On Worst-Case Regret of Linear Thompson Sampling

Nima Hamidi

Stanford University

Collaborator: Mohsen Bayati

Preprint: arXiv 2006.06790

Overview

1. Problem Definition
2. Confidence-based Policies
3. Failure of LinTS ☹
4. Positive Results ☺

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

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

Evaluation Metric

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

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

Evaluation Metric

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

- The **cumulative regret** is defined as

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

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

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

Algorithms

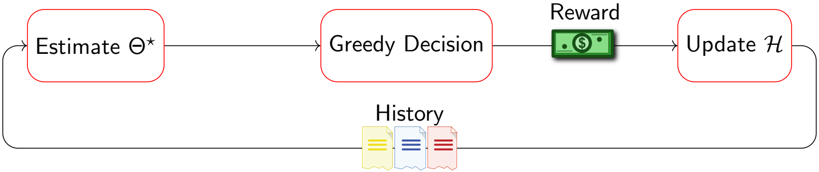
Greedy

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):

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

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

for $t = 1$ to T **do**

  Pull $\widetilde{A}_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widehat{\Theta}_{t-1} \rangle$

  Observe the reward r_t

  Compute $\mathbf{V}_t = \lambda \mathbb{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$

  Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$

end for

Greedy

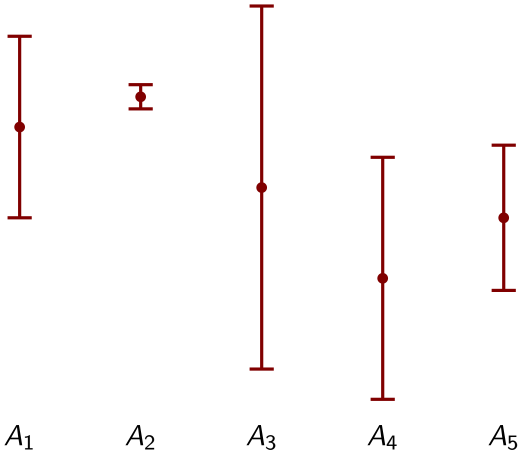
Algorithm 1 Greedy algorithm

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 $\mathbf{V}_t = \lambda \mathbb{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
5: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i \right)$
6: **end for**

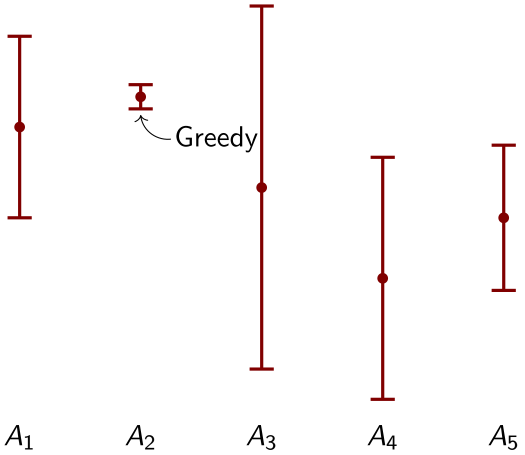
Greedy makes wrong decisions due to **over**- or **under-estimating** the true rewards.

- 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** $\mathcal{C}_t(\rho)$ to be

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

where

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

Optimism in Face of Uncertainty (OFU) Algorithm

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

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

where

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

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

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

Optimism in Face of Uncertainty (OFU) Algorithm

Algorithm 2 OFUL algorithm

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

Optimism in Face of Uncertainty (OFU) Algorithm

Algorithm 2 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 $\mathbf{V}_t = \lambda \mathbb{I} + \sum_{i=1}^t \widetilde{A}_i \widetilde{A}_i^\top$

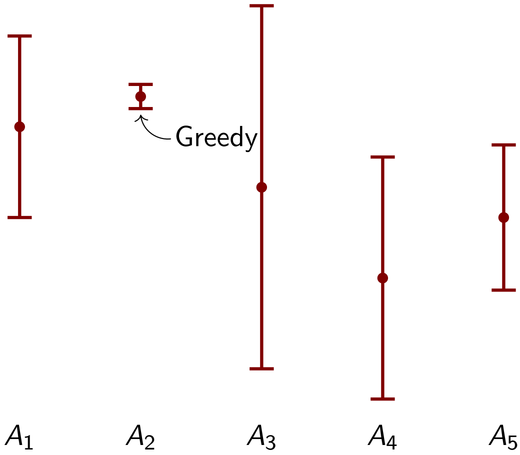
5: Compute $\widehat{\Theta}_t = \mathbf{V}_t^{-1} \left(\sum_{i=1}^t \widetilde{A}_i r_i \right)$

6: **end for**

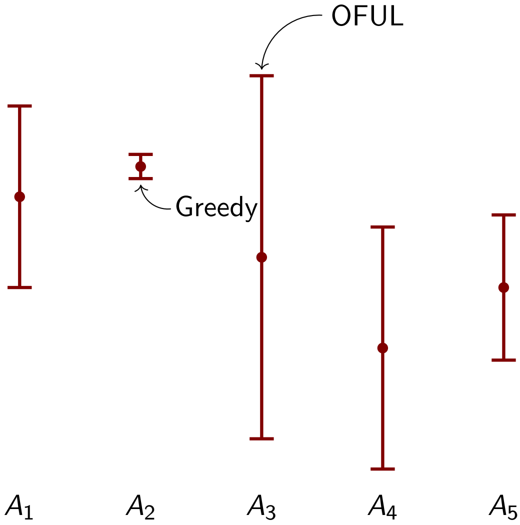
It can be shown that

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

Optimism in Face of Uncertainty (OFU) Algorithm



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 3 LinTS algorithm

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

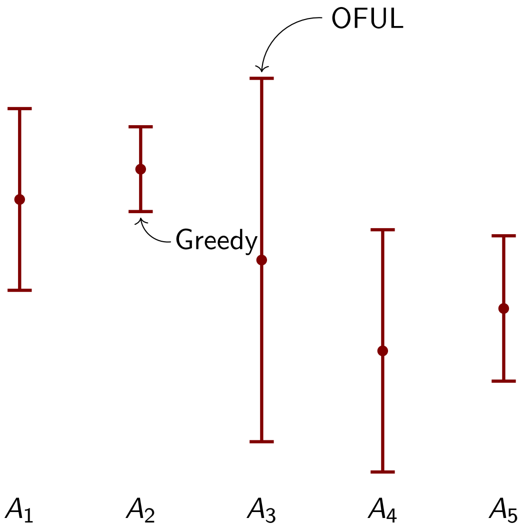
Linear Thompson Sampling (LinTS) Algorithm

- Under **normality**, LinTS becomes:

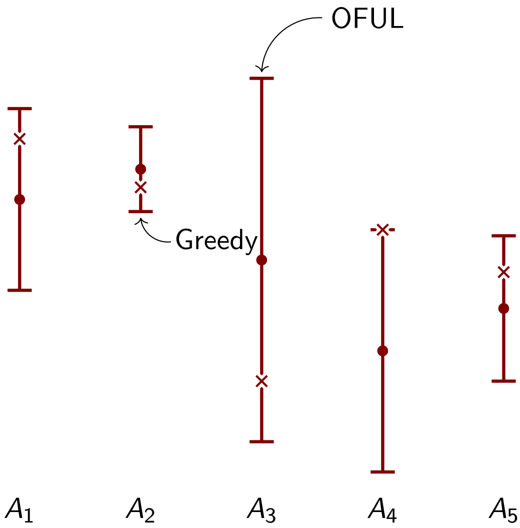
Algorithm 4 LinTS algorithm under normality

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

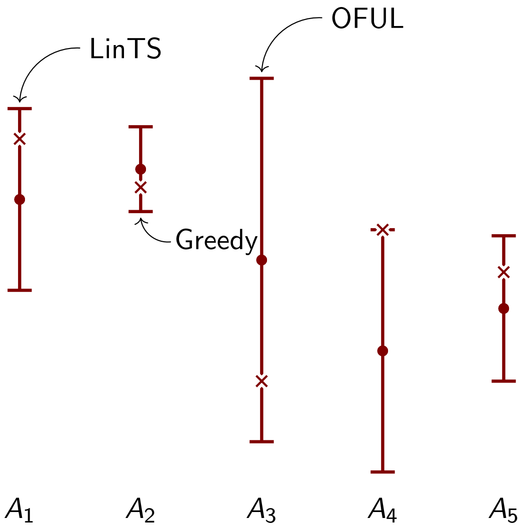
Linear Thompson Sampling (LinTS) Algorithm



Linear Thompson Sampling (LinTS) Algorithm

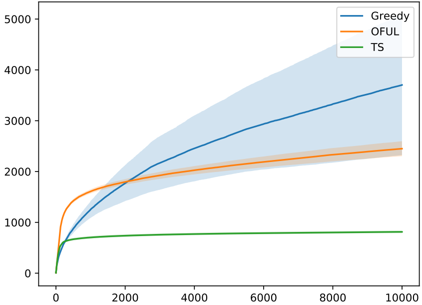


Linear Thompson Sampling (LinTS) Algorithm



Why Is LinTS Popular?

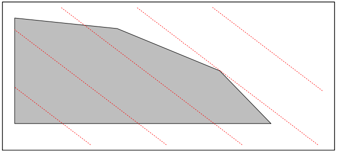
- **Empirical superiority**:
  - $d = 120$, $\Theta^\star \sim \mathcal{N}(0, \mathbb{I}_d)$,
  - $k = 10$, $X \sim \mathcal{N}(0, \mathbb{I}_{12})$,
  - Each A_t contains X as a block[1].



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

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!

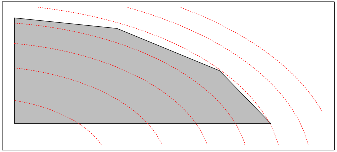


Photo credit: Russo and Van Roy 2014

Comparison of Regret Bounds

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

Under some conditions, the regret of OFUL is bounded by

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

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 5 LinTS algorithm under normality

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

A Worst-Case Regret Bound for LinTS

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

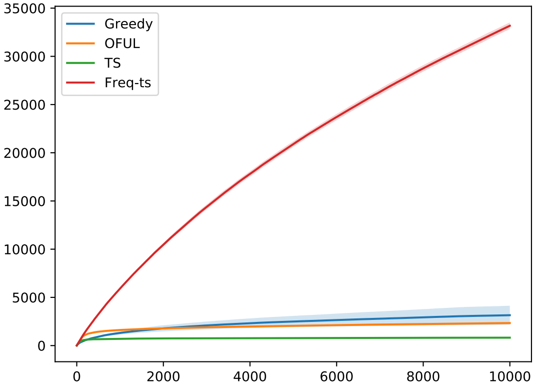
If $\beta \propto \sqrt{d}$, then

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

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

Empirical Performance of Inflated LinTS

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



A General Regret Bound

Randomized OFUL

- 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 $1 - \frac{1}{T^2}$.

Randomized OFUL

- 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 6 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 $\mathbf{V}_t$ and $\widehat{\Theta}_t$ via Eqs. (1) and (2)
5: **end for**

ROFUL Representations

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

Definition

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

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

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

- We can decompose it as

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

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

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

- We can decompose it as

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

- Hence, we have

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

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

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

- We can decompose it as

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

- Hence, we have

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

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

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

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

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

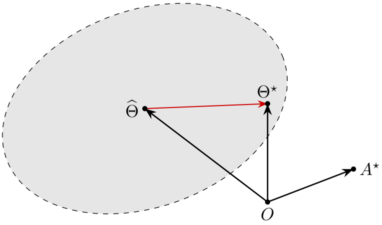
Omniscient Adversary and LinTS

- 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

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



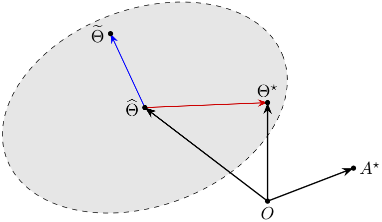
Omniscient Adversary and LinTS

- 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.

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*

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

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:

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

Reducing the Inflation Parameter

Reducing the Inflation Parameter

- Recall that a sufficient condition for optimism is that

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

with probability $p > 0$.

Reducing the Inflation Parameter

- Recall that a sufficient condition for optimism is that

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

with probability $p > 0$.

- Also, we have that

$$\langle A_t^\star, C \rangle \sim \mathcal{N}(0, \beta^2 \|A_t^\star\|_{\mathbf{V}_{t-1}}^2).$$

Reducing the Inflation Parameter

- Recall that a sufficient condition for optimism is that

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

  with probability $p > 0$.

- Also, we have that

$$\langle A_t^\star, C \rangle \sim \mathcal{N}(0, \beta^2 \|A_t^\star\|_{\mathbf{V}_{t-1}}^2).$$

- And, in the **worst-case**, we have

$$\langle A_t^\star, E \rangle \geq \rho \|A_t^\star\|_{\mathbf{V}_{t-1}}.$$

Reducing the Inflation Parameter

- Recall that a sufficient condition for optimism is that

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

with probability $p > 0$.

- Also, we have that

$$\langle A_t^\star, C \rangle \sim \mathcal{N}(0, \beta^2 \|A_t^\star\|^2_{\mathbf{V}_{t-1}}).$$

- And, in the **worst-case**, we have

$$\langle A_t^\star, E \rangle \geq \rho \|A_t^\star\|_{\mathbf{V}_{t-1}}.$$

- What if we assume that $A_t^\star$ is in a **random** direction?

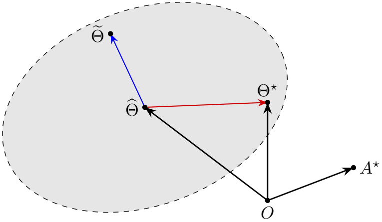
Diversity Assumption

Assumption (Optimal arm diversity)

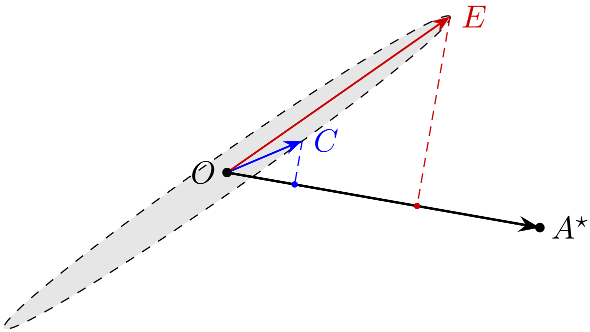
Assume that for any $V \in \mathbb{R}^d$ with $\|V\|_2 = 1$, we have

$$\mathbb{P}\left(\langle A_t^\star, V \rangle > \frac{\nu}{\sqrt{d}} \|A_t^\star\|_2 \right) \leq \frac{1}{t^3},$$

for some fixed $\nu \in [1, \sqrt{d}]$.



Diversity is not Sufficient



Improved Worst-Case Regret Bound for LinTS

Define **thinness** of a matrix $\mathbf{\Sigma}$ to be

$$\psi(\mathbf{\Sigma}) := \sqrt{\frac{d \cdot \|\mathbf{\Sigma}\|_{\text{op}}}{\|\mathbf{\Sigma}\|_*}}.$$

Improved Worst-Case Regret Bound for LinTS

Define **thinness** of a matrix $\mathbf{\Sigma}$ to be

$$\psi(\mathbf{\Sigma}) := \sqrt{\frac{d \cdot \|\mathbf{\Sigma}\|_{\text{op}}}{\|\mathbf{\Sigma}\|_*}}.$$

Assumption

For $\Psi, \omega > 0$, we have

$$\mathbb{P}\left(\|A^\star\|_{\mathbf{V}_t^{-1}} < \omega \sqrt{\frac{\|\mathbf{V}_t^{-1}\|_*}{d}} \cdot \|A^\star\|_2 \right) \leq \frac{1}{t^3}$$

for any positive definite $\mathbf{V}_t^{-1}$ with $\psi(\mathbf{V}_t^{-1}) \leq \Psi$.

Main Results

For $\beta := \frac{\nu\Psi}{\omega} \cdot \frac{\rho}{\sqrt{d}}$, optimism holds. So, we have the following result:

Theorem

If $\sum_{t=1}^{T} \mathbb{P}\left(\psi(\mathbf{V}_t^{-1}) > \Psi\right) \leq C$, *we have*

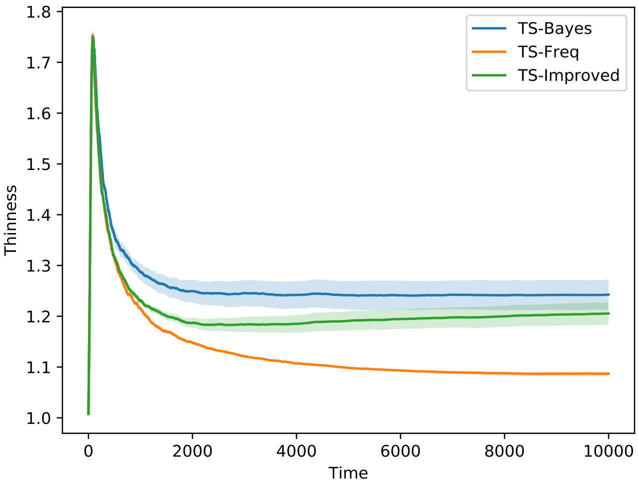
$$\text{Regret}(T, \Theta^\star, \pi^{TS}) \leq \mathcal{O}\left(\rho\beta\sqrt{dT\log(T)} + C\right).$$

Empirical Scrutiny on Thinness

Thinness in the simulations in Russo and Van Roy (2014):

Empirical Scrutiny on Thinness

Thinness in the simulations in Russo and Van Roy (2014):



Conclusion

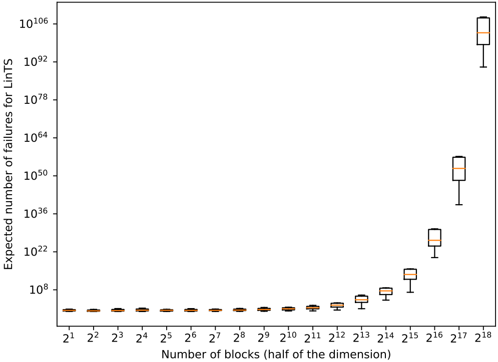
- Proved that LinTS without inflation can incur linear regret.
- Provided a general regret bound for confidence-based policies.
- Introduced sufficient conditions for reducing the inflation parameter.

Thank you!

Any questions?

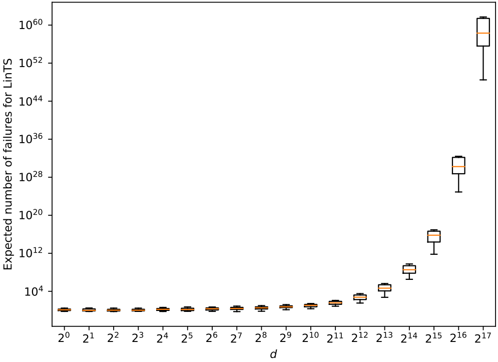
Failure of LinTS: Example 1

	Environment	LinTS
Prior	$\mathcal{N}(0, \mathbb{I}_d)$	$\mathcal{N}(0, \mathbb{I}_d)$
Noise	$\mathcal{N}(0, 0)$	$\mathcal{N}(0, 1)$



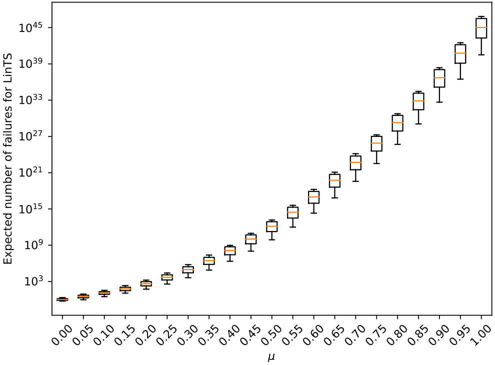
Failure of LinTS: Example 2

	Environment	LinTS
Prior	$\mathcal{N}(0.1 \cdot \mathbf{1}_d, \mathbb{I}_d)$	$\mathcal{N}(0, \mathbb{I}_d)$
Noise	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$



Failure of LinTS: Example 2

	Environment	LinTS
Prior	$\mathcal{N}(\mu \cdot \mathbf{1}_{2000}, \mathbb{I}_{2000})$	$\mathcal{N}(0, \mathbb{I}_{2000})$
Noise	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$



References I

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.

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

Shipra Agrawal and Navin Goyal. "Thompson Sampling for Contextual Bandits with Linear Payoffs.". In: *ICML (3)*. 2013, pp. 127–135.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. "Stochastic Linear Optimization under Bandit Feedback". In: *COLT*. 2008.

Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

References II

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.