Minimax Regret Bounds for Stochastic Linear Bandit Algorithms

Nima Hamidi

Stanford University

Advisor: Mohsen Bayati
Overview

1. Problem Definition
2. Confidence-based Policies
3. Failure of LinTS 😞
4. Positive Results 😊
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 $\Theta^* \in \mathbb{R}^d$ be fixed (and unknown).
- At time $t$, the action set $\mathcal{A}_t \subseteq \mathbb{R}^d$ is revealed to a policy $\pi$.
- The policy chooses $\tilde{A}_t \in \mathcal{A}_t$.
- It observes a reward $r_t = \langle \Theta^*, \tilde{A}_t \rangle + \varepsilon_t$.
- Conditional on the history, $\varepsilon_t$ has zero mean.
Stochastic Linear Bandit Problem

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

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

- The objective is to improve using past experiences.

- The **cumulative regret** is defined as

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

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

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

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

- The **cumulative regret** is defined as

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

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

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

- 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 $O(\text{poly}(\log T))$ 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 $O(d\sqrt{T \log T})$ **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:

- $\epsilon$-greedy algorithms:
  - Greedy algorithm
  - $\epsilon$-greedy (Goldenshluger and Zeevi 2013)
  - $\epsilon$-greedy and studentized test statistic for arm elimination (Kim, Lai, and Xu 2021)


- 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

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

- Construct an estimate $\hat{\Theta}_{t-1}$ for $\Theta^*$,

- Choose the action $A \in \mathcal{A}_t$ with largest $\langle A, \hat{\Theta}_{t-1} \rangle$. 

Diagram:

- Estimate $\Theta^*$
- Greedy Decision
- Reward
- Update $\mathcal{H}$
- History
Greedy

The **ridge estimator** is used to obtain \( \hat{\Theta}_t \) (for a fixed \( \lambda \)):

\[
\hat{\Theta}_t := \left( \lambda \mathbb{I} + \sum_{i=1}^{t} \tilde{A}_i \tilde{A}_i^\top \right)^{-1} \left( \sum_{i=1}^{t} \tilde{A}_i r_i \right) \in \mathbb{R}^d.
\]
The **ridge estimator** is used to obtain $\hat{\Theta}_t$ (for a fixed $\lambda$):

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

The following matrix also encodes the uncertainty about each direction:

$$V_t := \lambda I + \sum_{i=1}^{t} \tilde{A}_i \tilde{A}_i^\top \in \mathbb{R}^{d \times d}.$$
The **ridge estimator** is used to obtain $\hat{\Theta}_t$ (for a fixed $\lambda$):

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

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

$$V_t := \lambda I + \sum_{i=1}^{t} \tilde{A}_i \tilde{A}_i^\top \in \mathbb{R}^{d \times d}.$$ 

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

$$\|X\|_{V_t^{-1}} := \sqrt{X^\top V_t^{-1} X}.$$
Algorithm 1 Greedy algorithm

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

Algorithm 1 Greedy algorithm

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

Guarantees for OFUL require $\rho$ to be of order $\tilde{O}(\sqrt{d})$. 
Greedy vs OFUL

\[ \langle A_1, \hat{\theta}_t - 1 \rangle \propto \| A_1 \|_{V^{-1}} \]
\[ \langle A_1, \hat{\Theta}_{t-1} \rangle \]
Greedy vs OFUL

\[ \alpha \| A_1 \|_{\nu_{t-1}} \]

\[ A_1 \quad A_2 \quad A_3 \quad A_4 \quad A_5 \]
Greedy vs OFUL

\[ \langle A_1, \hat{\Theta}_{t-1} \rangle \propto \|A_1\| V_{t-1} \]

Greedy

\[ A_1 \quad \quad A_2 \quad \quad A_3 \quad \quad A_4 \quad \quad A_5 \]
Greedy vs OFUL

\[ \langle A_1, \hat{\Theta}_{t-1} \rangle \propto \left\| A_1 \right\| V^{-1} t^{-1} \]

- Greedy
- 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 $\Theta^*$ is sampled from a **prior distribution**.

- LinTS gets the **prior distribution** and **noise distributions** as input.

- LinTS samples from the **posterior** distribution of $\Theta^*$.

**Algorithm 3**

1: for $t = 1$ to $T$
2: Sample $\tilde{\Theta}_t \sim P(\Theta^* | H_{t-1})$
3: Pull $A_t := \text{arg max}_{A \in A_t} \langle A, \tilde{\Theta}_t \rangle$
4: Observe the reward $r_t$
5: Update $H_t \leftarrow H_{t-1} \cup \{ (A_t, r_t) \}$
6: end for
Linear Thompson Sampling (LinTS) Algorithm

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

Algorithm 3 LinTS algorithm

1: \textbf{for} $t = 1$ to $T$ \textbf{do}
2: \quad Sample $\tilde{\Theta}_t \sim P(\Theta^* | \mathcal{H}_{t-1})$
3: \quad Pull $A_t := \text{arg max}_{A \in A_t} \langle A, \tilde{\Theta}_t \rangle$
4: \quad Observe the reward $r_t$
5: \quad Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(A_t, r_t)\}$
6: \textbf{end for}
Under \textbf{normality}, LinTS becomes:

\begin{algorithm}
\caption{LinTS algorithm under normality}
\begin{algorithmic}[1]
\STATE \textbf{for} $t = 1$ to $T$ \textbf{do}
\STATE Sample $\tilde{\Theta}_t \sim \mathcal{N}(\hat{\Theta}_{t-1}, V_{t-1}^{-1})$
\STATE Pull $A_t := \arg \max_{A \in \mathcal{A}_t} \langle A, \tilde{\Theta}_t \rangle$
\STATE Observe the reward $r_t$
\STATE Compute $V_t = \lambda \mathbb{I} + \sum_{i=1}^t \tilde{A}_i \tilde{A}_i^\top$
\STATE Compute $\hat{\Theta}_t = V_t^{-1} \left( \sum_{i=1}^t \tilde{A}_i r_i \right)$
\STATE \textbf{end for}
\end{algorithmic}
\end{algorithm}
Linear Thompson Sampling (LinTS) Algorithm

A_1 \rightarrow OFUL

\rightarrow Greedy

A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5
Linear Thompson Sampling (LinTS) Algorithm

![Diagram of LinTS Algorithm]

- OFUL
- Greedy

Where:
- $A_1$
- $A_2$
- $A_3$
- $A_4$
- $A_5$
Linear Thompson Sampling (LinTS) Algorithm
Why Is LinTS Popular?

- **Empirical superiority:**
  - $d = 120$, $\Theta^* \sim \mathcal{N}(0, I_d)$,
  - $k = 10$, $X \sim \mathcal{N}(0, I_{12})$,
  - 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 $A_t$ is a polytope · · ·
  - 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

\[ \text{Regret}(T, \Theta^*, \pi^{OFUL}) \leq O(d \sqrt{T \log T}). \]
Comparison of Regret Bounds

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

*Under some conditions, the regret of OFUL is bounded by*

\[
\text{Regret}(T, \Theta^*, \pi^{OFUL}) \leq O(d\sqrt{T \log T}).
\]

**Theorem (Russo and Van Roy 2014)**

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

\[
\text{BayesRegret}(T, \mathcal{P}, \pi^{LinTS}) \leq O(d\sqrt{T \log T}).
\]
Comparison of Regret Bounds

Theorem (Dani, Hayes, and Kakade 2008)

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

$$\inf_{\pi} \text{BayesRegret}(T, \mathcal{P}, \pi) \geq \Omega(d\sqrt{T}).$$
Comparison of Regret Bounds

Theorem (Dani, Hayes, and Kakade 2008)

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

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

Theorem (Li, Wang, and Zhou 2019)

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

$$\inf_{\pi} \text{BayesRegret}(T, \mathcal{P}, \pi) \geq \Omega(d \sqrt{T \log T}).$$
Comparison of Regret Bounds

Theorem (Dong and Van Roy 2018)

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

$$\text{BayesRegret}(T, P, \pi^{\text{LinTS}}) \leq O(d \sqrt{T \log T}).$$
Comparison of Regret Bounds

Theorem (Dong and Van Roy 2018)

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

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

Theorem (Hamidi and Bayati 2021)

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

$$\text{BayesRegret}(T, \mathcal{P}, \pi) \leq O(d \sqrt{T \log T}).$$
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?
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($\beta$) algorithm under normality

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

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

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

$$\text{Regret}(T, \Theta^*, \pi^{\text{LinTS} (\beta)}) \leq \tilde{O}(d \sqrt{d T}).$$

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

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

<table>
<thead>
<tr>
<th>Environment</th>
<th>What LinTS assumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>( \mathcal{N}(0, I_d) )</td>
</tr>
<tr>
<td>Noise</td>
<td>( \mathcal{N}(0, 0) )</td>
</tr>
</tbody>
</table>
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

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

The counter-example satisfies the following properties:

<table>
<thead>
<tr>
<th>Environment</th>
<th>What LinTS assumes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>( \mathcal{N}(0, I_d) )</td>
</tr>
<tr>
<td>Noise</td>
<td>( \mathcal{N}(0, 0) )</td>
</tr>
</tbody>
</table>

LinTS can fail even by just improving the data. We need more robust guarantees.
Construction of Counter-examples

- **Fact 1:** $\tilde{\Theta}_t$ and $\Theta^*$ are identically distributed conditional on $\mathcal{H}_{t-1}$.
Construction of Counter-examples

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

- **Fact 2:** $\tilde{\Theta}_t$ and $\Theta^*$ are identically distributed unconditionally.
Construction of Counter-examples

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

- **Fact 2:** $\tilde{\Theta}_t$ and $\Theta^*$ are identically distributed unconditionally.

- This can break under distributional mismatch.
Construction of Counter-examples

- We let $\Theta^* \sim \mathcal{N}(0, I_d)$.

- Under mismatch, there is a bandit problem with:
  
  - $\mathbb{E}[\tilde{\Theta}_t] = c \mathbf{1}$ for some $c = \mathcal{O}(1) > 0$,
  
  - $\langle \tilde{\Theta}_t - \mathbb{E}[\tilde{\Theta}_t], \frac{1}{\sqrt{d}} \rangle$ is $\mathcal{O}(1)$-sub-Gaussian.

- Therefore, we have $\langle \tilde{\Theta}_t, \frac{1}{\sqrt{d}} \rangle = c \sqrt{d} + \mathcal{O}(1)$ w.h.p.
Construction of Counter-examples

- Now, let $A_t := \{0, A\}$ where $A := -1/\sqrt{d}$.

- $A$ is the optimal arm with probability $\frac{1}{2}$.
Construction of Counter-examples

- Now, let $A_t := \{0, A\}$ where $A := -\frac{1}{\sqrt{d}}$.

- $A$ is the optimal arm with probability $\frac{1}{2}$.

- However, LinTS will choose $A$ only if

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

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

- Now, let $A_t := \{0, A\}$ where $A := -1/\sqrt{d}$.

- $A$ is the optimal arm with probability $\frac{1}{2}$.

- However, LinTS will choose $A$ only if
  \[ \langle \tilde{\Theta}_t, A \rangle = -c\sqrt{d} + O(1)\text{-sub-Gaussian} > 0. \]

- 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 $\tilde{M}_t$ that maps each $A \in \mathcal{A}_t$ to $\mathbb{R}$ such that

$$|\tilde{M}_t(A) - \langle A, \hat{\Theta}_{t-1} \rangle| \leq \rho \|A\|_{V_{t-1}^{-1}}$$

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

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

$$|\tilde{M}_t(A) - \langle A, \hat{\Theta}_{t-1} \rangle| \leq \rho \|A\|_{V_{t-1}^{-1}}$$

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 $\tilde{A}_t := \arg \max_{A \in \mathcal{A}_t} \tilde{M}_t(A)$
3:   Observe the reward $r_t$
4:   Compute $V_t = \lambda \mathbb{I} + \sum_{i=1}^{t} \tilde{A}_i \tilde{A}_i^\top$
5:   Compute $\hat{\Theta}_t = V_t^{-1} \left( \sum_{i=1}^{t} \tilde{A}_i r_i \right)$
6: end for
```
Examples of worth functions:

- **Greedy**: \( \tilde{M}_t(A) = \langle A, \hat{\Theta}_{t-1} \rangle \)
- **OFUL**: \( \tilde{M}_t(A) = \langle A, \hat{\Theta}_{t-1} \rangle + \rho \|A\|_{V_{t-1}}^{-1} \)
- **LinTS**: \( \tilde{M}_t(A) = \langle A, \tilde{\Theta}_{t-1} \rangle \)
A General Regret Bound

Definition (Optimism – Informal)

We say a worth function \( \tilde{M}_t \) is optimistic if

\[
\sup_{A \in A_t} \tilde{M}_t(A) \geq \sup_{A \in A_t} \langle A, \Theta^* \rangle
\]

with probability at least \( p \).
Definition (Optimism – Informal)

We say a worth function $\tilde{M}_t$ is optimistic if

$$\sup_{A \in \mathcal{A}_t} \tilde{M}_t(A) \geq \sup_{A \in \mathcal{A}_t} \langle A, \Theta^* \rangle$$

(1)

with probability at least $p$.

Theorem

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

$$\text{BayesRegret}(T, \mathcal{P}, \pi^{\text{ROFUL}}) \leq \tilde{O}\left(\rho \sqrt{dT/p}\right).$$
Improving LinTS

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

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

**Algorithm 7** Improved LinTS algorithm

1: for $t = 1$ to $T$ do
2:   if $\psi(V_t^{-1}) \leq \psi$ then
3:     Sample $\tilde{\Theta}_t \sim \mathcal{N}(\hat{\Theta}_{t-1}, \beta^2 V_t^{-1})$
4:   else
5:     Sample $\tilde{\Theta}_t \sim \mathcal{N}(\hat{\Theta}_{t-1}, \rho^2 V_t^{-1})$
6:   end if
7:   Pull $A_t := \arg \max_{A \in A_t} \langle A, \tilde{\Theta}_t \rangle$
8:   Observe the reward $r_t$
9:   Compute $V_t$ and $\hat{\Theta}_t$ as before.
10: end for
Main Result

The inflation parameter $\beta$ 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 $\sum_{t=1}^{T} \mathbb{P}(\psi(V_t^{-1}) > \Psi) \leq C$, we have

$$\text{Regret}(T, \Theta^*, \pi^{TS}) \leq O\left(\rho \beta \sqrt{dT \log(T)} + C\right).$$
Empirical Scrutiny on Thinness

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

- $\Theta^* \sim \mathcal{N}(0, I_{100})$ and $\varepsilon_t \sim \mathcal{N}(0, 1)$,
- $A_t$ consists of $k = 50$ random vectors in $\text{Unif}\left(-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\right)^d$. 
Empirical Scrutiny on Thinness

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

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

![Graphs showing Thiness and Instantaneous Regret over time for different algorithms.](image-url)
Intuitions Behind Improved LinTS
A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

\[ \tilde{M}_t(A) = \langle A, \tilde{\Theta}_t \rangle. \]
A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by
  \[ \tilde{M}_t(A) = \langle A, \tilde{\Theta}_t \rangle. \]

- We can decompose it as
  \[ \tilde{M}_t(A) = \langle A, \tilde{\Theta}_t - \hat{\Theta}_{t-1} \rangle + \langle A, \hat{\Theta}_{t-1} - \Theta^* \rangle + \langle A, \Theta^* \rangle. \]
A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

\[ \tilde{M}_t(A) = \langle A, \tilde{\Theta}_t \rangle. \]

- We can decompose it as

\[ \tilde{M}_t(A) = \langle A, \tilde{\Theta}_t - \hat{\Theta}_{t-1} \rangle + \langle A, \hat{\Theta}_{t-1} - \Theta^* \rangle + \langle A, \Theta^* \rangle. \]

- Hence, letting \( A^*_t := \arg \max_{A \in A_t} \langle A, \Theta^* \rangle \), we have

\[
\sup_{A \in A_t} \tilde{M}_t(A) - \sup_{A \in A_t} \langle A, \Theta^* \rangle \geq \tilde{M}_t(A^*_t) - \langle A^*_t, \Theta^* \rangle = \langle A^*_t, \tilde{\Theta}_t - \hat{\Theta}_{t-1} \rangle + \langle A^*_t, \hat{\Theta}_{t-1} - \Theta^* \rangle.
\]
A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by
  \[ \tilde{M}_t(A) = \langle A, \tilde{\Theta}_t \rangle. \]

- We can decompose it as
  \[ \tilde{M}_t(A) = \langle A, \tilde{\Theta}_t - \hat{\Theta}_{t-1} \rangle + \langle A, \hat{\Theta}_{t-1} - \Theta^* \rangle + \langle A, \Theta^* \rangle. \]

- Hence, letting \( A^*_t := \arg \max_{A \in \mathcal{A}_t} \langle A, \Theta^* \rangle \), we have
  \[
  \sup_{A \in \mathcal{A}_t} \tilde{M}_t(A) - \sup_{A \in \mathcal{A}_t} \langle A, \Theta^* \rangle \geq \tilde{M}_t(A^*_t) - \langle A^*_t, \Theta^* \rangle = \langle A^*_t, \tilde{\Theta}_t - \hat{\Theta}_{t-1} \rangle + \langle A^*_t, \hat{\Theta}_{t-1} - \Theta^* \rangle. \]

\[ \text{Error term} \]
A Sufficient Condition for Optimism

- Recall that the worth function for LinTS is given by

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

- We can decompose it as

$$\tilde{M}_t(A) = \langle A, \tilde{\Theta}_t - \hat{\Theta}_{t-1} \rangle + \langle A, \hat{\Theta}_{t-1} - \Theta^* \rangle + \langle A, \Theta^* \rangle.$$ 

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

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

$$= \langle A_t^*, \tilde{\Theta}_t - \hat{\Theta}_{t-1} \rangle + \langle A_t^*, \hat{\Theta}_{t-1} - \Theta^* \rangle.$$  

**Compensation term**  

**Error term**
A Sufficient Condition for Optimism

Define

- Error vector $E := \Theta^* - \hat{\Theta}_{t-1}$
- Compensator vector $C := \tilde{\Theta}_t - \tilde{\Theta}_{t-1}$

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

$$\langle A_t^*, C \rangle \geq \langle A_t^*, E \rangle.$$
Reducing the Inflation Parameter

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

- We have $C \sim \mathcal{N}(0, \beta^2 V_{t-1}^{-1})$ which implies that $\|C\|_{V_{t-1}} \approx \beta \sqrt{d}$.

- On the other hand, recall that $\|E\|_{V_{t-1}} \approx \sqrt{d}$.

- Next note that with high probability

$$\langle A^*_t, C \rangle \propto \beta \|A^*_t\|_{V_{t-1}^{-1}}.$$  

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

$$\langle A^*_t, E \rangle \propto \sqrt{d} \|A^*_t\|_{V_{t-1}^{-1}}.$$
Reducing the Inflation Parameter

- We have $C \sim \mathcal{N}(0, \beta^2 V_{t-1}^{-1})$ which implies that $\|C\|_{V_{t-1}} \approx \beta \sqrt{d}$.

- On the other hand, recall that $\|E\|_{V_{t-1}} \approx \sqrt{d}$.

- Next note that with high probability
  \[
  \langle A^*_t, C \rangle \propto \beta \|A^*_t\|_{V_{t-1}^{-1}}.
  \]

- Finally, in the worst case, we may get (by Cauchy-Schwartz)
  \[
  \langle A^*_t, E \rangle \propto \sqrt{d} \|A^*_t\|_{V_{t-1}^{-1}}.
  \]

- What if we assume that $A^*_t$ 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

\[
\mathbb{P} \left( \langle A^*_t, V \rangle > \frac{\nu}{\sqrt{d}} \|A^*_t\|_2 \right) \leq \frac{1}{t^3},
\]

for some fixed \( \nu \in [1, \sqrt{d}] \).
Diversity is not Sufficient
Improved Worst-Case Regret Bound for LinTS

Define **thinness** of a matrix $\Sigma$ to be

$$\psi(\Sigma) := \sqrt{d \cdot \|\Sigma\|_{\text{op}}} \cdot \frac{1}{\|\Sigma\|_*}.$$
Improved Worst-Case Regret Bound for LinTS

Define **thinness** of a matrix $\Sigma$ to be

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

**Assumption**

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

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

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

Nima Hamidi
Regret Bounds for Linear Bandit Algs.
Stanford University
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
Thank you!

Any questions?
## Failure of LinTS: Example 1

<table>
<thead>
<tr>
<th>Environment</th>
<th>LinTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>$\mathcal{N}(0, \mathbb{I}_d)$</td>
</tr>
<tr>
<td>Noise</td>
<td>$\mathcal{N}(0, 0)$</td>
</tr>
</tbody>
</table>

### Figure

![Graph showing the expected number of failures for LinTS](image-url)
Failure of LinTS: Example 2

<table>
<thead>
<tr>
<th>Environment</th>
<th>LinTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>( \mathcal{N}(0.1 \cdot 1_d, I_d) )</td>
</tr>
<tr>
<td>Noise</td>
<td>( \mathcal{N}(0, 1) )</td>
</tr>
</tbody>
</table>

Expected number of failures for LinTS

\[
\begin{align*}
2^{10} & \quad 10^4 \\
2^{11} & \quad 10^{12} \\
2^{12} & \quad 10^{20} \\
2^{13} & \quad 10^{28} \\
2^{14} & \quad 10^{36} \\
2^{15} & \quad 10^{52} \\
2^{16} & \quad 10^{60} \\
\end{align*}
\]
## Failure of LinTS: Example 2

<table>
<thead>
<tr>
<th>Environment</th>
<th>LinTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>( \mathcal{N}(\mu \cdot 1_{2000}, I_{2000}) )</td>
</tr>
<tr>
<td>Noise</td>
<td>( \mathcal{N}(0, 1) )</td>
</tr>
</tbody>
</table>

The expected number of failures for LinTS is visualized in the graph below:

![Graph showing expected number of failures for LinTS](image-url)


References III


