Minimax Regret Bounds for Stochastic Linear Bandit Algorithms

Nima Hamidi

Stanford University

Advisor: Mohsen Bayati

Overview

Summary of Results

- Low-rank matrix estimation
  - On Low-rank Trace Regression under General Sampling Distribution (Submitted)
- Multi-armed bandits with many arms
  - Personalizing Many Decisions with High-dimensional Covariates (Neurips 2019)
  - The Unreasonable Effectiveness of Greedy Algorithms in Multi-Armed Bandit with Many Arms (Neurips 2020)

Summary of Results

- Low-rank matrix estimation
  - On Low-rank Trace Regression under General Sampling Distribution (Submitted)
- Multi-armed bandits with many arms
  - Personalizing Many Decisions with High-dimensional Covariates (Neurips 2019)
  - The Unreasonable Effectiveness of Greedy Algorithms in Multi-Armed Bandit with Many Arms (Neurips 2020)
- Stochastic linear bandits
  - A General Framework to Analyze Stochastic Linear Bandit (Submitted)
  - On Worst-case Regret of Linear Thompson Sampling (Submitted)
  - The Randomized Elliptical Potential Lemma with an Application to Linear Thompson Sampling (Submitted)

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.

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.

This model includes the following important special cases:

- **Multi-armed bandits (MAB)**
- **Contextual bandits**

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

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

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

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

- Regret grows at most **linearly in T** and grows **sublinearly (typically as $\sqrt{T}$)** for well-designed algorithms.

Summary of Results on Stochastic Linear Bandits

- Introducing of a meta algorithm, called **Randomized OFUL (ROFUL)** with the following special cases:
  - OFUL (Linear variant of UCB Lai and Robbins 1985)
  - Linear TS
  - Sieved-Greedy (a new algorithm)

- Introducing a notion of **optimism** under which near-optimal Bayesian and frequentist regret bounds can be obtained for ROFUL.

- Proving $\mathcal{O}(\text{poly}(\log T))$ regret bounds for ROFUL (and thus OFUL and LinTS) under **a so-called margin condition**. (Similar to Goldenshluger and Zeevi 2013)

Summary of Results on Stochastic Linear Bandits

- Proving a **stochastic variant** of elliptical potential which leads to an $\mathcal{O}(d\sqrt{T \log T})$ **Bayesian regret bound** for LinTS (with changing action sets).

- Proving that the **worst-case regret of LinTS** can grow linearly in T.

- Presenting **robust conditions** under which the worst-case regret of LinTS can be improved.

Algorithms

Overview of Algorithms

There are several algorithms proposed for linear/contextual bandits:

- ϵ-greedy algorithms:
  - Greedy algorithm
  - ϵ-greedy (Goldenshluger and Zeevi 2013)
  - ϵ-greedy and studentized test statistic for arm elimination (Kim, Lai, and Xu 2021)
- UCB-based algorithms (Abbasi-Yadkori, Pál, and Szepesvári 2011; Auer 2003; Dani, Hayes, and Kakade 2008; Li, Wang, and Zhou 2019)
- Thompson sampling / randomized algorithms:
  - Bayesian analysis (Russo and Van Roy 2016, 2014)
  - Frequentist analysis (Abeille and Lazaric 2017; Agrawal and Goyal 2013)

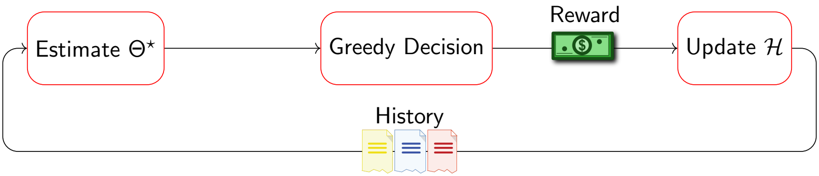
Greedy

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):

$$\widehat{\Theta}_t := \left(\lambda \mathbb{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top\right)^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i\right) \in \mathbb{R}^d.$$

Greedy

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

$$\widehat{\Theta}_t := \left(\lambda\mathbb{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top\right)^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i\right) \in \mathbb{R}^d.$$

The following matrix also encodes the **uncertainty about each direction**:

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

Greedy

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

$$\widehat{\Theta}_t := \left(\lambda\mathbb{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top\right)^{-1} \left(\sum_{i=1}^{t} \widetilde{A}_i r_i\right) \in \mathbb{R}^d.$$

The following matrix also encodes the **uncertainty about each direction**:

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

The magnitude of the estimation error in the direction X is proportional to

$$\|X\|_{\mathbf{V}_t^{-1}} := \sqrt{X^\top \mathbf{V}_t^{-1} X}.$$

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

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 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**.

Optimism in Face of Uncertainty (OFU) Algorithm

- The variant of UCB (Lai and Robbins 1985) for linear bandits.
- Key idea: **be optimistic** when estimating the reward of actions.

Optimism in Face of Uncertainty (OFU) Algorithm

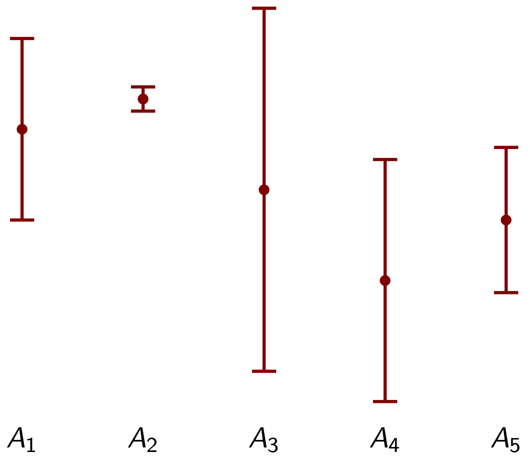
- The variant of UCB (Lai and Robbins 1985) for linear bandits.
- Key idea: **be optimistic** when estimating the reward of actions.

Algorithm 2 OFUL algorithm

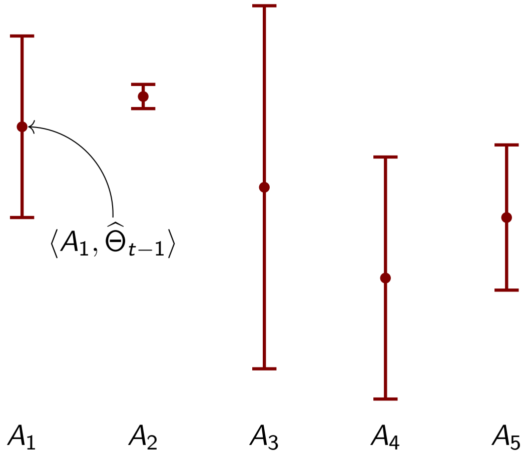
1: **for** $t = 1$ to T **do**
2: Pull $\widetilde{A}_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widehat{\Theta}_t \rangle + \rho \|A\|_{\mathbf{V}_{t-1}^{-1}}$
3: Observe the reward r_t
4: Compute $\mathbf{V}_t = \lambda \mathbb{I} + \sum_{i=1}^{t} \widetilde{A}_i \widetilde{A}_i^\top$
5: Compute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

Guarantees for OFUL require ρ to be of order $\widetilde{\mathcal{O}}(\sqrt{d})$.

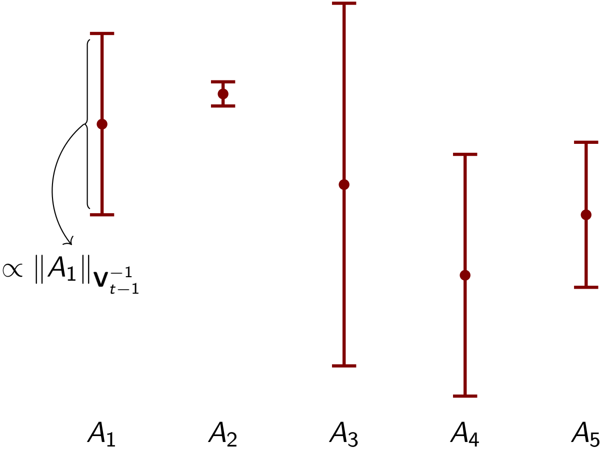
Greedy vs OFUL



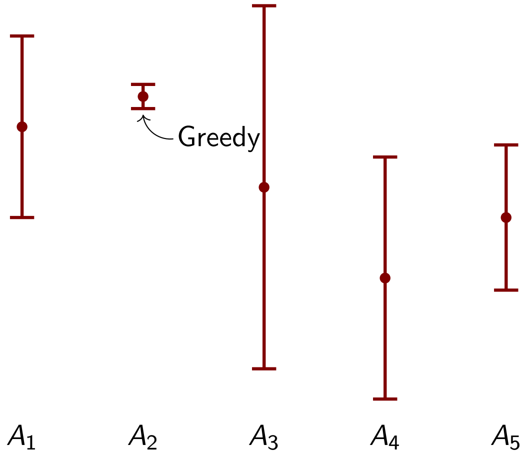
Greedy vs OFUL



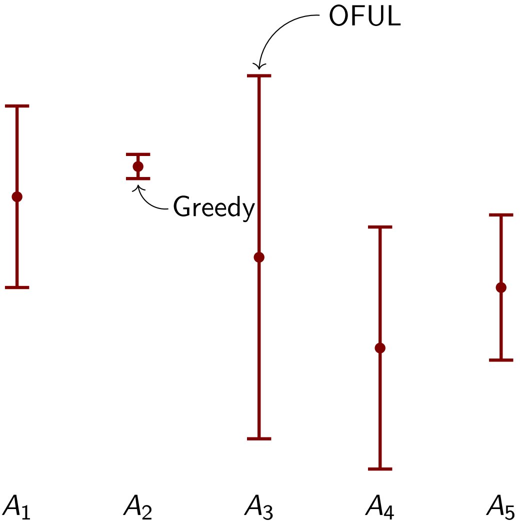
Greedy vs OFUL



Greedy vs OFUL



Greedy vs OFUL



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 is a **Bayesian heuristic** and assumes $\Theta^\star$ is sampled from a **prior distribution**.
- LinTS gets the **prior distribution** and **noise distributions** as input.
- LinTS samples from the **posterior** distribution of $\Theta^\star$.

Linear Thompson Sampling (LinTS) Algorithm

- Key idea: use **randomization** to address under-estimation.
- LinTS is a **Bayesian heuristic** and assumes $\Theta^\star$ is sampled from a **prior distribution**.
- LinTS gets the **prior distribution** and **noise distributions** as input.
- 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

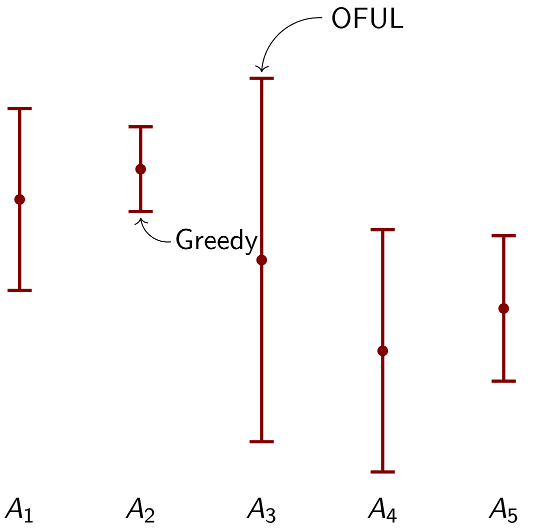
Linear Thompson Sampling (LinTS) Algorithm

- Under **normality**, LinTS becomes:

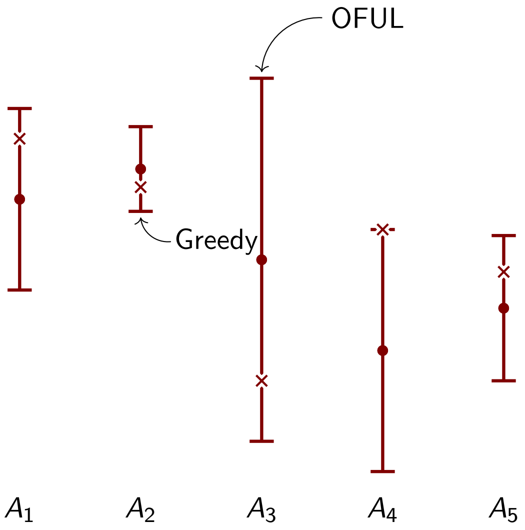
Algorithm 4 LinTS algorithm under normality

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**

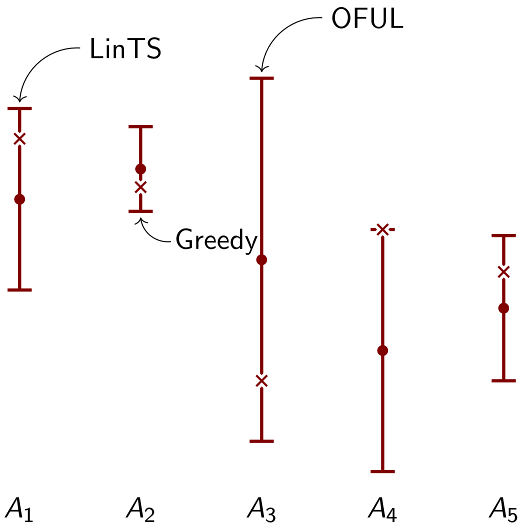
Linear Thompson Sampling (LinTS) Algorithm



Linear Thompson Sampling (LinTS) Algorithm

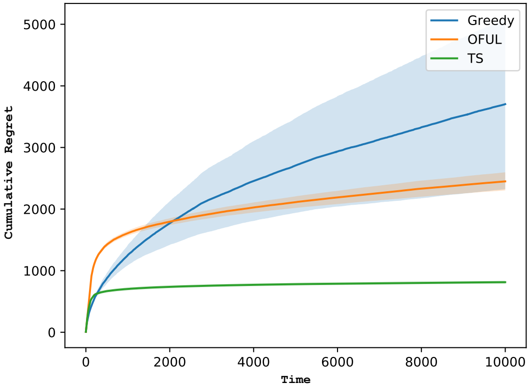


Linear Thompson Sampling (LinTS) Algorithm



Why Is LinTS Popular?

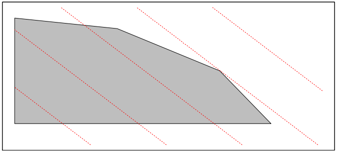
- **Empirical superiority**:
  - $d = 120$, $\Theta^\star \sim \mathcal{N}(0, \mathbb{I}_d)$,
  - $k = 10$, $X \sim \mathcal{N}(0, \mathbb{I}_{12})$,
  - Each A_t contains X as a block[1].



[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!

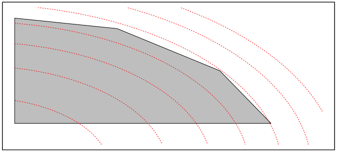


Photo credit: Russo and Van Roy 2014

Comparison of Regret Bounds

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

Under some conditions, the regret of OFUL is bounded by

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

Comparison of Regret Bounds

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

Under some conditions, the regret of OFUL is bounded by

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

Theorem (Russo and Van Roy 2014)

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

$$\mathrm{BayesRegret}(T, \mathcal{P}, \pi^{LinTS}) \leq \mathcal{O}(d\sqrt{T}\log T).$$

Comparison of Regret Bounds

Theorem (Dani, Hayes, and Kakade 2008)

There is a Bayesian linear bandit problem with a **fixed action set** *that satisfies*

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

Comparison of Regret Bounds

Theorem (Dani, Hayes, and Kakade 2008)

There is a Bayesian linear bandit problem with a **fixed action set** *that satisfies*

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

Theorem (Li, Wang, and Zhou 2019)

There is a Bayesian linear bandit problem with **changing action sets** *that satisfies*

$$\inf_{\pi} \mathrm{BayesRegret}(T, \mathcal{P}, \pi) \geq \Omega(d\sqrt{T \log T}).$$

Comparison of Regret Bounds

Theorem (Dong and Van Roy 2018)

When **the action set is fixed**, *the Bayesian regret of LinTS is bounded by*

$$\mathrm{BayesRegret}(T, \mathcal{P}, \pi^{LinTS}) \leq \mathcal{O}(d\sqrt{T \log T}).$$

Comparison of Regret Bounds

Theorem (Dong and Van Roy 2018)

When **the action set is fixed***, the Bayesian regret of LinTS is bounded by*

$$\mathrm{BayesRegret}(T, \mathcal{P}, \pi^{LinTS}) \leq \mathcal{O}(d\sqrt{T \log T}).$$

Theorem (Hamidi and Bayati 2021)

Under mild assumptions, the Bayesian regret of LinTS is bounded by (even when **the action sets changes***)*

$$\mathrm{BayesRegret}(T, \mathcal{P}, \pi) \leq \mathcal{O}(d\sqrt{T \log T}).$$

Worst-Case Regret Bounds for LinTS

- Near-optimal worst-case (and Bayesian) regret bounds are known for OFUL.
- Near-optimal Bayesian regret bounds are also known for LinTS.
- **Question:** can one prove a similar **worst-case regret bound for LinTS**?

Worst-Case Regret Bounds for LinTS

- Near-optimal worst-case (and Bayesian) regret bounds are known for OFUL.
- Near-optimal Bayesian regret bounds are also known for LinTS.
- **Question:** can one prove a similar **worst-case regret bound for LinTS**?
- The only known results require **inflating** the posterior variance.

A Worst-Case Regret Bound for LinTS

Algorithm 5 LinTS(β) 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**

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

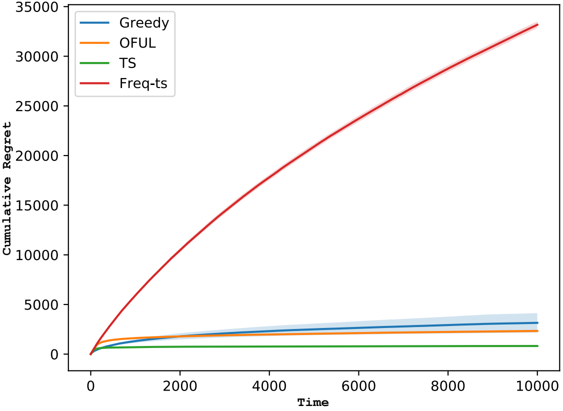
If $\beta \propto \sqrt{d}$, then

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

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

Empirical Performance of Inflated LinTS

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



When LinTS Fails!

Construction of Counter-examples

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

Theorem (Hamidi and Bayati 2020)

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

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

Construction of Counter-examples

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

Theorem (Hamidi and Bayati 2020)

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

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

The counter-example satisfies the following properties:

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

Construction of Counter-examples

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

Theorem (Hamidi and Bayati 2020)

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

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

The counter-example satisfies the following properties:

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

LinTS can fail even **by just improving the data**. We need more **robust guarantees**.

Construction of Counter-examples

- **Fact 1:** $\widetilde{\Theta}_t$ and $\Theta^\star$ are identically distributed conditional on $\mathcal{H}_{t-1}$.

Construction of Counter-examples

- **Fact 1:** $\widetilde{\Theta}_t$ and $\Theta^\star$ are identically distributed conditional on $\mathcal{H}_{t-1}$.
- **Fact 2:** $\widetilde{\Theta}_t$ and $\Theta^\star$ are identically distributed unconditionally.

Construction of Counter-examples

- **Fact 1:** $\widetilde{\Theta}_t$ and $\Theta^\star$ are identically distributed conditional on $\mathcal{H}_{t-1}$.
- **Fact 2:** $\widetilde{\Theta}_t$ and $\Theta^\star$ are identically distributed unconditionally.
- This can break under distributional mismatch.

Construction of Counter-examples

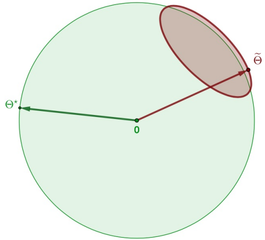
- We let $\Theta^\star \sim \mathcal{N}(0, \mathbb{I}_d)$.

- Under mismatch, there is a bandit problem with:

  - $\mathbb{E}[\widetilde{\Theta}_t] = c\mathbf{1}$ for some $c = \mathcal{O}(1) > 0$,

  - $\left\langle \widetilde{\Theta}_t - \mathbb{E}[\widetilde{\Theta}_t], \frac{\mathbf{1}}{\sqrt{d}} \right\rangle$ is $\mathcal{O}(1)$-sub-Gaussian.

  - Therefore, we have $\left\langle \widetilde{\Theta}_t, \frac{\mathbf{1}}{\sqrt{d}} \right\rangle = c\sqrt{d} + \mathcal{O}(1)$ w.h.p.



Construction of Counter-examples

- Now, let $\mathcal{A}_t := \{0, A\}$ where $A := -\mathbf{1}/\sqrt{d}$.
- A is the optimal arm with probability $\frac{1}{2}$.

Construction of Counter-examples

- Now, let $\mathcal{A}_t := \{0, A\}$ where $A := -\mathbf{1}/\sqrt{d}$.
- A is the optimal arm with probability $\frac{1}{2}$.
- However, LinTS will choose A only if

$$\langle \widetilde{\Theta}_t, A \rangle = -c\sqrt{d} + \mathcal{O}(1)\text{-sub-Gaussian} > 0.$$

- This happens with probability $\exp(-Cd)$ for some constant $C > 0$.

Construction of Counter-examples

- Now, let $\mathcal{A}_t := \{0, A\}$ where $A := -\mathbf{1}/\sqrt{d}$.
- A is the optimal arm with probability $\frac{1}{2}$.
- However, LinTS will choose A only if

$$\langle \widetilde{\Theta}_t, A \rangle = -c\sqrt{d} + \mathcal{O}(1)\text{-sub-Gaussian} > 0.$$

- This happens with probability $\exp(-Cd)$ for some constant $C > 0$.
- Also, choosing 0 will reveal **no new information**.
- So, show **the same action set for all** t.

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

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 (Optimism – Informal)

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

$$\sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) \geq \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \tag{1}$$

with probability at least p.

A General Regret Bound

Definition (Optimism – Informal)

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

$$\sup_{A \in \mathcal{A}_t} \widetilde{\mathrm{M}}_t(A) \geq \sup_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle \tag{1}$$

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

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

Improving LinTS

Define **thinness** of a positive definite matrix $\mathbf{V}^{-1}$ to be

$$\psi(\mathbf{V}^{-1}) := \sqrt{\frac{d \cdot \|\mathbf{V}^{-1}\|_{\mathrm{op}}}{\|\mathbf{V}^{-1}\|_{*}}} \in [1, \sqrt{d}].$$

Algorithm 7 Improved LinTS algorithm

---

for $t = 1$ to T **do**
    if $\psi(\mathbf{V}_t^{-1}) \leq \Psi$ **then**
        Sample $\widetilde{\Theta}_t \sim \mathcal{N}(\widehat{\Theta}_{t-1}, \beta^2 \mathbf{V}_{t-1}^{-1})$
    else
        Sample $\widetilde{\Theta}_t \sim \mathcal{N}(\widehat{\Theta}_{t-1}, \rho^2 \mathbf{V}_{t-1}^{-1})$
    end if
    Pull $A_t := \arg\max_{A \in \mathcal{A}_t} \langle A, \widetilde{\Theta}_t \rangle$
    Observe the reward r_t
    Compute $\mathbf{V}_t$ and $\widehat{\Theta}_t$ as before.
end for

Main Result

The inflation parameter β can be small if the optimal arm:

- is not aligned with any given direction; and
- takes advantage of a small thinness parameter appropriately.

Theorem (Informal)

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

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

Empirical Scrutiny on Thinness

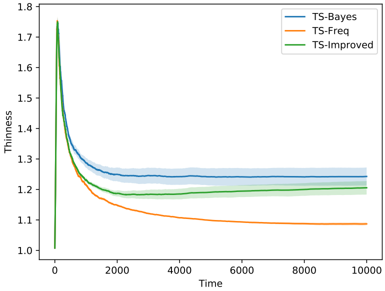
A case study – simulations in Russo and Van Roy (2014):

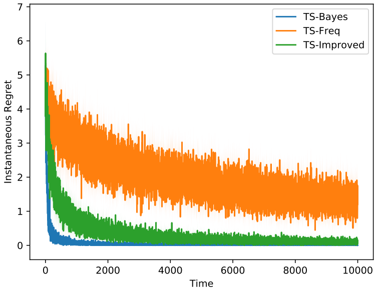
- $\Theta^\star \sim \mathcal{N}(0, \mathbb{I}_{100})$ and $\varepsilon_t \sim \mathcal{N}(0, 1)$,
- $\mathcal{A}_t$ consists of $k = 50$ random vectors in $\text{Unif}\big([-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]^d\big)$.

Empirical Scrutiny on Thinness

A case study – simulations in Russo and Van Roy (2014):

- $\Theta^\star \sim \mathcal{N}(0, \mathbb{I}_{100})$ and $\varepsilon_t \sim \mathcal{N}(0,1)$,
- $\mathcal{A}_t$ consists of $k = 50$ random vectors in $\mathrm{Unif}\left([-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]^d\right)$.





Intuitions Behind Improved LinTS

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

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

- 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, letting $A_t^\star := \arg\max_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle$, we have

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

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, letting $A_t^\star := \arg\max_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle$, we have

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

A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

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

- 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, letting $A_t^\star := \arg\max_{A \in \mathcal{A}_t} \langle A, \Theta^\star \rangle$, we have

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

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

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

Reducing the Inflation Parameter

- We have $C \sim \mathcal{N}(0, \beta^2 \mathbf{V}_{t-1}^{-1})$ which implies that $\|C\|_{\mathbf{V}_{t-1}} \approx \beta\sqrt{d}$.
- On the other hand, recall that $\|E\|_{\mathbf{V}_{t-1}} \approx \sqrt{d}$.

Reducing the Inflation Parameter

- We have $C \sim \mathcal{N}(0, \beta^2 \mathbf{V}_{t-1}^{-1})$ which implies that $\|C\|_{\mathbf{V}_{t-1}} \approx \beta\sqrt{d}$.
- On the other hand, recall that $\|E\|_{\mathbf{V}_{t-1}} \approx \sqrt{d}$.
- Next note that with high probability

$$\langle A_t^\star, C \rangle \propto \beta \|A_t^\star\|_{\mathbf{V}_{t-1}^{-1}}.$$

- Finally, in the **worst case**, we may get (by Cauchy-Schwartz)

$$\langle A_t^\star, E \rangle \propto \sqrt{d} \|A_t^\star\|_{\mathbf{V}_{t-1}^{-1}}.$$

Reducing the Inflation Parameter

- We have $C \sim \mathcal{N}(0, \beta^2 \mathbf{V}_{t-1}^{-1})$ which implies that $\|C\|_{\mathbf{V}_{t-1}} \approx \beta\sqrt{d}$.
- On the other hand, recall that $\|E\|_{\mathbf{V}_{t-1}} \approx \sqrt{d}$.
- Next note that with high probability

$$\langle A_t^\star, C \rangle \propto \beta \|A_t^\star\|_{\mathbf{V}_{t-1}^{-1}}.$$

- Finally, in the **worst case**, we may get (by Cauchy-Schwartz)

$$\langle A_t^\star, E \rangle \propto \sqrt{d} \|A_t^\star\|_{\mathbf{V}_{t-1}^{-1}}.$$

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

Diversity Assumption

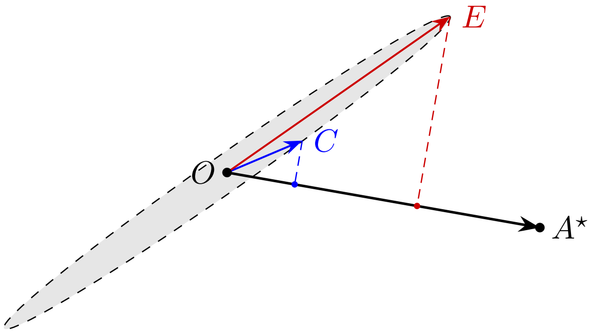
Assumption (Optimal arm diversity)

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

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

Improved Worst-Case Regret Bound for LinTS

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

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

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 $\psi(\mathbf{V}_t^{-1}) \leq \Psi$.

Conclusion

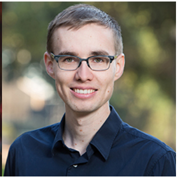
- Proved that LinTS without inflation can incur linear regret.
- Provided a general regret bound for confidence-based policies.
- Introduced sufficient conditions for reducing the inflation parameter.

Acknowledgements

Acknowledgements



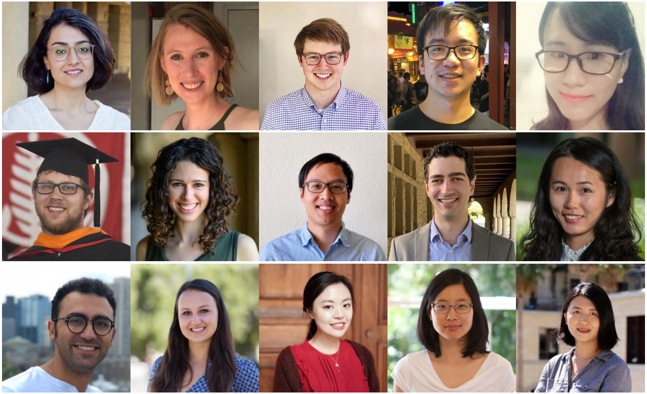




Acknowledgements

Acknowledgements

Acknowledgements



Acknowledgements

Acknowledgements

Acknowledgements

Acknowledgements

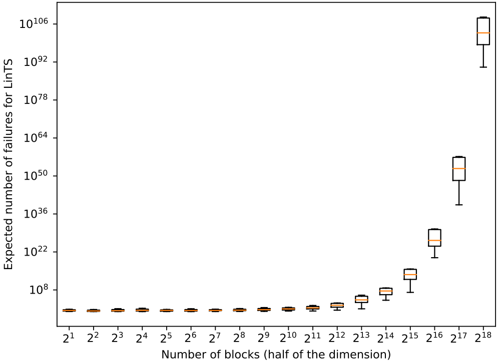
Acknowledgements

Thank you!

Any questions?

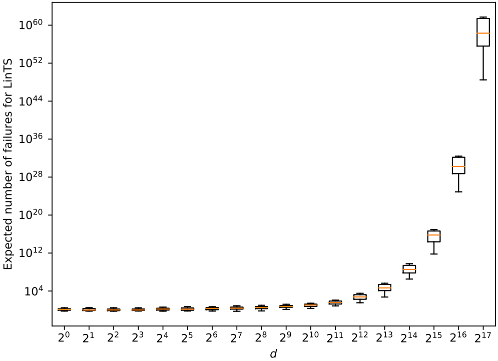
Failure of LinTS: Example 1

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



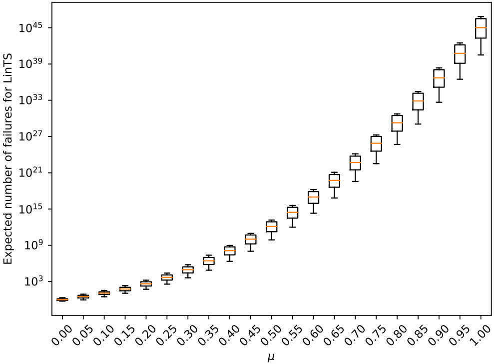
Failure of LinTS: Example 2

	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.
- Peter Auer. "Using confidence bounds for exploitation-exploration trade-offs". In: *Journal of Machine Learning Research* 3 (2003), pp. 397–422.
- Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. "Stochastic Linear Optimization under Bandit Feedback". In: *COLT*. 2008.

References II

Shi Dong and Benjamin Van Roy. "An information-theoretic analysis for Thompson sampling with many actions". In: *Advances in Neural Information Processing Systems*. 2018, pp. 4157–4165.

Alexander Goldenshluger and Assaf Zeevi. "A linear response bandit problem". In: *Stochastic Systems* 3.1 (2013), pp. 230–261.

Nima Hamidi and Mohsen Bayati. "On Worst-case Regret of Linear Thompson Sampling". In: *arXiv preprint arXiv:2006.06790* (2020). URL: https://arxiv.org/pdf/2006.06790.pdf.

Nima Hamidi and Mohsen Bayati. "The Randomized Elliptical Potential Lemma with an Application to Linear Thompson Sampling". In: *arXiv preprint arXiv:2102.07987* (2021). URL: https://arxiv.org/pdf/2102.07987.pdf.

Dong Woo Kim, Tze Leung Lai, and Huanzhong Xu. "Multi-Armed Bandits with Covariates: Theory and Applications". In: *Statistica Sinica* (2021).

References III

- Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.
- Yingkai Li, Yining Wang, and Yuan Zhou. "Nearly minimax-optimal regret for linearly parameterized bandits". In: *arXiv preprint arXiv:1904.00242* (2019).
- 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.
- 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.