

Under some conditions, the regret of OFUL is bounded by

$$\text{Regret}(T, \Theta^\star, \pi^{OFUL}) \leq \widetilde{\mathcal{O}}(d\sqrt{T}).$$

Theorem (Russo and Van Roy 2014)

Under minor assumptions, the Bayesian regret of LinTS is bounded by

$$\text{BayesRegret}(T, \pi^{LinTS}) \leq \widetilde{\mathcal{O}}(d\sqrt{T}).$$

Comparison of Regret Bounds

Theorem (Abbasi-Yadkori, Pál, and Szepesvári 2011)

Under some conditions, the regret of OFUL is bounded by

$$\text{Regret}(T, \Theta^\star, \pi^{OFUL}) \leq \widetilde{\mathcal{O}}(d\sqrt{T}).$$

Theorem (Russo and Van Roy 2014)

Under minor assumptions, the Bayesian regret of LinTS is bounded by

$$\text{BayesRegret}(T, \pi^{LinTS}) \leq \widetilde{\mathcal{O}}(d\sqrt{T}).$$

Theorem (Dani, Hayes, and Kakade 2008)

There is a Bayesian linear bandit problem that satisfies

$$\inf_{\pi} \text{BayesRegret}(T, \pi) \geq \Omega(d\sqrt{T}).$$

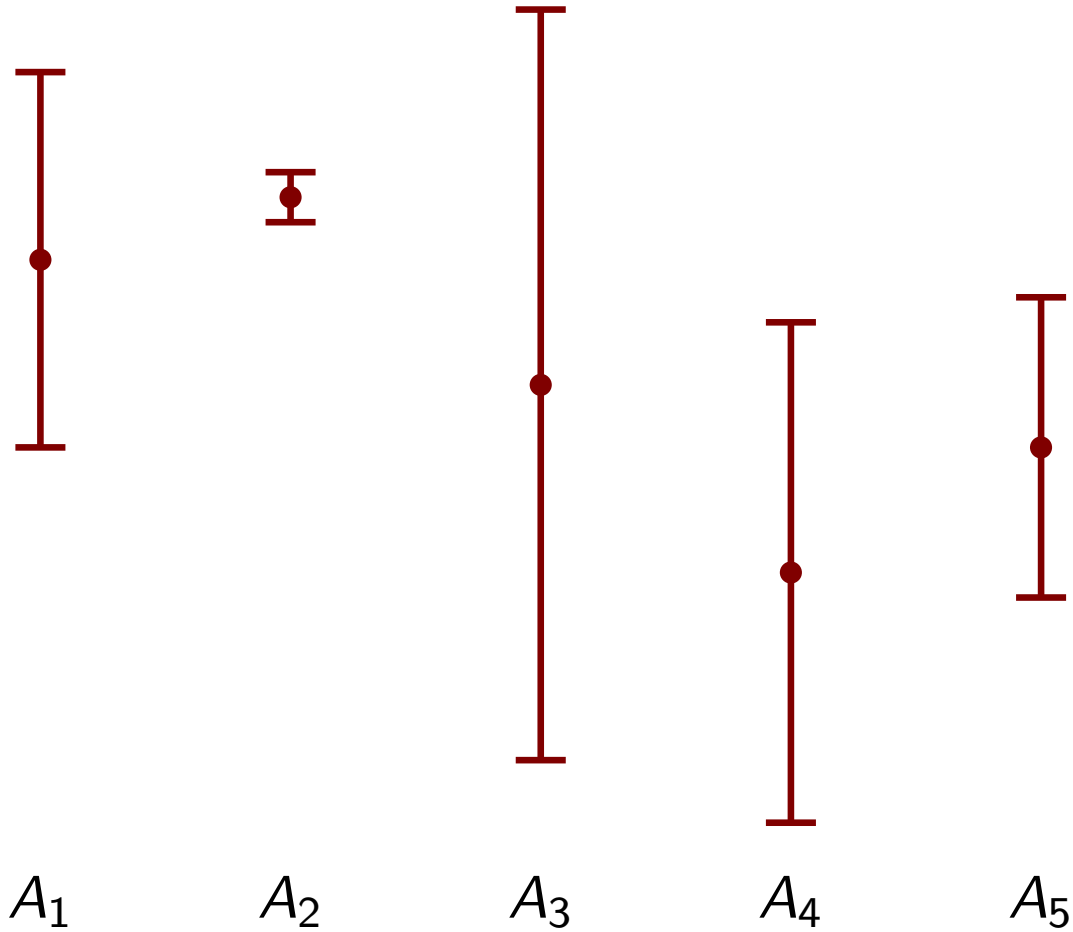
A Worst-Case Regret Bound for LinTS

- Question: can one prove a similar worst-case regret bound for LinTS?
- The only known results require **inflating** the posterior variance.

Algorithm 6 LinTS algorithm under normality

1: **for** $t = 1$ to T **do**
2: Sample $\widetilde{\Theta}_{t-1} \sim \mathcal{N}(\widehat{\Theta}_{t-1}, \beta^2 \mathbf{V}_{t-1}^{-1})$
3: Pull $A_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widetilde{\Theta}_{t-1} \rangle$
4: Update $\mathbf{V}_t$ and $\widehat{\Theta}_t$
5: **end for**

Inflated Linear Thompson Sampling (LinTS) Algorithm



A Worst-Case Regret Bound for LinTS

Theorem (Agrawal and Goyal 2013; Abeille and Lazaric 2017)

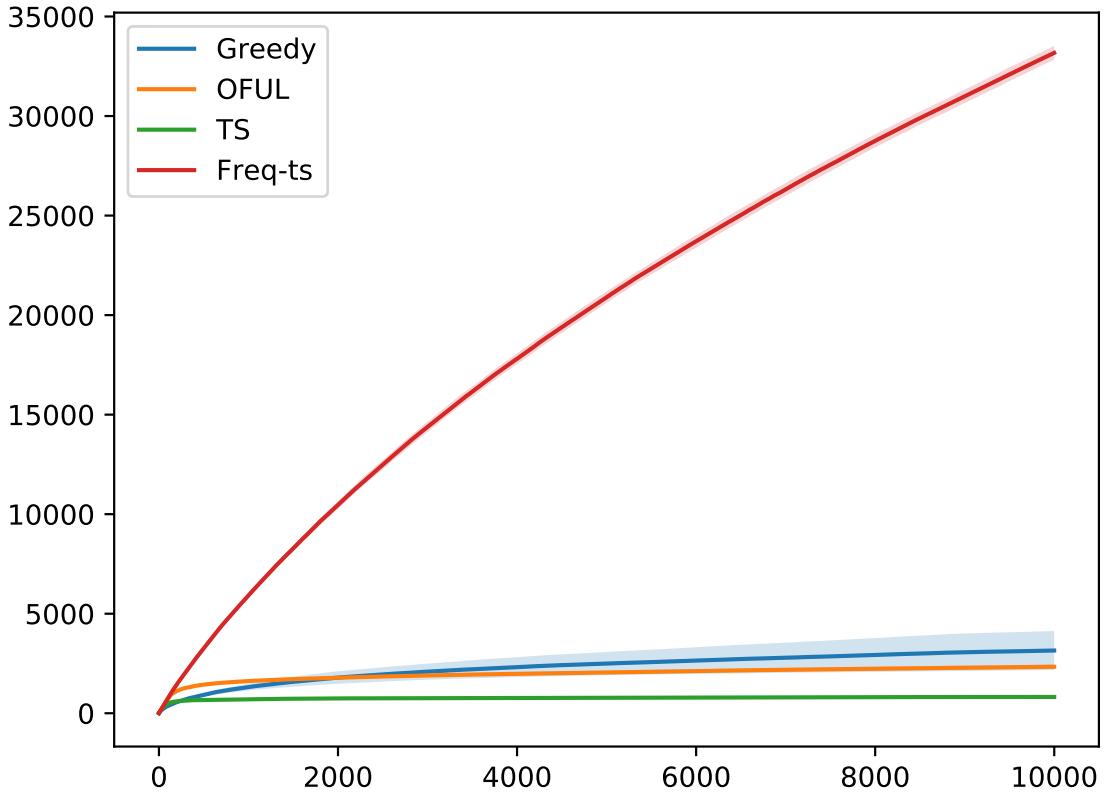
If $\beta \propto \sqrt{d}$, *then*

$$\text{Regret}(T, \Theta^\star, \pi^{LinTS}) \leq \widetilde{\mathcal{O}}(d\sqrt{dT}).$$

This result is far from optimal by a $\sqrt{d}$ factor.

Empirical Performance of Inflated LinTS

- Unfortunately, the inflated variant of LinTS performs poorly...



Bayesian Analyses are Brittle

We prove that the inflation is **necessary** for LinTS to work.

Theorem

There exists a linear bandit problem such that for $T \leq \exp(\Omega(d))$, *we have*

$$\text{BayesRegret}(T, \pi^{LinTS}) = \Omega(T).$$

Bayesian Analyses are Brittle

We prove that the inflation is **necessary** for LinTS to work.

Theorem

There exists a linear bandit problem such that for $T \leq \exp(\Omega(d))$*, we have*

$$\text{BayesRegret}(T, \pi^{LinTS}) = \Omega(T).$$

The counter-example satisfies the following properties:

- $\Theta^\star \sim \mathcal{N}(0, \mathbf{I}_d)$,
- LinTS uses the right prior,
- LinTS assumes noises are standard normal,
- $r_t = \langle \Theta^\star, A_t \rangle$. (i.e., **noiseless** data!)

Some Remarks on LinTS

- Under some assumptions on the action set we can prove that LinTS with only a logarithmic inflation works.
- Bastani, Simchi-Levi, and Zhu (2019) showed, in the dynamic pricing case, when only the prior mean is unknown but prior variance is known, regret loss can be only up to a constant.
- Jin, Xu, Shi, Xiao, and Gu (2020) showed a variant of TS achieves optimal regret bound in the standard k-armed bandit setting.

Improving OFUL

Weaker Optimism and Improving OFUL

- Unlike its name, OFUL is very **pessimistic**.
- It assumes that $\langle A, \widehat{\Theta}_t \rangle$ is **as small as it can be** for all arms.
- In practice, this may not be the case.
- One may be able to **use data more efficiently**.
- This is challenging since even choosing **the second most optimistic action** can lead to **linear regret**.

Optimism Baseline

Let $\mathsf{L}_t(A)$ be the **lower confidence-bound** for action A defined as

$$\mathsf{L}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle - \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}.$$

Optimism Baseline

Let $\mathsf{L}_t(A)$ be the **lower confidence-bound** for action A defined as

$$\mathsf{L}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle - \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}.$$

A simple observation:

Lemma

The following inequality holds with high probability:

$$\sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \geq \sup_{A \in \mathcal{A}_t} \mathsf{L}_t(A).$$

Then, define the **baseline** as $\mathsf{B}_t := \sup_{A \in \mathcal{A}_t} \mathsf{L}_t(A)$.

Sieved-Greedy

Let $\mathsf{L}_t(A)$ and $\mathsf{U}_t(A)$ be the **confidence-bounds** for action A given by

$$\mathsf{U}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle + \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$$
$$\mathsf{L}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle - \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}.$$

Sieved-Greedy

Let $\mathsf{L}_t(A)$ and $\mathsf{U}_t(A)$ be the **confidence-bounds** for action A given by

$$\mathsf{U}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle + \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$$
$$\mathsf{L}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle - \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}.$$

For a fixed **sieving-rate** $\alpha \in (0, 1]$, define

$$\mathcal{A}'_t := \left\{ A \in \mathcal{A}_t : \mathsf{U}_t(A) - \mathsf{B}_t \geq \alpha \left(\sup_{A' \in \mathcal{A}_t} \mathsf{U}_t(A') - \mathsf{B}_t \right) \right\}.$$

Sieved-Greedy

Let $\mathsf{L}_t(A)$ and $\mathsf{U}_t(A)$ be the **confidence-bounds** for action A given by

$$\mathsf{U}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle + \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$$
$$\mathsf{L}_t(A) := \langle A, \widehat{\Theta}_{t-1} \rangle - \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}.$$

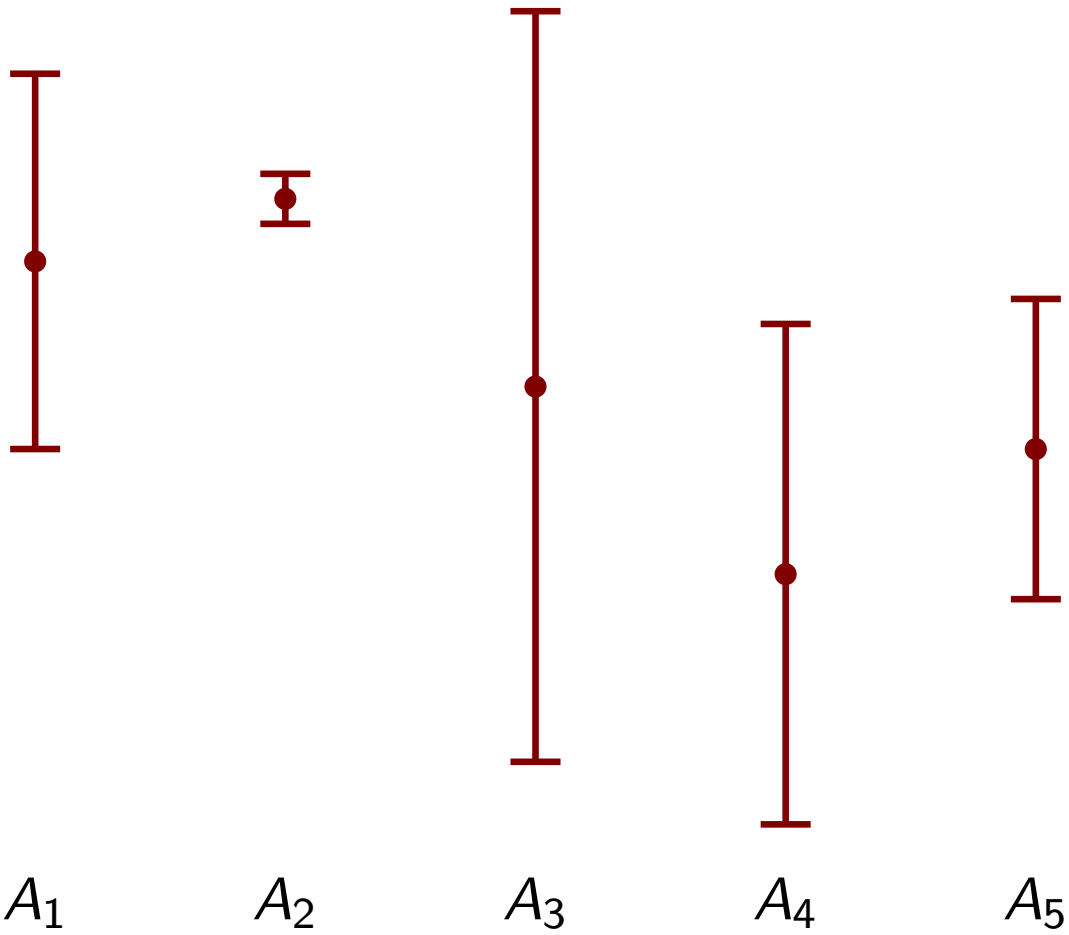
For a fixed **sieving-rate** $\alpha \in (0, 1]$, define

$$\mathcal{A}'_t := \left\{ A \in \mathcal{A}_t : \mathsf{U}_t(A) - \mathsf{B}_t \geq \alpha \left(\sup_{A' \in \mathcal{A}_t} \mathsf{U}_t(A') - \mathsf{B}_t \right) \right\}.$$

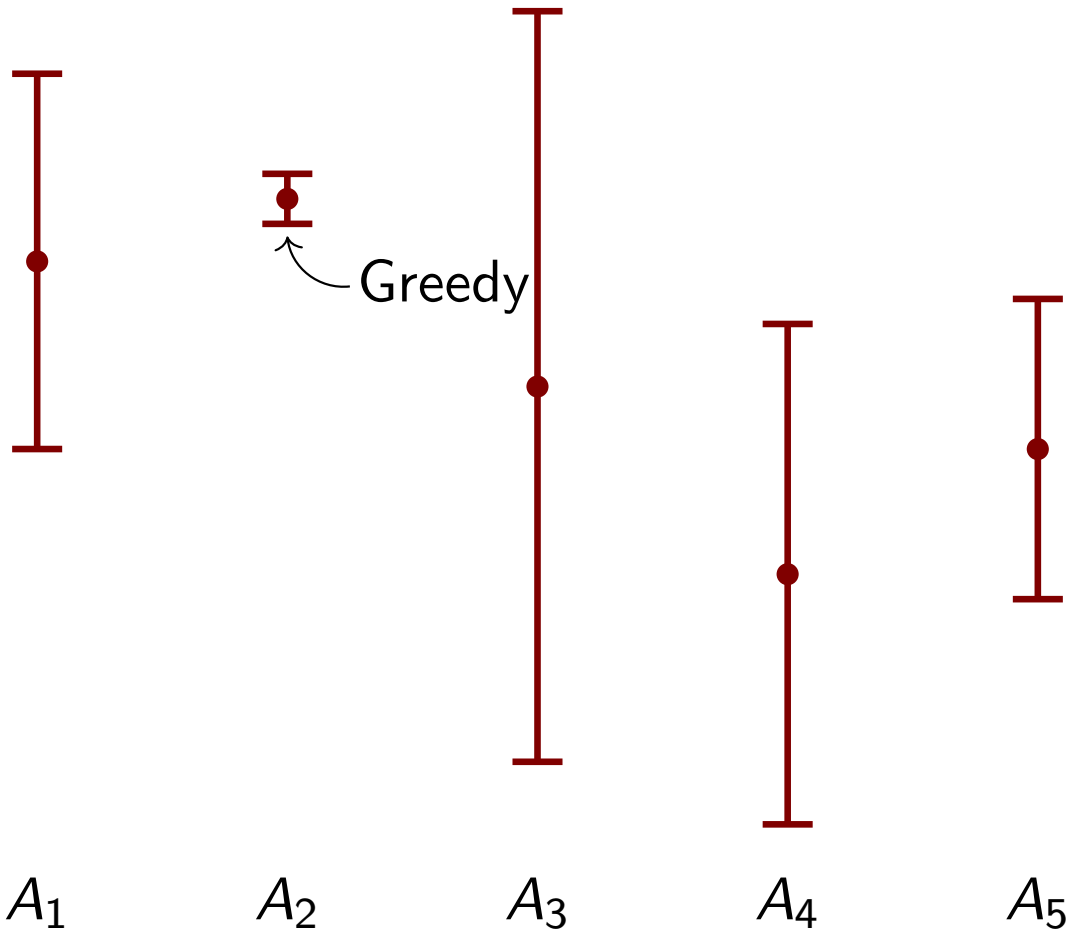
Sieved-greedy (SG) then chooses

$$\widetilde{A}_t \in \arg\max_{A \in \mathcal{A}'_t} \langle A, \widehat{\Theta}_{t-1} \rangle.$$

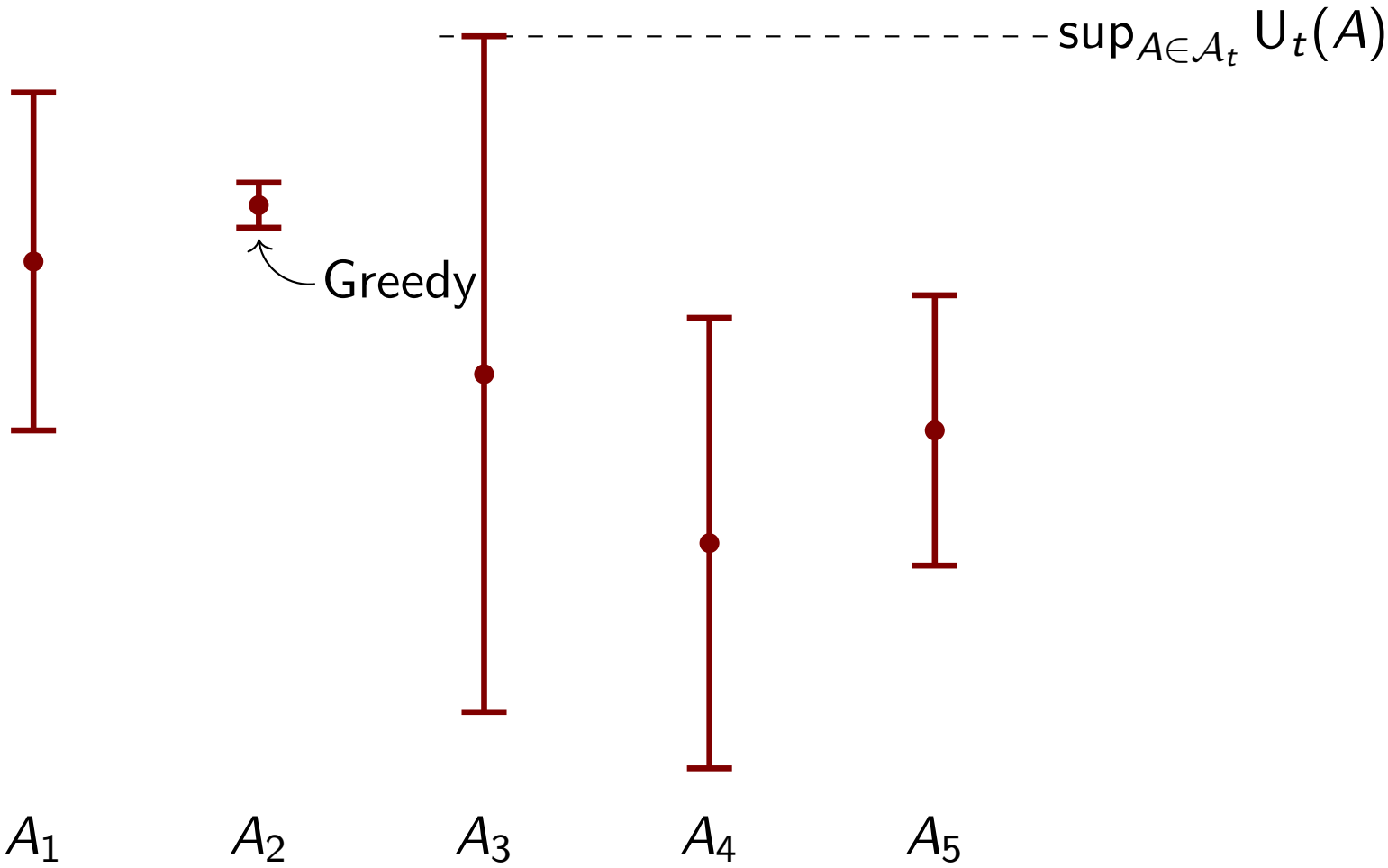
Sieved-Greedy



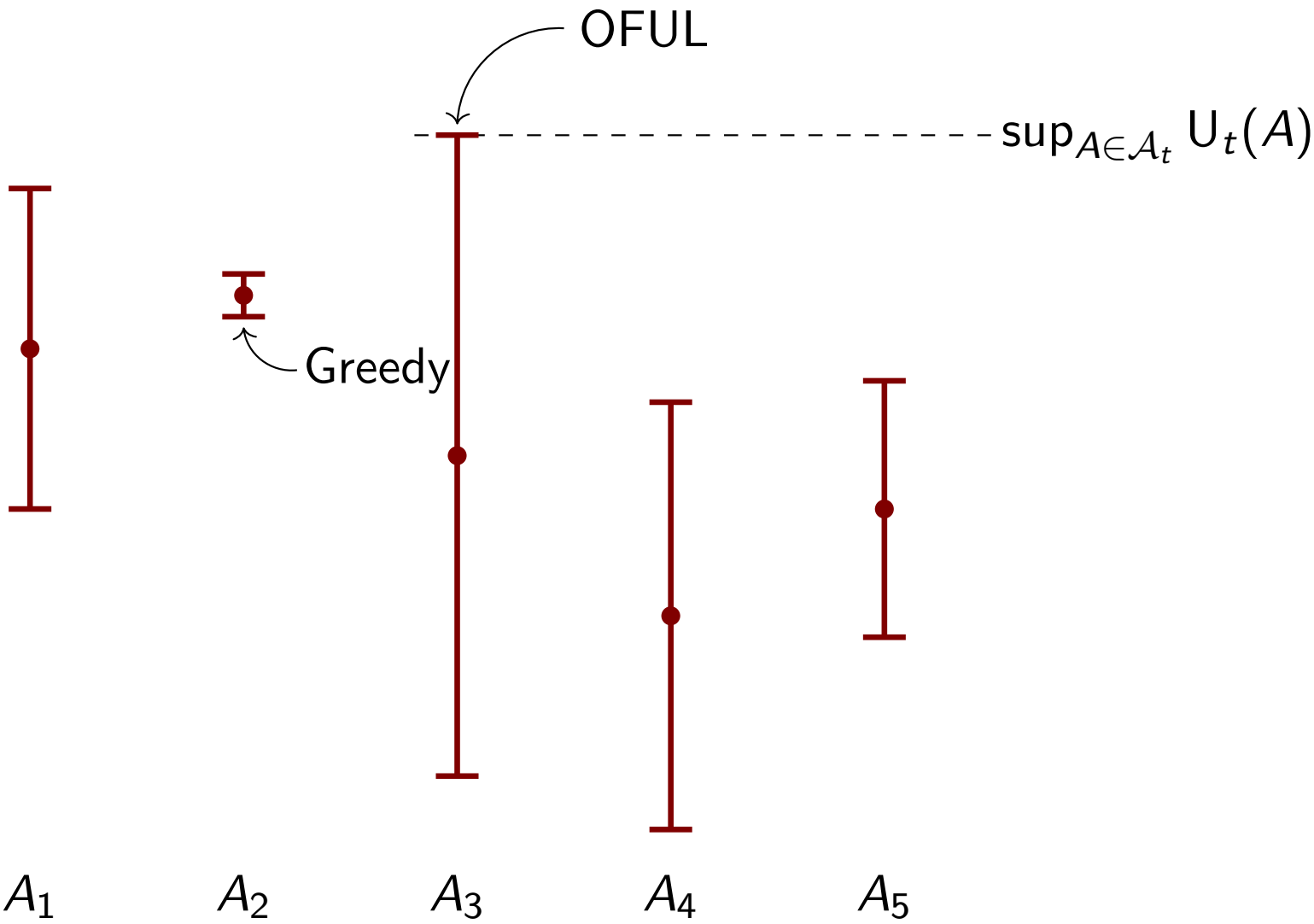
Sieved-Greedy



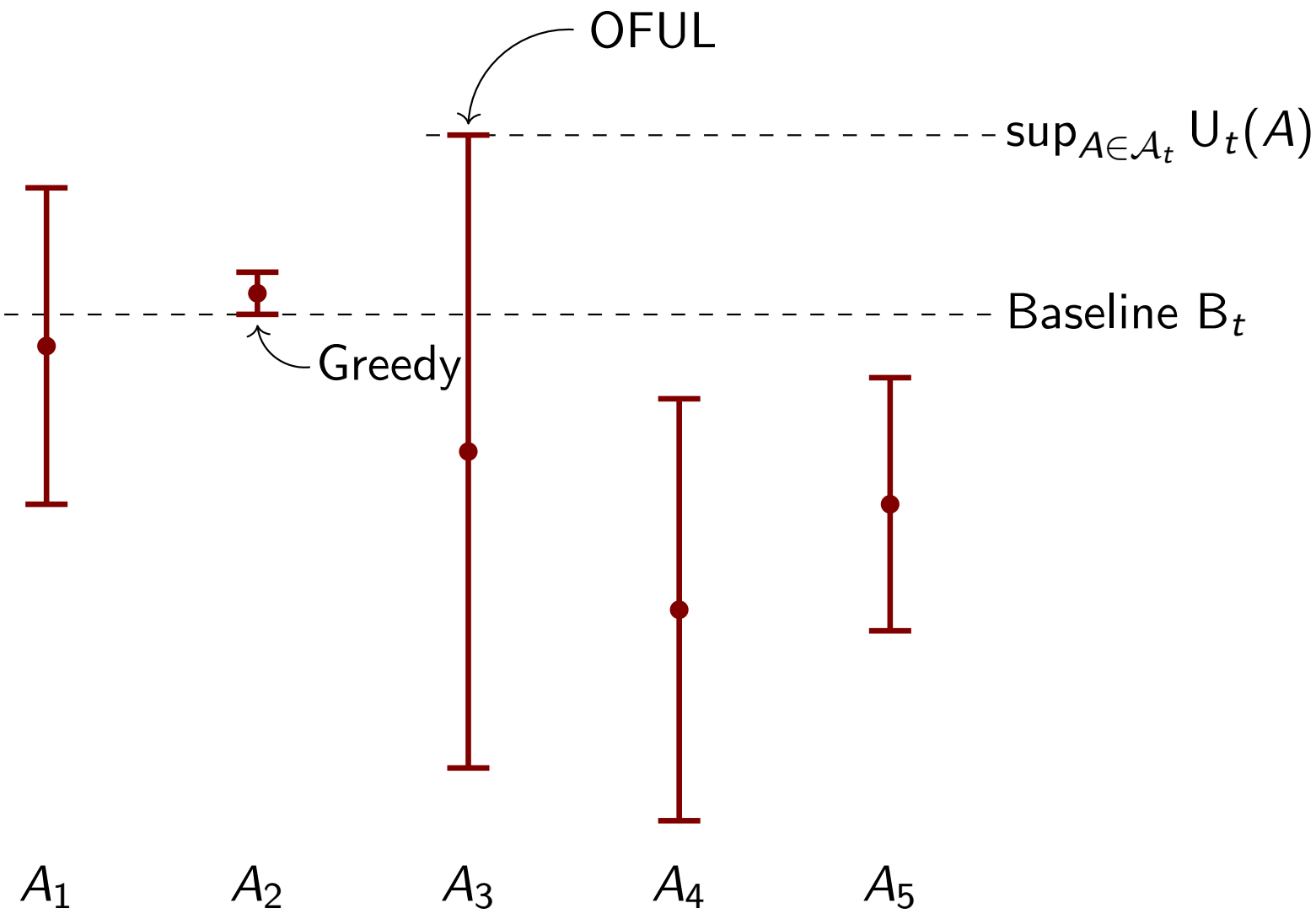
Sieved-Greedy



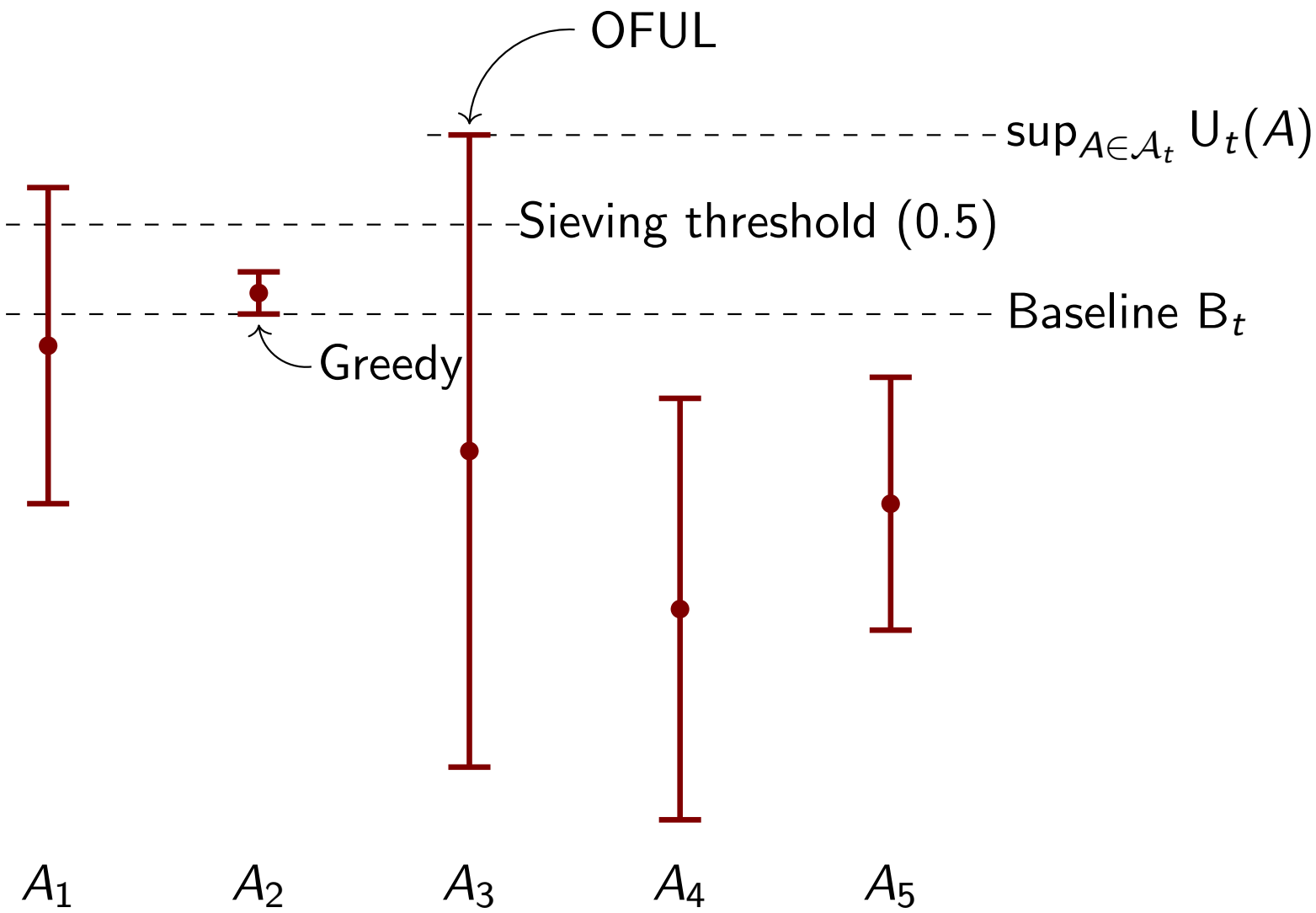
Sieved-Greedy



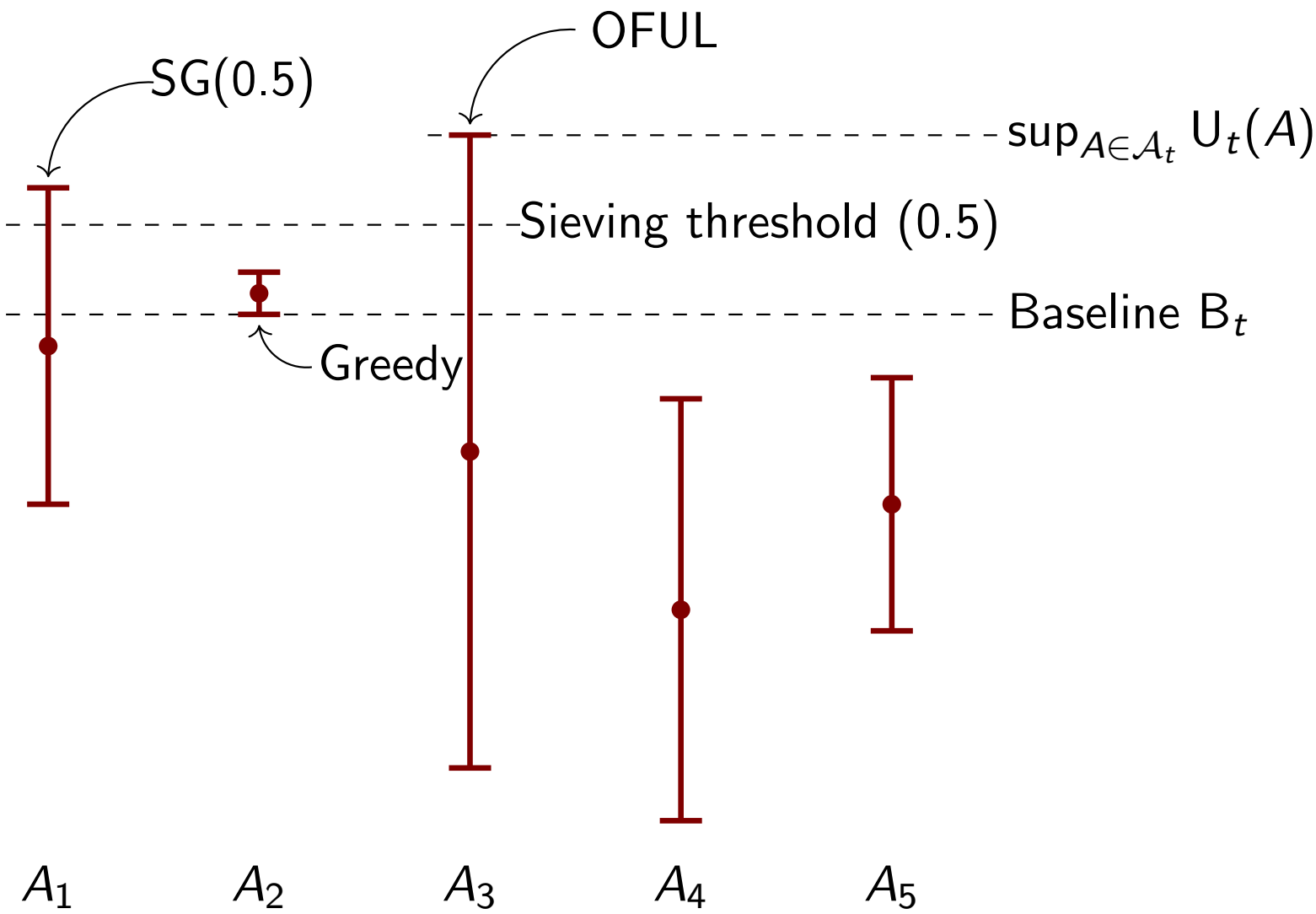
Sieved-Greedy



Sieved-Greedy



Sieved-Greedy



Sieved-Greedy

Two instances of Sieved-greedy:

- For $\alpha = 1$, Sieved-greedy is equivalent to OFUL.
- For $\alpha = 0$, Sieved-greedy is the same as Greedy.

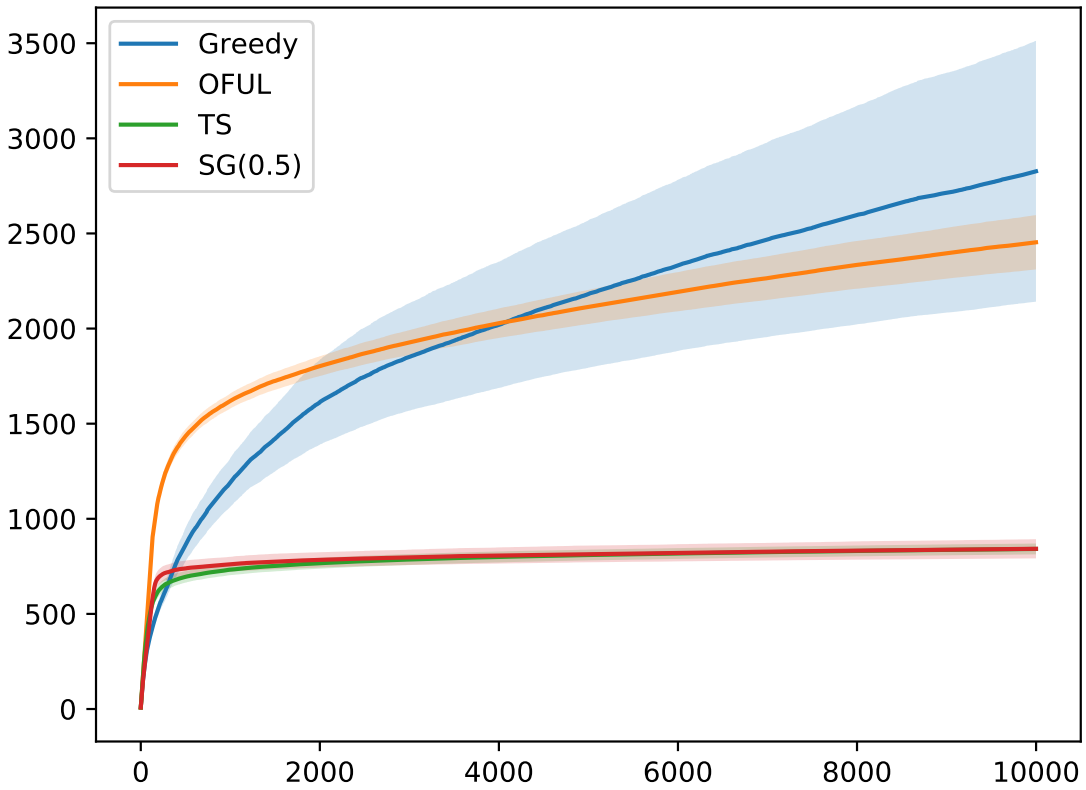
Sieved-Greedy

Theorem

For $\alpha > 0$, the regret of Sieved-greedy is bounded by

$$\text{Regret}(T, \Theta^\star, \pi^{\text{SG}(\alpha)}) \leq \widetilde{\mathcal{O}}\left(\frac{d\sqrt{T}}{\alpha}\right).$$

Simulations



A General Regret Bound

- By a **worth function**, we mean a function $\widetilde{\mathrm{M}}_t$ that maps each $A \in \mathcal{A}_t$ to $\mathbb{R}$ such that

$$|\widetilde{\mathrm{M}}_t(A) - \langle A, \widehat{\Theta}_{t-1} \rangle| \leq \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$$

  with probability at least $1 - \frac{1}{T^2}$.

A General Regret Bound

- By a **worth function**, we mean a function $\widetilde{\mathrm{M}}_t$ that maps each $A \in \mathcal{A}_t$ to $\mathbb{R}$ such that

$$|\widetilde{\mathrm{M}}_t(A) - \langle A, \widehat{\Theta}_{t-1} \rangle| \leq \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$$

  with probability at least $1 - \frac{1}{T^2}$.

- Next, define **Randomized OFUL (ROFUL)** to be:

Algorithm 7 ROFUL algorithm

1: **for** $t = 1$ to T **do**
2: Pull $\widetilde{A}_t := \arg\max_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A)$
3: Observe the reward r_t
4: Compute $\mathbf{V}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
5: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$
6: **end for**

A General Regret Bound

Examples of worth functions:

- Greedy: $\widetilde{\mathrm{M}}_t(A) = \langle A, \widehat{\Theta}_{t-1} \rangle$
- OFUL: $\widetilde{\mathrm{M}}_t(A) = \langle A, \widehat{\Theta}_{t-1} \rangle + \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$
- LinTS: $\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_{t-1} \rangle$

A General Regret Bound

Examples of worth functions:

- Greedy: $\widetilde{\mathsf{M}}_t(A) = \langle A, \widehat{\Theta}_{t-1} \rangle$
- OFUL: $\widetilde{\mathsf{M}}_t(A) = \langle A, \widehat{\Theta}_{t-1} \rangle + \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$
- LinTS: $\widetilde{\mathsf{M}}_t(A) = \langle A, \widetilde{\Theta}_{t-1} \rangle$
- Sieved-greedy:

$$\widetilde{\mathsf{M}}_t(A) = \begin{cases} \mathsf{U}_t(A) & \text{if } A = \widetilde{A}_t \\ \mathsf{L}_t(A) & \text{otherwise} \end{cases}$$

Weaker Optimism and Regret Bound

Definition (Optimism in expectation)

We say a worth function $\widetilde{\mathsf{M}}_t$ is **optimistic in expectation (OIE)** if

$$\mathbb{E}\left[\left(\sup_{A\in\mathcal{A}_t} \widetilde{\mathsf{M}}_t(A) - \mathsf{B}_t\right)^2\right] \geq \mathsf{p}\,\mathbb{E}\left[\left(\sup_{A\in\mathcal{A}_t} \langle A, \Theta^\star\rangle - \mathsf{B}_t\right)^2\right].$$

Weaker Optimism and Regret Bound

Definition (Optimism in expectation)

We say a worth function $\widetilde{\mathrm{M}}_t$ is **optimistic in expectation (OIE)** if

$$\mathbb{E}\left[\left(\sup_{A\in\mathcal{A}_t}\widetilde{\mathrm{M}}_t(A)-\mathrm{B}_t\right)^2\right]\geq \mathrm{p}\,\mathbb{E}\left[\left(\sup_{A\in\mathcal{A}_t}\langle A,\Theta^\star\rangle-\mathrm{B}_t\right)^2\right].$$

Theorem

For a sequence of OIE worth functions $(\widetilde{\mathrm{M}}_t)_{t=1}^T$*, we have*

$$\mathrm{Regret}(T,\pi^{\mathrm{ROFUL}})\leq\widetilde{\mathcal{O}}\left(\rho\sqrt{\frac{dT}{\mathrm{p}}}\right).$$

Conclusion

- Proved that LinTS without inflation can incur linear regret.
- Provided a general regret bound for confidence-based policies.
- Introduced a new algorithm that is less pessimistic than OFUL while enjoying similar regret bounds.

References I

- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2-3 (2002), pp. 235–256.
- Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. "Stochastic Linear Optimization under Bandit Feedback". In: *COLT*. 2008.
- John Langford and Tong Zhang. "The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information". In: *Advances in Neural Information Processing Systems 20*. Ed. by J. C. Platt et al. Curran Associates, Inc., 2008, pp. 817–824.
- Paat Rusmevichientong and John N Tsitsiklis. "Linearly parameterized bandits". In: *Mathematics of Operations Research* 35.2 (2010), pp. 395–411.

References II

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits". In: *Advances in Neural Information Processing Systems*. 2011, pp. 2312–2320.
- Shipra Agrawal and Navin Goyal. "Thompson Sampling for Contextual Bandits with Linear Payoffs.". In: *ICML (3)*. 2013, pp. 127–135.
- Alexander Goldenshluger and Assaf Zeevi. "A linear response bandit problem". In: *Stochastic Systems* 3.1 (2013), pp. 230–261.
- Daniel Russo and Benjamin Van Roy. "Learning to Optimize via Posterior Sampling". In: *Mathematics of Operations Research* 39.4 (2014), pp. 1221–1243. DOI: 10.1287/moor.2014.0650.
- Daniel Russo and Benjamin Van Roy. "An information-theoretic analysis of thompson sampling". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2442–2471.

References III

Marc Abeille, Alessandro Lazaric, et al. "Linear Thompson sampling revisited". In: *Electronic Journal of Statistics* 11.2 (2017), pp. 5165–5197.

Thank you!

Any questions?

A General Regret Bound

Definition (Strong optimism)

We say a worth function $\widetilde{\mathrm{M}}_t$ is **optimistic** if

$$\sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) \geq \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \tag{3}$$

with probability at least p.

A General Regret Bound

Definition (Strong optimism)

We say a worth function $\widetilde{\mathrm{M}}_t$ is **optimistic** if

$$\sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) \geq \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \tag{3}$$

with probability at least p.

Theorem

Let $(\widetilde{\mathrm{M}}_t)_{t=1}^T$ be a sequence of optimistic worth functions. Then, the regret of ROFUL with this worth function is bounded by

$$\mathrm{Regret}(T, \pi^{\mathrm{ROFUL}}) \leq \widetilde{\mathcal{O}}\left(\rho \sqrt{\frac{dT}{\mathrm{p}}}\right).$$

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t \rangle.$$

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t \rangle.$$

- We can decompose it as

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \langle A, \widehat{\Theta}_{t-1} - \Theta^\star \rangle + \langle A, \Theta^\star \rangle.$$

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t \rangle.$$

- We can decompose it as

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \langle A, \widehat{\Theta}_{t-1} - \Theta^\star \rangle + \langle A, \Theta^\star \rangle.$$

- Hence, we have

$$\sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) - \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \geq \widetilde{\mathrm{M}}_t(A_t^\star) - \langle A_t^\star, \Theta^\star \rangle$$

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t \rangle.$$

- We can decompose it as

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \langle A, \widehat{\Theta}_{t-1} - \Theta^\star \rangle + \langle A, \Theta^\star \rangle.$$

- Hence, we have

$$\begin{aligned} \sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) - \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle &\geq \widetilde{\mathrm{M}}_t(A_t^\star) - \langle A_t^\star, \Theta^\star \rangle \\ &= \langle A_t^\star, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \langle A_t^\star, \widehat{\Theta}_{t-1} - \Theta^\star \rangle \,. \end{aligned}$$

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t \rangle.$$

- We can decompose it as

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \langle A, \widehat{\Theta}_{t-1} - \Theta^\star \rangle + \langle A, \Theta^\star \rangle.$$

- Hence, we have

$$\sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) - \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \geq \widetilde{\mathrm{M}}_t(A_t^\star) - \langle A_t^\star, \Theta^\star \rangle$$

$$= \langle A_t^\star, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \underbrace{\langle A_t^\star, \widehat{\Theta}_{t-1} - \Theta^\star \rangle}_{\text{Error term}}.$$

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t \rangle.$$

- We can decompose it as

$$\widetilde{\mathrm{M}}_t(A) = \langle A, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \langle A, \widehat{\Theta}_{t-1} - \Theta^\star \rangle + \langle A, \Theta^\star \rangle.$$

- Hence, we have

$$\sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) - \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \geq \widetilde{\mathrm{M}}_t(A_t^\star) - \langle A_t^\star, \Theta^\star \rangle$$

$$= \underbrace{\langle A_t^\star, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle}_{\text{Compensation term}} + \underbrace{\langle A_t^\star, \widehat{\Theta}_{t-1} - \Theta^\star \rangle}_{\text{Error term}}.$$

A Sufficient Condition for Optimism

Define

- Error vector $E := \Theta^\star - \widehat{\Theta}_{t-1}$
- Compensator vector $C := \widetilde{\Theta}_t - \widehat{\Theta}_{t-1}$

The optimism assumption holds if, with probability p, the following holds

$$\langle A_t^\star, C \rangle \geq \langle A_t^\star, E \rangle.$$

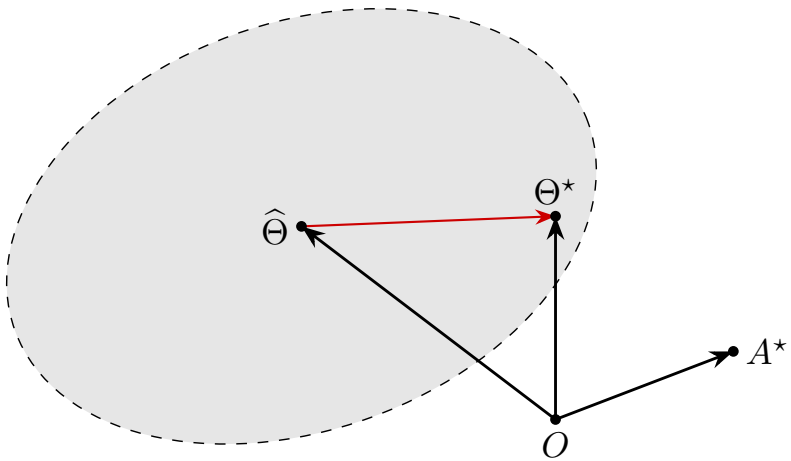
Omniscient Adversary and LinTS

- An **adversary** chooses $\mathcal{A}_t$ at time t.
- The adversary is **omniscient** if he knows $\widehat{\Theta}_{t-1}$ and $\Theta^\star$.

Omniscient Adversary and LinTS

- An **adversary** chooses $\mathcal{A}_t$ at time t.
- The adversary is **omniscient** if he knows $\widehat{\Theta}_{t-1}$ and $\Theta^\star$.
- He chooses $A = -c\widehat{\Theta}_{t-1} + E$ so that

$$\langle A, \Theta^\star \rangle > 0 \qquad \text{and} \qquad \langle A, \widehat{\Theta}_{t-1} \rangle < -\frac{1}{2} \cdot \|A\|_{\mathbf{V}_{t-1}^{-1}} \cdot \underbrace{\|E\|_{\mathbf{V}_{t-1}}}_{\approx\sqrt{d}} \ll 0.$$



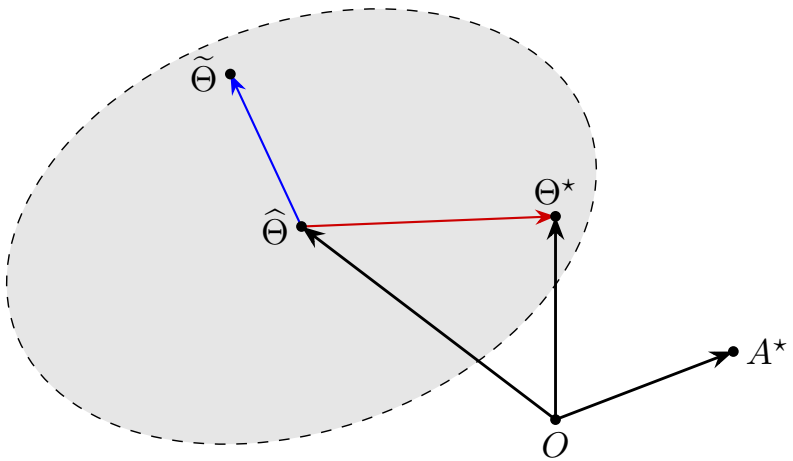
Omniscient Adversary and LinTS

- The adversary sets $\mathcal{A}_t = \{0, A\}$.
- LinTS chooses A if and only if

$$\langle A, \widetilde{\Theta}_t \rangle = \langle A, \widetilde{\Theta}_t - \widehat{\Theta}_{t-1} \rangle + \langle A, \widehat{\Theta}_{t-1} \rangle > 0.$$

- This requires

$$\langle A, C \rangle \sim \mathcal{N}(0, \mathbf{V}_{t-1}^{-1}) > \frac{1}{2} \cdot \|A\|_{\mathbf{V}_{t-1}^{-1}} \cdot \underbrace{\|E\|_{\mathbf{V}_{t-1}}}_{\approx\sqrt{d}}.$$



Omniscient Adversary and LinTS

- Next, we have

$$\mathbb{P}(\langle A, \widetilde{\Theta}_t \rangle > 0) \leq \exp\left(-\Omega(d)\right)!$$

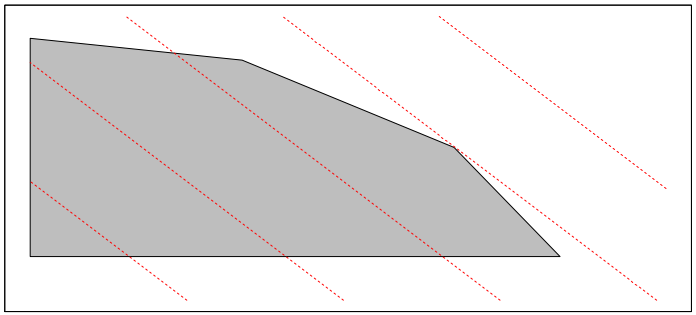
- LinTS chooses the optimal arm A w.p. **exponentially small in** $\Omega(d)$.
- When $\widetilde{A}_t = 0$, the reward contains **no new information** about $\Theta^\star$.
- The adversary reveals the same action set in the next rounds.
- The regret will grow **linearly**.

Bayesian Analyses are Brittle

- The key point was the **adversary's knowledge of** E.
- This can be relaxed by **slightly modifying** the noise distribution.
- **Reducing the noise variance** reveals information about E.

Why is LinTS Popular?

- **Computation efficiency**: when $\mathcal{A}_t$ is a polytope $\cdots$
  - LinTS solves an LP problem,



  - OFUL becomes an NP-hard problem!

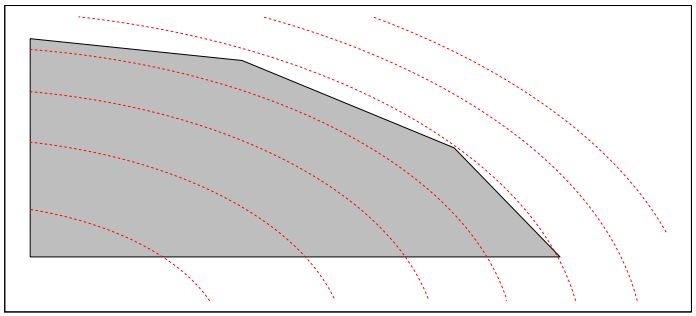


Photo credit: Russo and Van Roy 2014