On Worst-case Regret of Linear Thompson Sampling

Nima Hamidi and Mohsen Bayati

Stanford University

October 12, 2020

Preprint: arXiv 2006.06790

Overview

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.

Main Goal

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

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

Main Goal

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

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

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

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

Algorithms

Related Literature

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

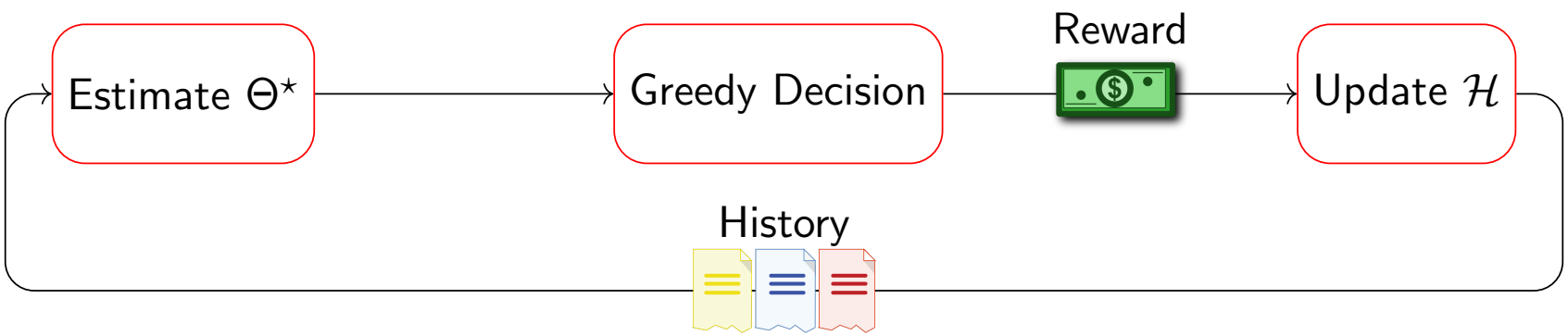
Greedy

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

- Using the set of observations

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

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



Greedy

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

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

and

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

Greedy

Algorithm 1 Greedy algorithm

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

Greedy

Algorithm 2 Greedy algorithm

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

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

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

Optimism in Face of Uncertainty (OFU) Algorithm

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

Optimism in Face of Uncertainty (OFU) Algorithm

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

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

where

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

Optimism in Face of Uncertainty (OFU) Algorithm

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

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

where

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

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

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

Optimism in Face of Uncertainty (OFU) Algorithm

Algorithm 3 OFUL algorithm

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

Optimism in Face of Uncertainty (OFU) Algorithm

Algorithm 3 OFUL algorithm

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

It can be shown that

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

Linear Thompson Sampling (LinTS) Algorithm

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

Linear Thompson Sampling (LinTS) Algorithm

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

Algorithm 4 LinTS algorithm

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

Linear Thompson Sampling (LinTS) Algorithm

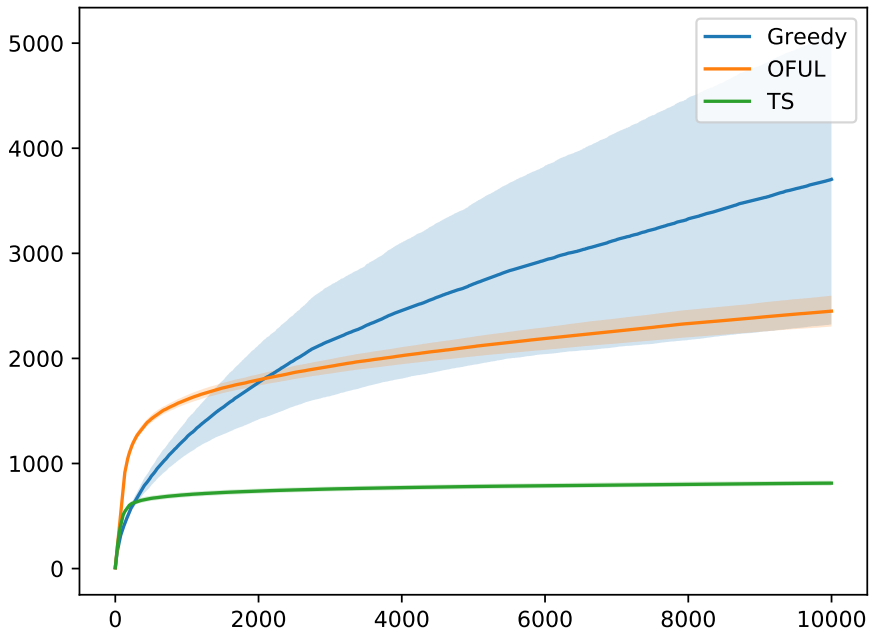
- Under **normality**, LinTS becomes:

Algorithm 5 LinTS algorithm under normality

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

Why Is LinTS Popular?

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



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

Theoretical Results

Comparison of Regret Bounds

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

Under some conditions, the regret of OFUL is bounded by

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

Comparison of Regret Bounds

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

Under some conditions, the regret of OFUL is bounded by

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

Theorem (Russo and Van Roy 2014)

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

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

Comparison of Regret Bounds

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

Under some conditions, the regret of OFUL is bounded by

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

Theorem (Russo and Van Roy 2014)

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

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

Theorem (Dani, Hayes, and Kakade 2008)

There is a Bayesian linear bandit problem that satisfies

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

A Worst-Case Regret Bound for LinTS

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

Algorithm 6 LinTS algorithm under normality

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

A Worst-Case Regret Bound for LinTS

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

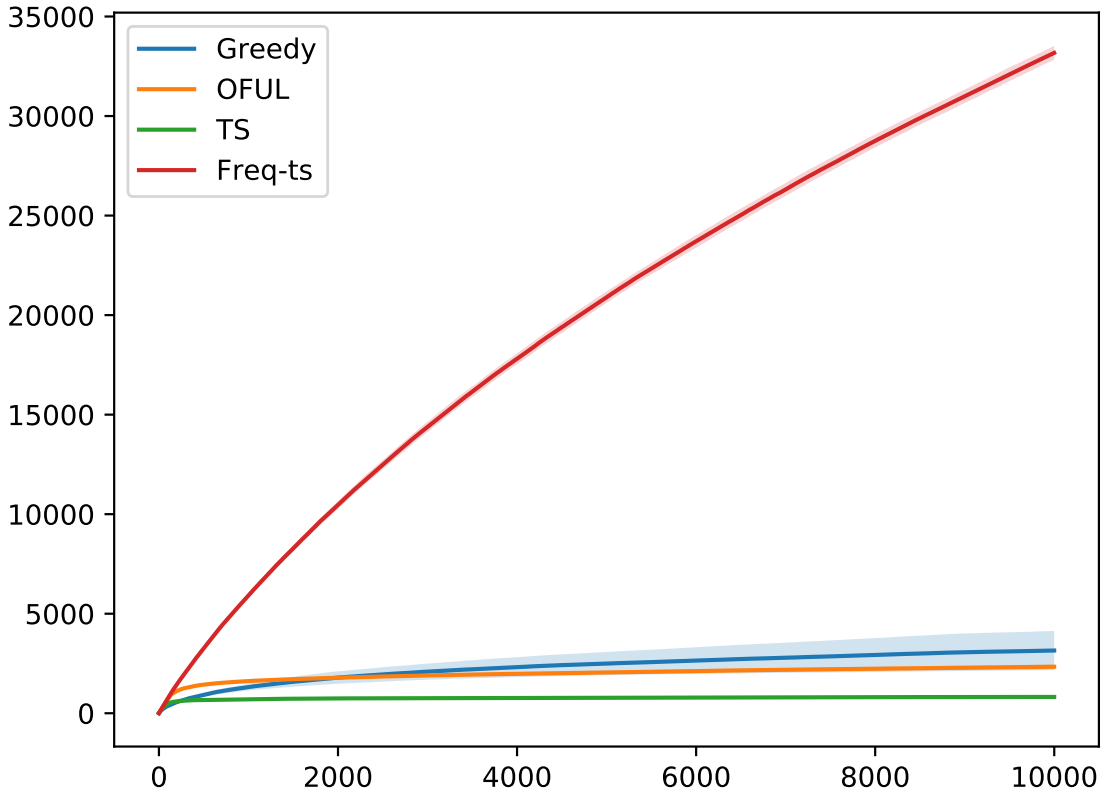
If $\beta \propto \sqrt{d}$, *then*

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

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

Empirical Performance of Inflated LinTS

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



Bayesian Analyses are Brittle

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

Theorem

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

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

Bayesian Analyses are Brittle

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

Theorem

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

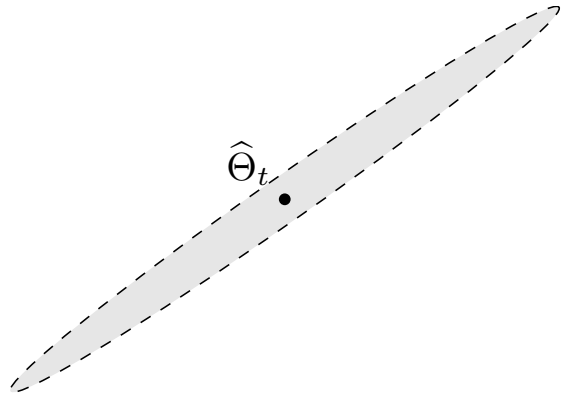
$$\text{BayesRegret}(T, \pi^{LinTS}) = \Omega(T).$$

The counter-example satisfies the following properties:

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

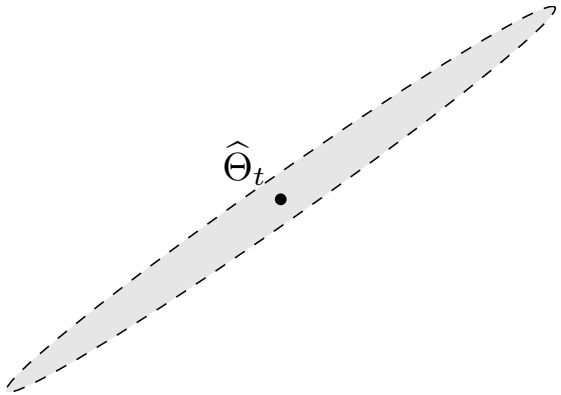
Sufficient Conditions for Reducing Inflation

- If the optimal action is sufficiently **diverse**,
- When the confidence set is not **thin**,
- The inflation can be reduced to $\widetilde{\mathcal{O}}(1)$.



Sufficient Conditions for Reducing Inflation

- If the optimal action is sufficiently **diverse**,
- When the confidence set is not **thin**,
- The inflation can be reduced to $\widetilde{\mathcal{O}}(1)$.



Theorem

If $\sum_{t=1}^{T} \mathbb{P}\big(\mathbf{V}_t^{-1}$ *is thin*$\big) \leq C$, *we have*

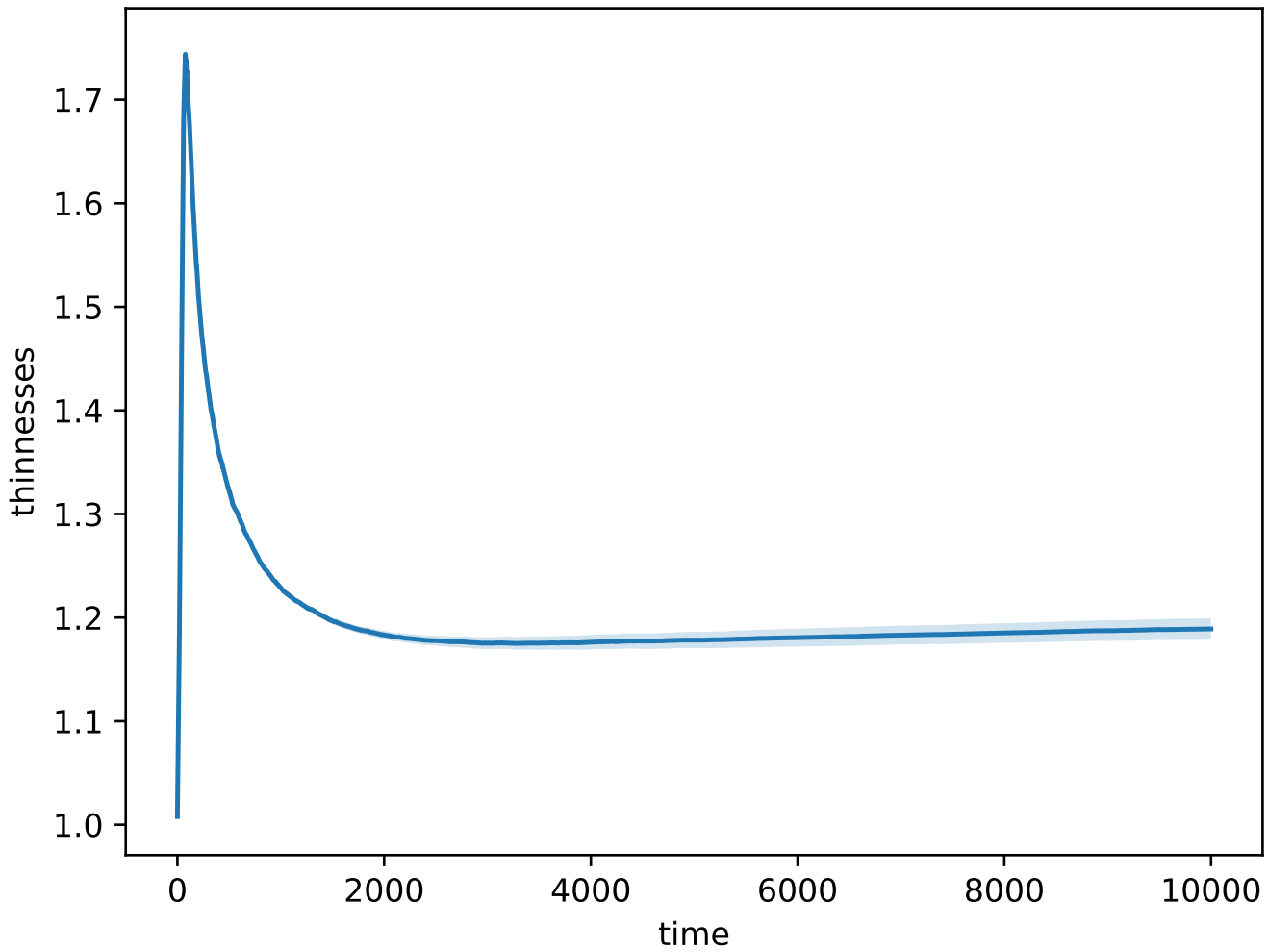
$$\text{Regret}(T, \Theta^\star, \pi^{TS}) \leq \mathcal{O}\Big(\rho\beta\sqrt{dT\log(T)} + C\Big).$$

Sufficient Conditions for Reducing Inflation

Thinness in our simulations:

Sufficient Conditions for Reducing Inflation

Thinness in our simulations:



Conclusion

- We proved that LinTS without inflation can incur linear regret.
- This happens even if one reduces the noise variance.
- We proved that diversity and thickness assumptions can help to reduce the inflation.

Thank you!

Any questions?