Thompson Sampling for Linear Bandits

Lokesh Kashyap (22924)

December 3/12/2024

1 Abstract

This project explores Thompson Sampling for Linear Bandits, a probabilistic algorithm addressing the contextual bandit problem where rewards depend linearly on context vectors. Balancing exploration and exploitation, the study compares Thompson Sampling with LinUCB, a deterministic confidence-bound-based method. Results reveal that LinUCB converges faster in low-dimensional settings, while Thompson Sampling excels in high-dimensional and dynamic environments due to its robust handling of uncertainty. Both methods achieve sublinear regret, validating their efficiency. The project proposes enhancements like hybrid models and adaptations for delayed feedback and non-stationary contexts, underscoring their real-world applicability in recommendation systems, online advertising, and personalized healthcare.

2 Introduction

The contextual bandit problem serves as a foundational framework for sequential decision-making in dynamic environments. Unlike traditional multi-armed bandits, contextual bandits leverage contextual information, enabling more informed decision-making. This adaptability is vital for applications such as recommendation systems, online advertising, and personalized healthcare, where balancing exploration (gathering knowledge) with exploitation (using knowledge) is key to maximizing rewards.

Thompson Sampling, a Bayesian algorithm utilizing posterior sampling, has gained prominence for its ability to navigate stochastic environments. Its probabilistic framework allows efficient handling of uncertainty, especially in high-dimensional settings where deterministic approaches may falter. This project investigates Thompson Sampling's effectiveness in linear bandit scenarios, where expected rewards are modeled as linear functions of context vectors.

The study further contrasts Thompson Sampling with LinUCB, a deterministic confidence-bound-based method, analyzing their performance across varying context dimensions. Experimental results highlight the strengths and limitations of both algorithms, providing insights into their adaptability. Building on these findings, this project explores potential extensions to enhance their real-world relevance, including hybrid models and adjustments for delayed feedback and evolving contexts.

3 Methodology

3.1 Problem Setup

In the contextual bandit framework, at each time step t, the algorithm observes a context vector $b_i(t) \in \mathbb{R}^d$ for each arm $i \in \{1, \ldots, N\}$ and selects an arm $a(t)$. The reward for arm i is given by:

$$r_i(t) = b_i(t)^T \mu + \eta, \quad \eta \sim \mathcal{N}(0, R^2),$$

where:

- $b_i(t)$: Context vector for arm i, representing features.
- $\mu \in \mathbb{R}^d$: Unknown parameter vector, constant across time.
- η: Gaussian noise, modeling stochasticity in the reward.

The algorithm aims to minimize the **cumulative regret** over T rounds, defined as:

$$R(T) = \sum_{t=1}^{T} \left(\max_i \mathbb{E}[r_i(t)] - \mathbb{E}[r_{a(t)}(t)] \right),$$

where $\mathbb{E}[r_i(t)]$ is the expected reward of arm i, and $\mathbb{E}[r_{a(t)}(t)]$ is the expected reward of the selected arm $a(t)$.

3.2 Thompson Sampling Algorithm

1. **Prior Initialization**: Initialize the prior for μ as:
$$\mu \sim \mathcal{N}(0, \lambda I),$$
where $\lambda > 0$ is a regularization parameter, and I is the identity matrix.

2. **Posterior Sampling**: At each time step t, sample a parameter vector $\tilde{\mu}(t)$ from the posterior:
$$\tilde{\mu}(t) \sim \mathcal{N}(\hat{\mu}, B^{-1}),$$
where:
   - $\hat{\mu} = B^{-1} b_{\text{sum}}$ is the posterior mean.
   - B is the precision matrix, initialized as $B = \lambda I$.

3. **Arm Selection**: Select the arm $a(t)$ that maximizes the predicted reward:
$$a(t) = \arg\max_i \left(b_i(t)^T \tilde{\mu}(t) \right).$$

4. **Posterior Update**: After observing the reward $r_{a(t)}(t)$ and the context $b_{a(t)}(t)$, update:
$$B \leftarrow B + b_{a(t)}(t) b_{a(t)}(t)^T, \quad b_{\text{sum}} \leftarrow b_{\text{sum}} + b_{a(t)}(t) r_{a(t)}(t).$$

4 Results and Discussion

4.1 Theoretical Regret Bound

In addition to the empirical evaluation, the theoretical regret bound for the Thompson Sampling algorithm demonstrates its efficiency in balancing exploration and exploitation. The regret bound is expressed as:

$$\mathcal{R}(T) = O\left(d^{3/2} \sqrt{T} \left(\ln(T) + \sqrt{\ln(T) \ln\left(\frac{1}{\delta}\right)} \right) \right),$$

or equivalently:

$$\mathcal{R}(T) = O\left(d\sqrt{T \log(N)} \left(\ln(T) + \sqrt{\ln(T) \ln\left(\frac{1}{\delta}\right)} \right) \right),$$

where:

- d: Dimension of the context vector.
- T: Time horizon.
- N: Number of arms.
- δ: Confidence parameter, representing the probability of failure.

4.1.1 Relevance to Empirical Results

The theoretical regret bounds are consistent with the sublinear cumulative regret observed in the experiments (Figure 3 and Figure 4). While the empirical results show that LinUCB achieves faster initial convergence, Thompson Sampling's probabilistic approach is more effective in handling stochastic rewards and high uncertainty, which aligns with the theoretical guarantees.

4.2 Cumulative Regret Plot

The cumulative regret for both Thompson Sampling and LinUCB is shown in Figure 3 and Figure 4. The plot demonstrates that both algorithms achieve sublinear cumulative regret over time, indicating that they effectively balance exploration and exploitation. Thompson Sampling exhibits slower initial convergence due to its probabilistic exploration, but it adapts well to stochastic environments, especially in high-noise settings. LinUCB, on the other hand, converges faster in low-noise settings due to its deterministic confidence-bound approach. Overall, Thompson Sampling performs better in handling uncertainty, while LinUCB is more efficient in stable, low-noise environments.

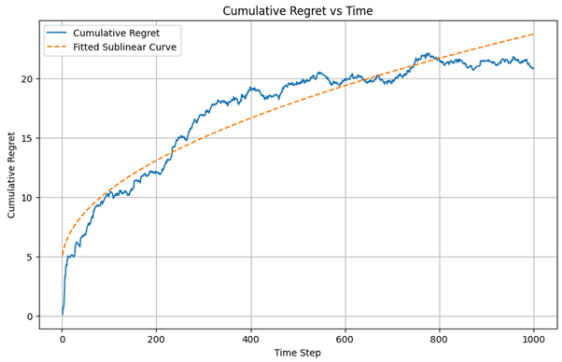

Figure 1: Cumulative Regret for Thompson Sampling.

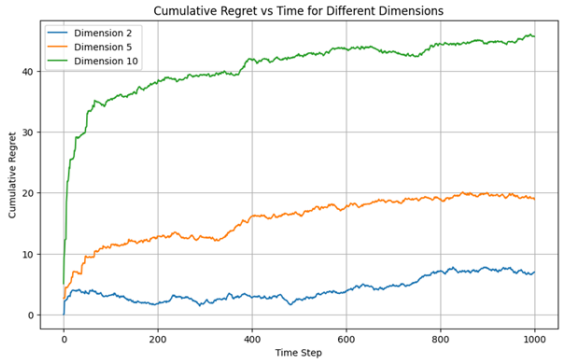

Figure 2: Cumulative Regret for Thompson Sampling for different dimension.

Dimension	True Parameter (μ)	Learned Parameter ($\hat{\mu}$)
1	1.22	1.20
2	0.30	0.32
3	-0.23	-0.24

Table 1: Comparison of True and Learned Parameters.

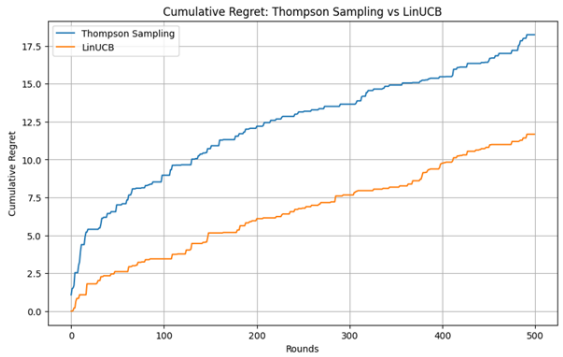

Figure 3: Cumulative Regret for Thompson Sampling VS LinUCB for Context$_dim = 3$.

5 Challenges and Future Work

5.1 Challenges

Implementing Thompson Sampling for linear bandits presented several challenges:

- **Computational Complexity:** The matrix operations involved in posterior updates (e.g., calculating μ_{hat} and B^{-1}) become computationally expensive as the context dimension d and time horizon T increase.
- **Dynamic Contexts:** Simulating and adapting to dynamic contexts required careful handling to ensure the algorithm could learn effectively without overfitting to specific patterns in the context vectors.
- **Noisy Rewards:** Handling stochasticity in rewards ($\eta \sim \mathcal{N}(0, R^2)$) made parameter estimation and arm selection more challenging, especially in high-noise scenarios.
- **Regret Analysis:** The theoretical analysis of cumulative regret required a deep understanding of concentration inequalities and martingale properties, which were non-trivial to derive and interpret.

5.2 Future Work

There are several avenues to extend the work in this project:

- **Hybrid Models:** Integrating shared and arm-specific features, as seen in hybrid LinUCB models, could enhance Thompson Sampling's performance by leveraging global patterns while preserving arm-specific nuances.
- **Delayed Feedback:** Adapting Thompson Sam-

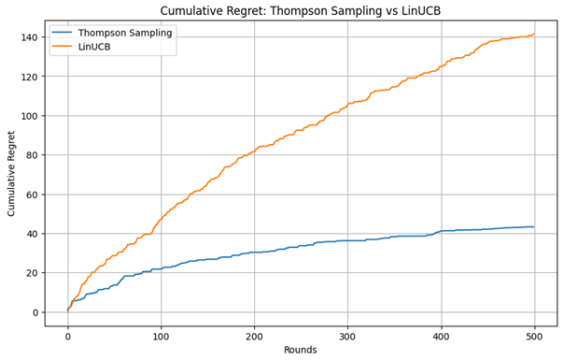

Figure 4: Cumulative Regret for Thompson Sampling VS LinUCB for Context$_dim = 10$.

pling to handle scenarios where rewards are observed with a delay, such as in online advertising or recommendation systems, would improve its applicability.

- **Non-Stationary Contexts:** Developing mechanisms to handle non-stationary environments, such as sliding window updates or time-weighted learning, could make the algorithm robust to changing conditions.
- **Non-Linear Reward Structures:** Extending the algorithm to support non-linear relationships between contexts and rewards, using kernel methods or neural networks, would broaden its scope to more complex real-world problems.
- **Scalability:** Optimizing matrix operations and leveraging parallel processing or approximate inference techniques could make the algorithm scalable to high-dimensional or large-scale problems.

6 Conclusion

This project explored Thompson Sampling for linear bandits, focusing on its implementation, theoretical analysis, and comparison with LinUCB. Thompson Sampling's Bayesian approach effectively balances exploration and exploitation, achieving sublinear regret and demonstrating robustness in dynamic and noisy environments. The algorithm successfully estimated the true parameter vector, showcasing its capability to adapt to varying contexts and uncertainties.

The comparison with LinUCB highlighted the strengths and weaknesses of each approach. While LinUCB's deterministic confidence bounds provided faster convergence in low-noise settings, Thompson Sampling excelled in handling stochastic rewards and dynamic contexts. This study also emphasized the theoretical guarantees of Thompson Sampling, including its regret bound, which aligns closely with the state-of-the-art methods.

Future extensions, such as hybrid models, handling delayed feedback, and adapting to non-linear rewards, offer exciting opportunities to broaden the algorithm's applicability to real-world problems. Overall, Thompson Sampling remains a powerful and flexible algorithm for contextual bandit problems, with significant potential for further improvements and adaptations.

7 References

References

[1] S. Agrawal and N. Goyal, *Analysis of Thompson Sampling for the Multi-Armed Bandit Problem*, In Proceedings of the 25th Annual Conference on Learning Theory (COLT), 2012.

[2] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári, *Parametric Bandits: The Generalized Linear Case*, In Advances in Neural Information Processing Systems (NIPS), 2010.

[3] L. Li, W. Chu, J. Langford, and R. Schapire, *A Contextual-Bandit Approach to Personalized News Article Recommendation*, In Proceedings of the 19th International World Wide Web Conference (WWW), 2010.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, *Finite-time Analysis of the Multiarmed Bandit Problem*, Machine Learning, 47(2–3), pp. 235–256, 2002.