Randomized Multiarm Bandits: An Improved Adaptive Data Collection Method

Zhigen Zhao[1], Tong Wang[1], and Bo Ji[2]

[1]Department of Statistics, Operations, and Data Science, Temple University.
Email: zhaozhg@temple.edu
[2]Department of Computer Science, Virginia Tech

Abstract

In many scientific experiments, multi-armed bandits are used as an adaptive data collection method. However, this adaptive process can lead to a dependence that renders many commonly used statistical inference methods invalid. An example of this is the sample mean, which is a natural estimator of the mean parameter but can be biased. This can cause test statistics based on this estimator to have an inflated type I error rate, and the resulting confidence intervals may have significantly lower coverage probabilities than their nominal values. To address this issue, we propose an alternative approach called randomized multiarm bandits (rMAB). This combines a randomization step with a chosen MAB algorithm, and by selecting the randomization probability appropriately, optimal regret can be achieved asymptotically. Numerical evidence shows that the bias of the sample mean based on the rMAB is much smaller than that of other methods. The test statistic and confidence interval produced by this method also perform much better than its competitors.

Keywords: Multiarmed bandits. Biased estimator. Optimal regret. Type I error. Statistical inference.

1 Introduction

Adaptive data collection has been widely used to optimize experimental resources and enhance user experience. It has been applied in various domains, including clinical trials, recommendation systems, and online advertising (Zeng et al. [2016], Lamprier et al. [2019], Li et al. [2010], Suhr et al. [2015], Jonker et al. [2016]). For instance, in clinical trials, adaptive designs allow researchers to address uncertainties in the planning phase and modify the trial's characteristics based on the accumulating information (Suhr et al. [2015], Jonker et al. [2016]). This approach is especially beneficial for rare diseases, such as orphan diseases, where the pool of subjects is limited. Examples of successful applications of the adaptive designs can be seen in recent trials such as BATTLE for lung cancer (Kim et al. [2011]) and I-SPY2 for breast cancer (Barker et al. [2009]). Additionally, adaptive data collection techniques are employed in scenarios such

as finding the strategy that maximizes the reward of a gambler in the presence of multiple slot machines (Lai and Robbins [1985], Katehakis and Arthur F. Veinott [1987], Sutton and Barto [2018]). In the tech industry, these data-driven decisions for continuous improvement are referred to as A/B testing (Bubeck and Cesa-Bianchi [2012]). Recent research has uncovered a major issue concerning the relationship between decision-making processes and past information in adaptive data collection. This connection leads to a bias when attempting to calculate mean parameters using sample means (Xu et al. [2013], Nie et al. [2018], Neel and Roth [2018], Shin et al. [2019]). As discussed in Section 2, numerical evidence shows that this bias causes a variety of problems, such as misidentifying treatment effects, increasing false positive errors in hypothesis testing, and reducing the accuracy of confidence intervals. Therefore, it is essential to address the pressing need to reduce the size of the bias while still preserving the inherent features of adaptive data collection techniques.

The emergence of bias can be attributed to the fact that the choice of an arm at a given time is based on historical data. To address this issue, several approaches have been proposed in recent years to reduce such dependence. For example, in Nie et al. [2018], a method called cMLE was introduced, which uses an MCMC-based approach. At each time step, a Gumbel noise is added to the statistic to determine the arm selection for that particular time. In another study Neel and Roth [2018], a method based on Differential Privacy (DP) was developed, providing a bound on bias in relation to the DP parameter. Both of these approaches introduce noise into the adaptive sequence to reduce bias; however, this noise addition may affect the regret to some degree. Therefore, there is a strong interest in developing bias-mitigating approaches that minimize the loss of regret, finding a balance between the two objectives.

This paper presents a novel approach, referred to as rMAB, which seeks to reduce the reliance on historical data in multi-armed bandit (MAB) algorithms. The rMAB method combines a randomization step (independent of historical data) with an MAB algorithm (dependent on historical data). This approach has several advantages: (i) the randomization component weakens the correlation between data points, leading to a significant decrease in the bias of the sample mean and improved inferential properties; (ii) by selecting the randomization probability appropriately, the regret of the rMAB method remains optimal asymptotically, guaranteeing the preservation of performance guarantees.

This paper is structured as follows. Section 2 provides an overview of the adaptive data collection framework and examines the effects of bias on statistical inferences. Section 3 introduces the randomization-based approach, rMAB, which effectively reduces bias, and provides upper bounds on the regret for these algorithms. Section 4 presents a numerical comparison between rMAB and existing methods. Section 5 offers concluding remarks and a discussion of the findings. The technical proofs and additional numerical results are included in the Appendix.

2 Challenges in Statistical Inference for Adaptive Data Collection

Assuming a multi-armed bandit (MAB) algorithm with K arms, each corresponding to a distribution P_k for $k = 1, 2, \ldots, K$, we denote $I_t \in \{1, 2, \ldots, K\}$ as the index of the arm chosen at time t. A random real-valued reward $X_{I_t}(t)$ is generated from the distribution P_{I_t}. The mean

and variance of the rewards obtained when drawing from arm k are denoted by μ_k and σ_k^2, respectively. It is important to note that rewards obtained by repeatedly playing a specific arm are independent and identically distributed, and they are independent of rewards from other arms. The regret $R(T)$ for a specific algorithm over a total time horizon T is the difference between the total reward that could have been obtained by always choosing the best arm ($\max_k \sum_{t=1}^{T} X_k(t)$) and the actual cumulative reward achieved by the chosen arms ($\sum_{t=1}^{T} X_{I_t}(t)$) during the time horizon T. It is defined as:

$$R(T) = \max_k \sum_{t=1}^{T} X_k(t) - \sum_{t=1}^{T} X_{I_t}(t).$$

Without any loss of generality, assume that the first arm is the optimal one and define $\Delta_k = \mu_1 - \mu_k$ as the difference between the mean reward of the optimal arm and the k-th arm. We denote $N_k(T)$ as the number of times the k-th arm is pulled within the time horizon T, i.e., $N_k(T) = \sum_{t=1}^{T} \mathbf{1}(I_t = k)$, where $\mathbf{1}(\cdot)$ is the indicator function. The expected pseudo-regret can then be expressed as:

$$\mathbb{E}[R(T)] = \mathbb{E}\left[T\mu_1 - \sum_{k=2}^{K} N_k(T)\mu_k\right] = \sum_{k=2}^{K} \mathbb{E}[N_k(T)]\Delta_k,$$

which is the sum of the expected number of times each suboptimal arm k is pulled, multiplied by the corresponding difference Δ_k.

For arm k, the sample mean is usually used to estimate the parameter μ_k. This is expressed as:

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{t=1}^{T} X_{I_t}(t)\mathbf{1}(I_t = k). \tag{1}$$

In the case of independent data, $\hat{\mu}_k$ is an unbiased estimator of μ_k. However, when the data is collected adaptively, this estimator is no longer unbiased (Neel and Roth [2018]). The adaptive nature of the data collection process creates a connection between the selection of arms and the past data, resulting in distorted estimates.

Such bias can have a significant impact on the validity of statistical inference methods. To illustrate this point, consider the following experimental scenario:

$$\text{Gaussian: } K = 2,\ X_t \sim N(\mu_{I_t}, \sigma^2) \text{ where } \sigma = 1,\ \mu_1 = 1,\ \mu_2 = 0.5,\ t = 1, 2, \cdots, T = 500. \tag{2}$$

We test the hypothesis

$$H_0 : \mu_1 - \mu_2 = 0.5, vs \quad H_1 : \mu_1 - \mu_2 > 0.5 \tag{3}$$

at a significance level of $\alpha = 0.05$. We employ the two-sample T-test statistic to calculate the p-value based on the data collected. This process is repeated 1,000 times to evaluate the type I error rates for four popular MAB algorithms: Greedy (Bubeck and Cesa-Bianchi [2012]), ϵ_t-Greedy (Auer [2002]), Thompson Sampling (TS, Thompson [1933]), and lil-UCB (Jamieson et al. [2014]). The obtained type I error rates for these algorithms are: Greedy (0.373), ϵ_t-Greedy

(0.101), Thompson Sampling (0.226), and lil-UCB (0.108). These error rates are higher than the desired nominal level of 0.05.

We also construct a 95% confidence interval for $\mu_1 - \mu_2$ under the same experimental conditions. The coverage probabilities for Greedy, ϵ_t-Greedy, Thompson Sampling, and lil-UCB are 0.533, 0.892, 0.764, and 0.903, respectively. These probabilities are substantially lower than the expected level, suggesting a lack of confidence in the estimated intervals.

Indeed, the examples highlight the failure of traditional statistical inference methods when applied to adaptively collected data using commonly used MAB algorithms. The bias, inflated type I error, and low coverage probability pose challenges in drawing valid statistical inferences.

3 Randomized Multi-Armed Bandits

This section introduces a novel approach to adaptively collect data, with the aim of substantially reducing the bias in the sample means and improving the resulting inference methods, such as hypothesis testing and confidence interval estimation. The bias in the sample means is caused by the dependence between the choice of an arm at a given time and the historical data. To address this, we propose a method called randomized MAB (rMAB), which combines a randomization step with an MAB algorithm. At each time t, a certain probability λ_t is used to enter the randomization step. In this step, an arm is randomly chosen and data is collected from it. The remaining probability of $1 - \lambda_t$ is used for the following MAB algorithm to select the arm based on the historical data. By introducing randomization independent of the past data, the rMAB approach weakens the dependence between the decision-making function and the historical data. This randomization step helps reduce the bias in the sample means and improves the accuracy of statistical inference. The rMAB method provides a practical solution to address the issues of bias and validity in adaptive data collection.

A flowchart of the rMAB approach is shown in Figure 1, which demonstrates how the randomization step is incorporated into the MAB algorithm.

Input: $K \geq 2$; $t = 1, ...T$; randomization probability λ_t.
for $t = 1, ..., T$ **do**
  if $t \leq K$ **then**
    Pull t-th arm;
  else
    Draw a random number $u \sim Unif[0, 1]$;
    if $u \leq \lambda_t$ **then**
      Pull the arm through randomized allocation policy (Uniformly Sampling, Water-Filling)
    else
      Pull the arm through regular MAB algorithms (TS, UCB, and etc).
    end
  end
end

Algorithm 1: The rMAB algorithm.

We investigate two different strategies for randomization: uniform sampling (US) and a

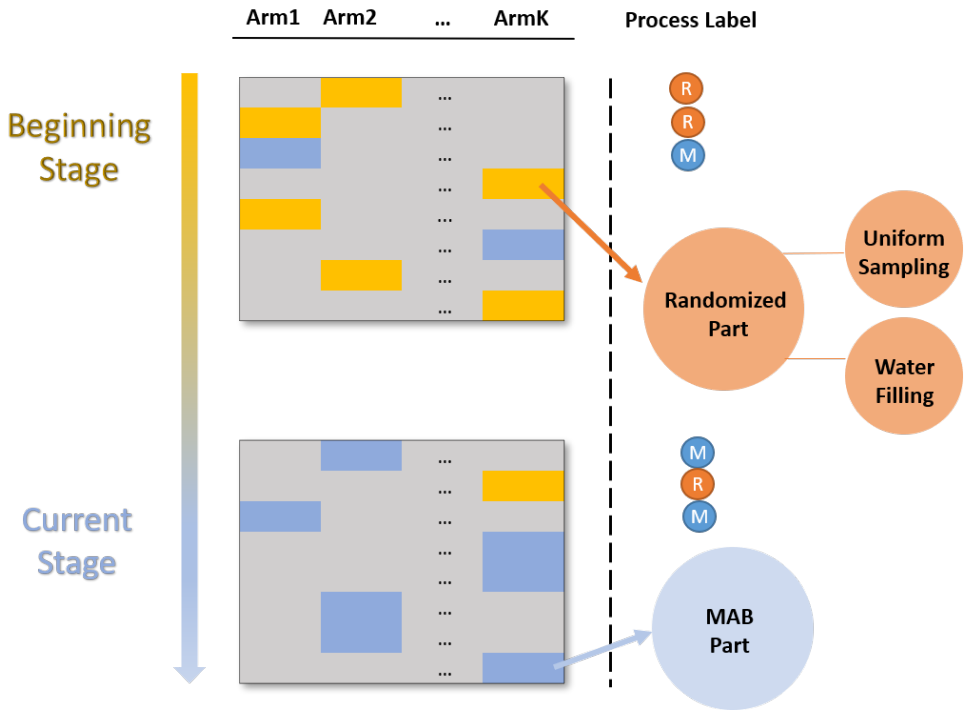

Figure 1: Flowchart of rMAB procedure.

water-filling algorithm (WF) based on the number of pulls. US randomly selects sub-optimal arms with an equal probability, ensuring a fair and unbiased selection process. WF, as proposed in Gai and Krishnamachari [2012], selects arms based on the smallest number of pulls, aiming to allocate more pulls to arms that have been selected less frequently. To denote the combination of the randomization step and the MAB algorithm, we use the notation rX(Y), where X is the MAB algorithm (e.g. UCB) and Y is the randomization strategy (US or WF). For example, rUCB (US) is the randomized algorithm that utilizes the UCB algorithm for the MAB part and uniform sampling for the randomization step. Similarly, rTS (WF) is the randomized algorithm that combines the Thompson Sampling algorithm with the water-filling randomization strategy. By considering different combinations of MAB algorithms and randomization strategies, we can explore a range of rMAB variants that offer flexibility and adaptability to different scenarios and datasets.

The trade-off between bias and regret is a key factor in the rMAB approach. Randomization reduces the bias by weakening the reliance on past data, but it also increases the regret due to decreased exploitation. The effect of the randomization step on regret is determined by the allocation probabilities λ_t. When λ_t is close to zero, there is minimal randomization, and both the regret and bias remain unchanged. As λ_t increases, the magnitude of the bias decreases, but the regret also increases. To achieve a balance, we suggest setting the allocation probability $\lambda_t = \frac{K}{t}$, which gives a higher allocation probability at the beginning of the experiment. This allows for more randomization in the initial phase of the experiment, reducing bias, while still keeping the regret at a reasonable level. It is important to note that in the long run, the total number of arms pulled due to randomization is limited to $(K-1)\log T$, which does not affect the asymptotic order of regret. Based on these considerations, we present the following theorems to analyze the properties of the rMAB approach and guide the selection of appropriate allocation probabilities.

Theorem 3.1 *Suppose there are total K arms and the reward of k-th arm follows the distribution Bernoulli(μ_k). Without loss of generality, assume that the first arm is optimal. Then the upper bound of expected pseudo-regret of the rTS is*

$$\mathbb{E}[R(T)] \leq \sum_{k=2}^{K} \left[\left(\frac{(1+\epsilon)^2}{d(\mu_k, \mu_1)} + 1 \right) \cdot \log T + O(1 + \frac{1}{\epsilon^2}) \right] \Delta_k,$$

where $d(\mu_k, \mu_1) = \mu_k \log \frac{\mu_k}{\mu_1} + (1-\mu_k) \log \frac{1-\mu_k}{1-\mu_1}$.

Similarly, if the reward of the k-th arm follows Gaussian distribution with a mean of μ_k and variance of one, the upper bound of the regret of the rTS is

$$\mathbb{E}[R(T)] \leq \sum_{k=2}^{K} \left[\left(\log T \Delta_k + \frac{18 \log(T\Delta_k^2)}{\Delta_k} \right) + \frac{13}{2\Delta_k} + O(1) \right].$$

According to Lai and Robbins [1985], the expected regret for multi-armed bandit problems has a lower bound of $\log T$. Theorem 3.1 demonstrates that the rTS method is rate optimal

in terms of regret, implying that it achieves optimal performance. The subsequent theorem provides an upper bound on the rate of the expected regret for the rUCB method, and shows that it is optimal. These results demonstrate the effectiveness and optimality of the proposed rMAB approaches in terms of bias reduction and regret optimization.

Theorem 3.2 *Suppose that there are total K arms and the reward of the k-th arm follows a distribution P_k with support in [0,1]. Without loss of generality, assume that the first arm is optimal. Then the upper bound of the expected regret of rUCB is*

$$\mathbb{E}[R(T)] \leq \sum_{k=2}^{K} \left[(\frac{8}{\Delta_k} + \Delta_k) \log T + (\frac{\pi^2}{3} + 1)\Delta_k \right].$$

4 Numerical results

In this section, we showcase the results of our numerical experiments, which assess the efficacy of various methods based on metrics such as biases of the sample mean, expected regrets, type I error rates, and coverage probabilities. These experiments provide valuable insights into how different methods perform across various scenarios. We analyze the following settings:

- Gaussian: $X_t \sim N(\mu_{I_t}, \sigma^2), \sigma^2 = 1$, $I_t \in \{1, 2, \cdots, K\}$,
  - $K = 2$, $\mu_1 = 1$, and $\mu_2 = 0.5$;
  - $K = 5$, $\mu_1 = 1$, $\mu_2 = 0.75$, $\mu_3 = 0.5$, $\mu_4 = 0.375$, and $\mu_5 = 0.25$.
- Bernoulli: $X_t \sim Bernoulli(p_{I_t})$,
  - $K = 2$, $p_1 = 0.8$ and $p_2 = 0.2$,
  - $K = 5$, $p_1 = 0.9, p_2 = 0.7, p_3 = 0.5, p_4 = 0.3$, and $p_5 = 0.1$.

4.1 Bias Reduction

This section compares the biases of three methods: the rMAB, the MAB algorithm with the Conditional Maximum Likelihood Estimator (cMLE, Nie et al. [2018]), and the MAB algorithms with $\epsilon-$differential privacy (DP, Neel and Roth [2018]). The MAB algorithms include Greedy, ϵ_t-Greedy, Thompson Sampling (TS), and lil-UCB. We use the code provided by the authors Nie et al. [2018] to assess the performance of cMLE. However, the code does not incorporate the ϵ_t-Greedy algorithm. To ensure a fair comparison, we choose the scale parameter of the Gumbel noise in cMLE and the differential privacy parameter so that the regrets of cMLE, DP, and rMAB are in the same range (Table 1). The DP method's parameter is selected to ensure its regret is comparable to that of the rMAB algorithm.

When $T = 100$, Table 1 shows that rMAB algorithms outperform their competitors significantly. For instance, for the fifth arm when using lilUCB as the MAB algorithm, the biases obtained with original lilUCB, DP, and cMLE are -0.336, -0.237, and -0.272 respectively. On the other hand, the biases using rlilUCB (US) and rlilUCB (WF) are -0.083 and -0.057 respectively.

MAB	True Mean	MAB	DP	rMAB(US)	rMAB(WF)	cMLE
lilUCB	1.0	-0.141	-0.049	-0.019	-0.031	-0.088
	0.75	-0.196	-0.158	-0.056	-0.047	-0.162
	0.5	-0.294	-0.193	-0.064	-0.046	-0.242
	0.375	-0.341	-0.264	-0.076	-0.043	-0.267
	0.25	-0.356	-0.240	-0.083	-0.057	-0.272
	Regret	0.310	0.324	0.335	0.343	0.338
TS	1.0	-0.232	-0.238	-0.138	-0.089	-0.086
	0.75	-0.393	-0.372	-0.184	-0.146	-0.223
	0.5	-0.410	-0.384	-0.173	-0.110	-0.253
	0.375	-0.410	-0.394	-0.129	-0.096	-0.248
	0.25	-0.407	-0.408	-0.148	-0.080	-0.277
	Regret	0.166	0.168	0.196	0.197	0.231
ϵ_t-Greedy	1	-0.336	-0.313	-0.146	-0.107	NA
	0.75	-0.386	-0.38	-0.189	-0.137	NA
	0.5	-0.317	-0.334	-0.143	-0.083	NA
	0.375	-0.284	-0.307	-0.101	-0.081	NA
	0.25	-0.226	-0.261	-0.087	-0.038	NA
	Regret	0.191	0.193	0.196	0.193	NA
Greedy	1.0	-0.459	-0.352	-0.161	-0.111	-0.279
	0.75	-0.452	-0.468	-0.209	-0.163	-0.407
	0.5	-0.396	-0.406	-0.162	-0.104	-0.429
	0.375	-0.349	-0.434	-0.110	-0.079	-0.411
	0.25	-0.331	-0.359	-0.111	-0.045	-0.277
	Regret	0.206	0.189	0.192	0.185	0.201

Table 1: When considering the Gaussian design with five arms, we simulated the biases of the original MAB algorithms, DP, rMAB(US), rMAB(WF), and cMLE by setting T to 100 and conducting 1,000 replications. The parameters for cMLE and DP were appropriately selected such that the regret of these two methods is similar to that of the rMAB algorithms.

We also plot the biases when T varies from 10 to 500 in Figure 2. The cMLE is not included because of its high computational cost when T is large. It is evident that rMAB algorithms (red dotdash line and green dotted line) outperform the DP method (blue dashed line) and the original MAB algorithms (black solid line) in all settings. In Figure 3, we plot the biases of various methods assuming the 5-arm Bernoulli model. A similar pattern is observed.

4.2 Regret

We carry out a numerical study to examine the regret of the rMAB algorithms. We set T as 500 and 10,000 to illustrate the long-term trend. The computation of regret for the cMLE is extremely time-consuming for large values of T, and therefore it has been omitted from the comparison. The results are shown in Figure 4 and others in the appendix. We can see that the

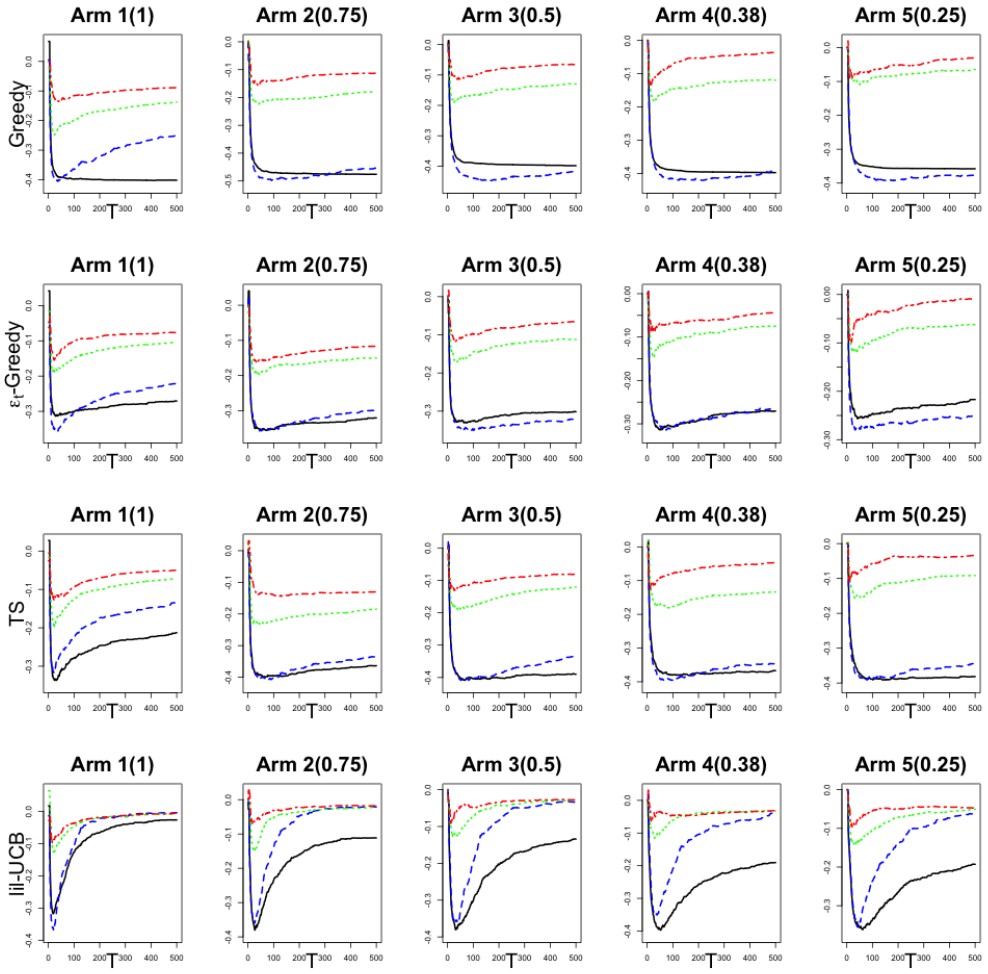

Figure 2: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Gaussian case with $T = 500$ and 1,000 replications. The means of the five arms are set as 1, 0.75, 0.5, 0.38, and 0.25 respectively. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dotdash).

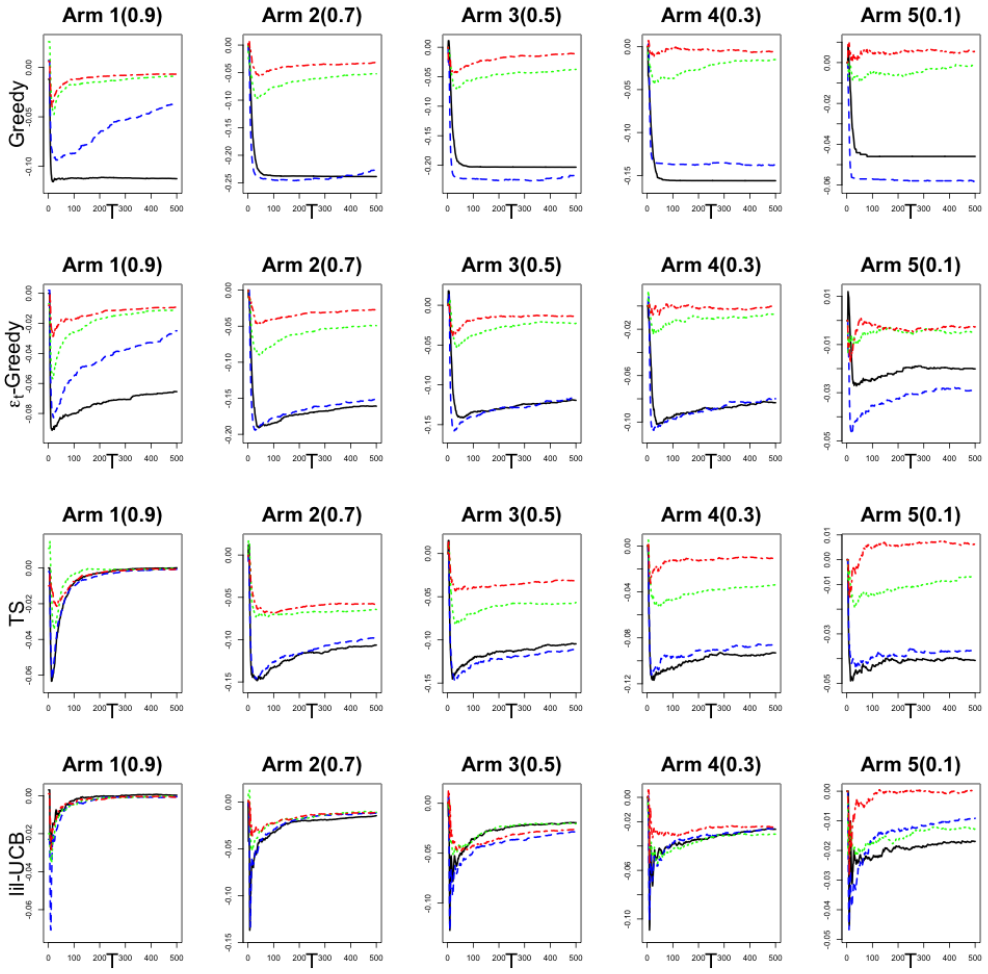

Figure 3: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Bernoulli case with $T = 500$ and 1,000 replications. The parameters of the five arms are set as 0.9, 0.7, 0.5, 0.3, and 0.1 respectively. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dotdash).

difference between rTS and TS is negiligible. It is noteworthy that TS reaches optimal levels of regret (Auer [2002], Agrawal and Gpyal [2013]). Thus, the extra regret incurred when using rTS is small.

It is remarkable that the Greedy algorithm has the capacity to reduce both bias and regret at the same time. We believe that the rMAB algorithm increases the rate of exploration, particularly in the initial stages, which helps to prevent being stuck in sub-optimal arms. This improvement in exploration leads to a decrease in regret.

4.3 Statistical Inference

We conduct a hypothesis test (Equation 3) to evaluate the properties of statistical inferene on the parameter $\mu_1 - \mu_2$. The number of replications of the experiment is 1,000. For each replication, we compute the two-sample T-statistic and its associated p-value. We reject the null hypothesis if the p-value is less than or equal to α, which is set to 0.05. Additionally, we construct the 95% confidence interval for the parameter $\mu_1 - \mu_2$ and calculate the empirical coverage probability based on 1,000 replications. The results are reported in Figure 5 and others in the appendix, with the left panel showing the Type I error rates of different methods and the right panel showing the coverage probabilities. It is evident that both the original MAB algorithm and the DP result in an inflated Type I error rate and a lower coverage probability than expected. The rMAB algorithm substantially improves to achieve the nominal level. In Figure 6, we reported these quantities of various methods for the 5-arm Bernoulli case. A similar pattern is observed. Due to the page limit, extensive numerical results are reported in Section 6.2 of the appendix.

The codes for the simulation studies are made available via github (`https://github.com/zhaozhg81/rMAB`).

5 Discussion

The presence of bias in the sample mean of the adaptively collected data can lead to significant issues in subsequent statistical inferences, such as an increase in type I error and low coverage probabilities when constructing confidence intervals. To address this, we propose a randomized Multi-Armed Bandit (rMAB) algorithm that aims to reduce the dependence and bias present in adaptively collected data. We focus on the rUCB and rTS algorithms, which have been shown to achieve optimality in terms of regret when the allocation probability is chosen correctly. Our numerical investigations demonstrate that the rMAB approach substantially reduces the magnitude of bias and exhibits favorable performance in subsequent inferential methods. These results suggest that rMAB may be a promising solution to address bias in adaptive data collection scenarios.

The concept of rMAB can be extended to other adaptive data collection settings. By incorporating randomization into the data collection process, we can potentially mitigate bias and improve the overall performance of statistical inference in a wide range of scenarios.

Existing research mainly focuses on mitigating bias through modifications in the data collection procedure, usually during the pre-data stage. However, in practical scenarios, we often encounter situations where the data has already been collected using standard MAB algorithms.

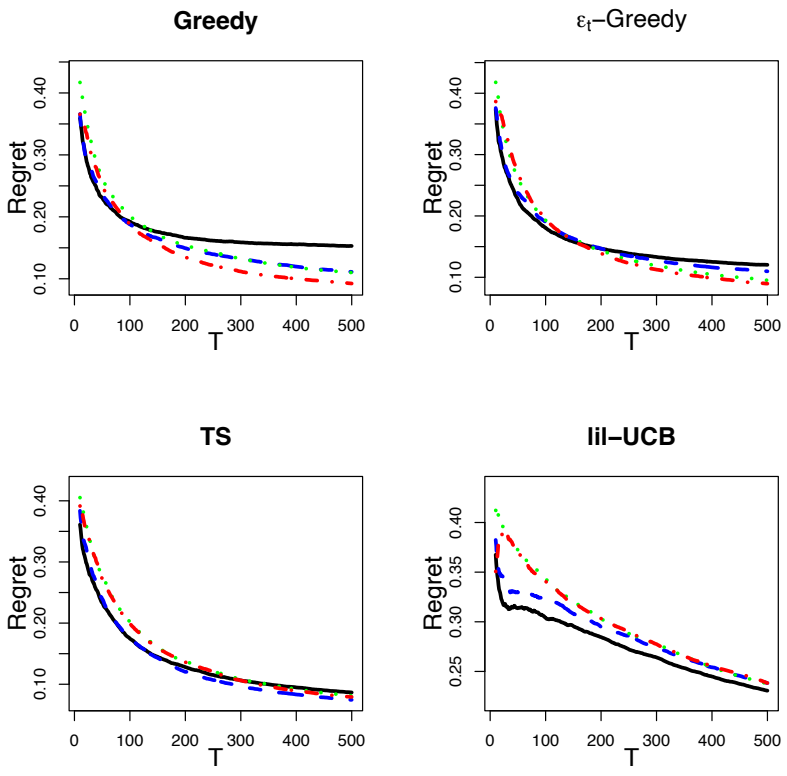

Figure 4: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 5-arm Gaussian case with $T = 500$ and 1,000 replications. The means of the five arms are set as 1, 0.75, 0.5, 0.38, and 0.25 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

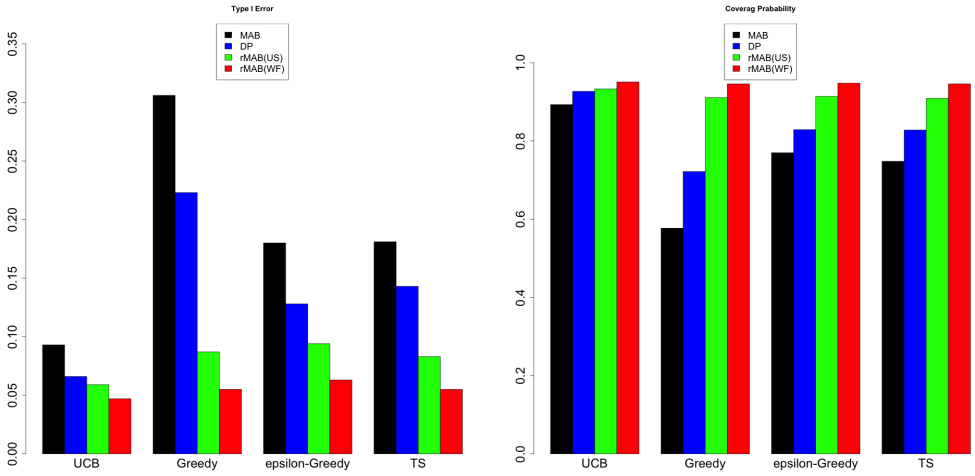

Figure 5: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $\mu_2 - \mu_1$ in the Gaussian design with $K = 5$. We set T to 500 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

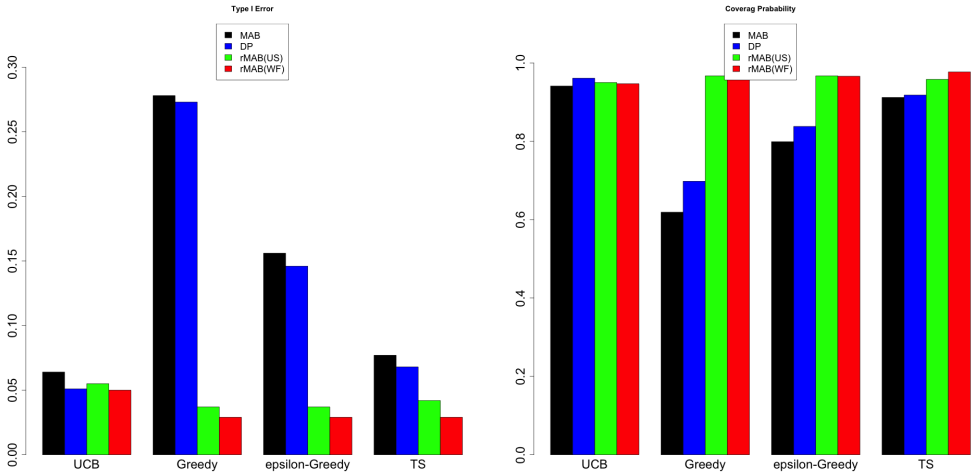

Figure 6: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $p_2 - p_1$ in the Bernoulli design with $K = 5$. We set T to 500 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

Addressing the issue of bias at the post-data stage is an important and challenging question that requires further investigation. Developing effective methodologies to tackle bias in adaptively collected data post-data collection is an important area for future research.

Acknowledgment

Tong Wang's research is partially supported by the grant NSF-IIS 1633283. Zhigen Zhao's research is partially supported by the grant NSF-IIS 1633283 and NSF-DMS-2311216.

References

Shipra Agrawal and Navin Gpyal. Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th International Con- ference on Artificial Intelligence and Statistics (AISTATS)*, volume 31, pages 99–107, 2013.

Peter Auer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47 (235):235–256, 2002.

A D Barker, C C Sigman, G J Kelloff, N M Hylton, D A Berry, and L J Esserman. I-SPY 2 : An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Nature*, 86(1):97–100, 2009. ISSN 0009-9236.

Sebastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Yi Gai and Bhaskar Krishnamachari. Online Learning Algorithms for Stochastic Water-Filling. In *2012 Information Theory and Applications Workshop*, pages 1–5. IEEE, 2012.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sebastien Bubeck. lil ' UCB : An Optimal Exploration Algorithm for Multi-Armed. In *JMLR: Workshop and Conference Proceedings*, volume 35, pages 1–17, 2014.

AH Jonker, A Mills, LPL Lau, S Ayme, and S Day. Small Population Clinical Trials Task Force Workshop Report and Recommendations July 2016 Small Population Clinical Trials : Challenges in the Field of Rare Diseases. Technical Report July, 2016.

Michael N. Katehakis and Jr. Arthur F. Veinott. The Multi-Armed Bandit Problem : Decomposition and Computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

Edward S. Kim, Roy S. Herbst, Ignacio I. Wistuba, J. Jack Lee, Suzanne E. Davis, and Waun K. Hong. The BATTLE Trial: Personalizing Therapy for Lung Cancer Edward. *Cancer Discovery*, 1(1):44–53, 2011.

T.L.Lai Lai and Herbert Robbins. Asymptotically Efficient Adaptive Rules. *Advances in Applied Mathematics*, 22:4–22, 1985.

Sylvain Lamprier, Thibault Gisselbrecht, and Patrick Gallinari. Contextual bandits with hidden contexts: a focused data capture from social media streams. *Data Mining and Knowledge Discovery*, 33(6):1853–1893, 2019. ISSN 1573756X.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 661–670, 2010.

Seth Neel and Aaron Roth. Mitigating Bias in Adaptive Data Gathering via Differential Privacy. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright*, pages 1–10, 2018.

Xinkun Nie, Jonathan Taylor, and James Zou. Why Adaptively Collected Data Have Negative Bias and How to Correct for It. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain*, volume 84, 2018.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? *Advances in Neural Information Processing Systems*, 32, 2019.

Ole B Suhr, Teresa Coelho, Juan Buades, Jean Pouget, Isabel Conceicao, John Berk, Hartmut Schmidt, Márcia Waddington-cruz, Josep M Campistol, Brian R Bettencourt, Akshay Vaishnaw, and Jared Gollob. Efficacy and safety of patisiran for familial amyloidotic polyneuropathy : a phase II multi-dose study. *Orphanet Journal of Rare Diseases*, pages 1–9, 2015. ISSN 1750-1172.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Min Xu, Tao Qin, and Tie-Yan Liu. Estimation Bias in Multi-Armed Bandit Algorithms for Search Advertising. In *Advances in Neural Information Processing Systems*, pages 2400–2408, 2013.

Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. Online context-aware recommendation with time varying multi-armed bandit. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:2025–2034, 2016.

6 Appendix

6.1 Technical Proofs

6.1.1 Notations

Without loss of generality, we assume that the first-arm is the optimal arm. Before proving Theorems 3.1 & 3.2, we define some notations:

- T: The total time horizontal.
- K: The total number of arms.
- Filtration $\mathcal{F}_{t-1} = \{I_\omega, r_{I_\omega}(\omega); I_\omega = 1, ..., K, \omega = 1, ..., t-1\}$: the history of plays until time $t-1$, where I_ω denotes the arm played at time t, and $r_k(t)$ denotes the reward observed for arm k at time t.
- $N_k(T) = \sum_{t=1}^{T} \mathbf{1}(I_t = k)$: the number of pulls for k-th arm.
- $\mathbb{E}[N_k(T)] = \sum_{t=1}^{T} P(I_t = k)$
- $\Delta_k = \mu_1 - \mu_k$
- $d(\mu_k, \mu_1) = \mu_k log\frac{\mu_k}{\mu_1} + (1-\mu_k) log\frac{1-\mu_k}{1-\mu_1}$
- $E_m(t)$: The event that the algorithm would enter into the MAB part.
- $E_r(t)$: The event that the algorithm would enter into the randomzied part.
- $E_m^{ts}(t)$: The event that the algorithm would enter into the MAB part and corresponding MAB algorithm is set as TS in advance.
- $E_m^{ucb}(t)$: The event that the algorithm would enter into the MAB part and corresponding MAB algorithm is set as UCB1 in advance.

For the k-th arm, the number of pulls $N_k(T)$ can be divided as 2 parts: $N_k^r(T)$, the total number of plays of arm k implemented by Random Process, and $N_k^m(T)$, the number of pulls from MAB Process. Then, we have

$$\mathbb{E}[N_k(T)] = \mathbb{E}\left\{\sum_{t=1}^{T}(\mathbf{1}(I_t = k, E_m(t)) + \mathbf{1}(I_t = k, E_r(t))\right\} = \mathbb{E}[N_k^r(T)] + \mathbb{E}[N_k^m(T)]. \tag{4}$$

Note that

$$\mathbb{E}[N_k^r(T)] = \sum_{t=1}^{T} \mathbb{P}(I_t = k | E_r(t)) \cdot \mathbb{P}(E_r(t)) \leq \sum_{t=1}^{T} \lambda(t) \leq K \log T. \tag{5}$$

Thus the remaining work is try to get the bound of $\mathbb{E}[N_k^m(T)]$.

6.1.2 Proof for Theorem 3.1.

In additional to the notations we introduce before, we further define the following notations for Thompson Sampling Algorithm:

- $n_k(t)$: The number of plays of arm k, until time $t-1$. Also equals to $N_K(t-1)$.
- $S_k(t)$: The number of successes(reward=1) observed at time t for the k-th arm.
- $F_k(t)$: The number of failures(reward=0) observed at time t for the k-th arm. Specifically, $n_k(t) = S_k(t) + F_k(t)$.
- $\hat{\mu}_k(t) = \dfrac{S_k(t)}{N_k(t)}$
- For each arm k, we will choose two thresholds x_k and y_k, such that $\mu_k < x_k < y_k < \mu_1$.
- $L_k(T) = \frac{\ln T}{d(x_k, y_k)}$
- $\theta_k(t)$: For each arm k, the sample from corresponding distribution at time t.
- $p_{k,t} = \mathbb{P}[\theta_1(t) > y_k | \mathcal{F}'_{t-1}; Pro(t) = TS]$
- τ_j: The time stamp at which j-th pull of the optimal arm happens.
- $E_k^\mu(t)$: the event $\hat{\mu}_k(t) > x_k$; $\overline{E_k^\mu(t)}$: the event $\hat{\mu}_k(t) \leq x_k$.
- $E_k^\theta(t)$: the event $\theta_k(t) > y_k$; $\overline{E_k^\theta(t)}$: the event $\theta_k(t) \leq y_k$.
- $E_{kj}^\theta(t)$: The event $\theta_k(t) \geq \theta_j(t)$.

Step 1.
Note that $\mathbb{E}[N_k^m(T)]$ can be written as

$$\begin{aligned}
&\mathbb{E}[N_k^m(T)] = \sum_{t=1}^{T} \mathbb{P}(I_t = k; E_m^{ts}(t)) \leq \sum_{t=1}^{T} \mathbb{P}(I_t = k | E_m^{ts}(t)) \\
&= \sum_{t=1}^{T} \mathbb{P}(I_t = k; \overline{E_k^\mu(t)} | E_m^{ts}(t)) + \sum_{t=1}^{T} \mathbb{P}(I_t = k; E_k^\mu(t) | E_m^{ts}(t)) \\
&= \sum_{t=1}^{T} \mathbb{P}(I_t = k; \overline{E_k^\mu(t)}; \overline{E_k^\theta(t)} | E_m^{ts}(t)) + \sum_{t=1}^{T} \mathbb{P}(I_t = k; \overline{E_k^\mu(t)}; E_k^\theta(t) | E_m^{ts}(t)) + \sum_{t=1}^{T} \mathbb{P}(I_t = k; E_k^\mu(t) | E_m^{ts}(t)).
\end{aligned} \tag{6}$$

Three terms in the last expression will be bounded in **Step 2**, **Step 3** and **Step 4** respectively.
Step 2. To get a bound for the first term, we state the following two lemmas.

Lemma 6.1 *Given* $\forall t \in [1,T]$, *and* $k \neq 1$,

$$\mathbb{P}(I_t = k; \overline{E_k^{\mu}(t)}; \overline{E_k^{\theta}(t)} | \mathcal{F}_{t-1}; E_m^{ts}(t)) \leq \frac{1-p_{k,t}}{p_{k,t}} \cdot \mathbb{P}(I_t = 1; \overline{E_k^{\mu}(t)}; \overline{E_k^{\theta}(t)} | \mathcal{F}_{t-1}; E_m^{ts}(t)),$$

where

$$p_{k,t} = \mathbb{P}(\theta_1(t) > y_k | \mathcal{F}_{t-1}; E_m^{ts}(t)) = \mathbb{P}(E_1^{\theta}(t) | \mathcal{F}_{t-1}; E_m^{ts}(t))$$

Lemma 6.2

$$\mathbb{E}[\frac{1}{p_{k,\tau_q}} | E_m^{ts}(\tau_q)] \leq \begin{cases} 1 + \frac{3}{\Delta_k'}, & q < \frac{8}{\Delta_k'}, \\ 1 + \Theta\big(e^{-{\Delta_k'}^2 q/2} + \frac{e^{-D_k q}}{(q+1){\Delta_k'}^2} + \frac{1}{e^{{\Delta_k'}^2 q/4} - 1}\big), & q \geq \frac{8}{\Delta_k'}, \end{cases}$$

where $\Delta_k' = \mu_1 - y_k$, $D_k = y_k \log \frac{y_k}{\mu_1} + (1-y_k) \log \frac{1-y_k}{1-\mu_1}$ *and* τ_q *represents the time stamp at which the* q*-th pull happens for arm* k.

Together with Lemma 6.1 and Lemma 6.2, we have the following,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(I_t = k; \overline{E_k^{\mu}(t)}; \overline{E_k^{\theta}(t)} | E_m^{ts}(\tau_q)) \leq \sum_{t=1}^{T} \frac{1-p_{k,t}}{p_{k,t}} \mathbb{P}(I_t = 1; \overline{E_k^{\mu}(t)}; \overline{E_k^{\theta}(t)} | E_m^{ts}(\tau_q)) \\
= &\sum_{t=1}^{T} \mathbb{E}[\frac{1-p_{k,t}}{p_{k,t}} \mathbf{1}(I_t = 1); \overline{E_k^{\mu}(t)}; \overline{E_k^{\theta}(t)} | E_m^{ts}(\tau_q)] \leq \sum_{q=0}^{T} \mathbb{E}[\frac{1-p_{k,\tau_q}}{p_{k,\tau_q}} \mathbf{1}(I_{\tau_q+1} = 1); \overline{E_k^{\mu}(\tau_q)}; \overline{E_k^{\theta}(\tau_q)} | E_m^{ts}(\tau_q)] \\
\leq &\sum_{q=0}^{T} \mathbb{E}[\frac{1-p_{k,\tau_q}}{p_{k,\tau_q}} | E_m^{ts}(\tau_q)] \\
= &\sum_{q=0}^{T} \mathbb{E}[\frac{1}{p_{k,\tau_q}} - 1 | E_m^{ts}(\tau_q)] = \sum_{q=0}^{8/{\Delta_k}'} \mathbb{E}[\frac{1}{p_{k,\tau_q}} - 1 | E_m^{ts}(\tau_q)] + \sum_{q=8/{\Delta_k}'}^{T} \mathbb{E}[\frac{1}{p_{k,\tau_q}} - 1 | E_m^{ts}(\tau_q)] \\
\leq &\frac{8}{{\Delta_k}'}(1 + \frac{3}{{\Delta_k}'} - 1) + \sum_{j=0}^{T} \Theta\big(e^{-{\Delta_k'}^2 q/2} + \frac{e^{-D_k q}}{(q+1){\Delta_k'}^2} + \frac{1}{e^{{\Delta_k'}^2 q/4} - 1}\big) \\
= &\frac{24}{{\Delta_k'}^2} + \sum_{j=0}^{T} \Theta\big(e^{-{\Delta_k'}^2 q/2} + \frac{e^{-D_k q}}{(q+1){\Delta_k'}^2} + \frac{1}{e^{{\Delta_k'}^2 q/4} - 1}\big). \qquad (7)
\end{aligned}$$

Step 3. Now, we consider the second term in (6). Note that

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}(I_t=k,\overline{E_k^{\mu}(t)};E_k^{\theta}(t)|E_m^{ts}(t))\\
=\;&\sum_{t=1}^{\tau}\mathbb{P}(I_t=k;\overline{E_k^{\mu}(t)};E_k^{\theta}(t)|E_m^{ts}(t))+\sum_{t=\tau+1}^{T}\mathbb{P}(I_t=k;\overline{E_k^{\mu}(t)};E_k^{\theta}(t)|E_m^{ts}(t))\\
\le\;&\sum_{t=1}^{\tau}\mathbb{P}(I_t=k|E_m^{ts}(t)]+\sum_{t=\tau+1}^{T}\mathbb{P}[I_t=k;\overline{E_k^{\mu}(t)};E_k^{\theta}(t)|E_m^{ts}(t))\\
\le\;&\sum_{t=1}^{\tau}\mathbb{P}(I_t=k|E_m^{ts}(t)]+\sum_{t=\tau+1}^{T}\mathbb{P}[I_t=k;E_k^{\theta}(t)|\overline{E_k^{\mu}(t)};E_m^{ts}(t)). \qquad (8)
\end{aligned}
$$

For the second term, note that

$$
\begin{aligned}
&\mathbb{P}(I_t=k;E_k^{\theta}(t)|\overline{E_k^{\mu}(t)};E_m^{ts}(t))\le\mathbb{P}(E_k^{\theta}(t)|\overline{E_k^{\mu}(t)};E_m^{ts}(t))=\mathbb{P}(Beta(S_k(t)+1,F_k(t)+1)>y_k|\overline{E_k^{\mu}(t)})\\
=\;&\mathbb{P}(Beta(q_k(t)\hat{\mu}_k(t)+1,q_k(t)(1-\hat{\mu}_k(t))+1)|\overline{E_k^{\mu}(t)})\le\mathbb{P}(Beta(q_k(t)x_k+1,q_k(t)(1-x_k)+1)>y_k)\\
=\;&F^{Bin}_{q_k(t)+1,y_k}(q_k(t)\cdot x_k)=\mathbb{P}(\sum_{j=1}^{q_k(t)+1}Z_j\le q_k(t)x_k)=\mathbb{P}(\sum_{j=1}^{q_k(t)+1}Z_j-(q_k(t)+1)y_k\le q_k(t)x_k-(q_k(t)+1)y_k)\\
\le\;&\mathbb{P}(\sum_{j=1}^{q_k(t)+1}Z_j-(q_k(t)+1)y_k\le-(q_k(t)+1)(y_k-x_k))\le e^{\frac{-2(q_k(t)+1)^2(y_k-x_k)^2}{q_k(t)+1}}=e^{-2(q_k(t)+1)(y_k-x_k)^2}\\
\le\;&e^{-(q_k(t)+1)d(x_k,y_k)}. \qquad (9)
\end{aligned}
$$

In the derivation of (9), Z_j represents the variable following Bernoulli$(q_k(t)\cdot x_k)$. Based on $L_k(T)=\frac{\ln T}{d(x_k,y_k)}$, if we choose $\tau=q_k(t)>L_k(t)$, then

$$(q_k(t)+1)d(x_k,y_k)>L_k(t)d(x_k,y_k)=\ln T$$

i.e.,

$$e^{-(q_k(t)+1)d(x_k,y_k)}\le e^{-lnT}=\frac{1}{T}.$$

As a result,

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}(I_t=k;\overline{E_k^{\mu}(t)};E_k^{\theta}(t)|E_m^{ts}(t))\le\sum_{t=1}^{\tau}\mathbb{P}(I_t=k|E_m^{ts}(t))+\sum_{t=\tau+1}^{T}\mathbb{P}(I_t=k;E_k^{\theta}(t)|\overline{E_k^{\mu}(t)};E_m^{ts}(t))\\
\le\;&\mathbb{E}[\sum_{t=1}^{\tau}\mathbf{1}(I_t=k|E_m^{ts}(t))]+\sum_{t=\tau+1}^{T}\frac{1}{T}\le L_k(T)+1. \qquad (10)
\end{aligned}
$$

Step 4. Now, we consider the thrid term in (6). Note that

$$\sum_{t=1}^{T}\mathbb{P}(I_t=k;E_k^{\mu}(t)|E_m^{ts}(t)) \leq \mathbb{E}\left\{\sum_{q=0}^{T-1}\sum_{t=\tau_q+1}^{\tau_{q+1}}\mathbf{1}(I_t=k;E_k^{\mu}(t)|E_m^{ts}(t))\right\}$$

$$= \mathbb{E}\left\{\sum_{q=0}^{T-1}\sum_{t=\tau_q+1}^{\tau_{q+1}}\mathbf{1}(I_t=k|E_m^{ts}(t))\cdot\mathbf{1}(E_k^{\mu}(t)|E_m^{ts}(t))\right\}$$

$$= \mathbb{E}\left\{\sum_{q=0}^{T-1}\mathbf{1}(\hat{\mu}_k(\tau_{q+1})>x_k|E_m^{ts}(t))\right\} \leq 1+\mathbb{E}\left\{\sum_{q=1}^{T-1}\mathbf{1}(E_k^{\mu}(\tau_{q+1})|E_m^{ts}(t))\right\}$$

$$= 1+\sum_{q=1}^{T-1}\mathbb{P}(E_k^{\mu}(\tau_{q+1})|E_m^{ts}(t)) = 1+\sum_{q=1}^{T-1}\mathbb{P}(\frac{S_k(\tau_{q+1})}{q+1}>\mu_k+x_k-\mu_k|E_m^{ts}(t))$$

$$= 1+\sum_{q=1}^{T-1}\mathbb{P}(S_k(\tau_{q+1})>(q+1)\mu_k+(q+1)(x_k-\mu_k)|E_m^{ts}(t))$$

$$\leq 1+\sum_{q=1}^{T-1}e^{-(q+1)d(x_k,\mu_k)} \leq 1+\frac{1}{d(x_k,\mu_k)}. \tag{11}$$

Note that the first equality holds because given $t\in[\tau_q+1,\tau_{q+1}]$, $I_t=k$ can only happen at time τ_{q+1} which means $I_t=k$ is independent of $E_k^{\mu}(t)$.

Step 5. Combine results in Steps 2, 3 and 4 together. For any $0<\epsilon<1$, choose $x_k\in(\mu_k,\mu_1)$ such that $d(x_k,\mu_1)=\frac{d(\mu_k,\mu_1)}{1+\epsilon}$. Choose $y_k\in(x_k,\mu_1)$ such that $d(x_k,y_k)=\frac{d(x_k,\mu_1)}{1+\epsilon}=\frac{d(\mu_k,\mu_1)}{(1+\epsilon)^2}$. Then,

$$L_k(T)=\frac{lnT}{d(x_k,y_k)}=(1+\epsilon)^2\frac{lnT}{d(\mu_k,\mu_1)}.$$

Some algebraic manipulations on $d(x_k,\mu_1)=\frac{d(\mu_k,\mu_1)}{1+\epsilon}$ leads to

$$x_k-\mu_k\geq\frac{\epsilon}{1+\epsilon}\cdot\frac{d(\mu_k,\mu_1)}{ln\big(\frac{\mu_1(1-\mu_k)}{\mu_k(1-\mu_1)}\big)}.$$

Hence,

$$\mathbb{E}[N_k^m(T)] = \sum_{t=1}^{T} \mathbb{P}(I_t = k; \overline{E_k^{\mu}(t)}; \overline{E_k^{\theta}(t)} | E_m^{ts}(t)) + \sum_{t=1}^{T} \mathbb{P}(I_t = k; \overline{E_k^{\mu}(t)}; E_k^{\theta}(t) | E_m^{ts}(t))$$

$$+ \quad \sum_{t=1}^{T} \mathbb{P}(I_t = k; E_k^{\mu}(t) | E_m^{ts}(t))$$

$$\leq \quad \frac{24}{{\Delta'_k}^2} + \sum_{j=0}^{T} \Theta\big(e^{-{\Delta'_k}^2 j/2} + \frac{e^{-D_k j}}{(j+1){\Delta'_k}^2} + \frac{1}{e^{{\Delta'_k}^2 j/4} - 1}\big) + L_k(T) + 1 + \frac{1}{d(x_k, \mu_k)} + 1$$

$$\leq \quad \{\frac{24}{{\Delta'_k}^2} + \Theta\big(\frac{1}{{\Delta'_k}^2} + \frac{1}{{\Delta'_k}^2 D} + \frac{1}{{\Delta'_k}^4} + \frac{1}{{\Delta'_k}^2}\big) + (1+\epsilon)^2 \frac{lnT}{d(\mu_k, \mu_1)} + O(\frac{1}{\epsilon^2})\}$$

$$= \quad O(1) + (1+\epsilon)^2 \frac{lnT}{d(\mu_k, \mu_1)} + O(\frac{1}{\epsilon^2}). \tag{12}$$

Finally, the expected average regret for the TS algorithm is

$$\mathbb{E}[N_k(T)] = \mathbb{E}[N_k^r(T)] + \mathbb{E}[N_k^m(T)] \leq \log T + O(1) + (1+\epsilon)^2 \frac{lnT}{d(\mu_k, \mu_1)} + O(\frac{1}{\epsilon^2})$$

$$\leq \quad (\frac{(1+\epsilon)^2}{d(\mu_k, \mu_1)} + 1) \log T + O(1 + \frac{1}{\epsilon^2}), \tag{13}$$

and

$$\mathbb{E}[R(T)] = \sum_{k=2}^{K} \Delta_k \mathbb{E}[N_k(T)] \leq \sum_{k=2}^{K} [\big(\frac{(1+\epsilon)^2}{d(\mu_k, \mu_1)} + 1\big) \log T + O(1 + \frac{1}{\epsilon^2})] \Delta_k. \tag{14}$$

For Gaussian distribution, similarly using the idea of Theorem 3 in Agrawal and Gpyal [2013], we can get the regret upper bound:

$$\mathbb{E}[R(T)] \leq \sum_{k=2}^{K} [\big(\log T \Delta_k + \frac{18 \log(T\Delta_k^2)}{\Delta_k}\big) + (e^{11} + 5 + \frac{13}{2\Delta_k})]. \tag{15}$$

6.1.3 Proof for Theorem 3.2.

Our proof of the regret bound is based on the same technique used in the UCB1 algorithm Auer [2002]. We demonstrate that the upper bound of regret is not adversely affected when the UCB1 algorithm is applied in the rMAB framework. For any positive constant value l, we can write

$E[N_k^m(T)]$ as:

$$\begin{aligned}
&\mathbb{E}[N_k^m(T)] = \mathbb{E}\sum_{t=1}^{T}\mathbf{1}(I_t = k; E_m^{ucb}(t)) \\
= \quad & 1 + \mathbb{E}\sum_{t=K+1}^{T}\mathbf{1}(I_t = k, E_m^{ucb}(t)) \leq 1 + \mathbb{E}\sum_{t=K+1}^{T}\mathbf{1}(I_t = k | E_m^{ucb}(t)) \qquad (16) \\
\leq \quad & 1 + \mathbb{E}\sum_{t=K+1}^{T}\mathbf{1}(I_t = k, N_k(t-1) < l | E_m^{ucb}(t)) + \mathbb{E}\sum_{t=K+1}^{T}\mathbf{1}(I_t = k, N_k(t-1) \geq l | E_m^{ucb}(t)).
\end{aligned}$$

Notice that

$$\mathbb{E}\sum_{t=K+1}^{T}\mathbf{1}(I_t = k, N_k(t-1) < l | E_m^{ucb}(t) = \mathbb{E}\sum_{t=K+1}^{\tau_l}\mathbf{1}(I_t = k, N_k(t-1) < l | E_m^{ucb}(t)) \leq l - 1. \quad (17)$$

which does not rely on $\lambda(t)$. Hence:

$$\mathbb{E}[N_k^m(T)] \leq l + \mathbb{E}\sum_{t=K+1}^{T}\mathbf{1}(I_t = k, N_k(t-1) \geq l | E_m^{ucb}(t)). \qquad (18)$$

Next apply *Chernoff-Hoeffding Inequality* on $\hat{\mu}_k(t)$ to get the bound. Follow the proof of UCB1, set $l = \dfrac{8\log T}{\Delta_k^2}$, then

$$\mathbb{E}[N_k^m(T)] \leq \frac{8\log T}{\Delta_k^2} + \frac{\pi^2}{3} + 1. \qquad (19)$$

Put everything together, we know that

$$\begin{aligned}
&\mathbb{E}[N_k(T)] = \mathbb{E}[N_k^r(T)] + \mathbb{E}[N_k^m(T)] \\
\leq \quad & \log T + \frac{8\log T}{\Delta_k^2} + \frac{\pi^2}{3} + 1 \leq (\frac{8}{\Delta_k^2} + 1)\log T + \frac{\pi^2}{3} + 1, \qquad (20)
\end{aligned}$$

and

$$\mathbb{E}[R(T)] \leq \sum_{k=2}^{K}\left[(\frac{8}{\Delta_k} + \Delta_k)\log T + (\frac{\pi^2}{3} + 1)\Delta_k\right]. \qquad (21)$$

Proof of Lemma 6.1.
Note that whether $\overline{E_k^{\mu}(t)}$ is true or not has already been decided by $\mathcal{F}_{t-1}$, while $\overline{E_k^{\theta}(t)}$ does not. Thus we can assume $\mathcal{F}_{t-1}$ is such that $\overline{E_k^{\mu}(t)}$ is true (otherwise the probability on the left hand side is 0 and the inequality is trivially true). It then suffices to prove the following inequalities (22), (23), and (24):

$$\mathbb{P}(I_t = k; \overline{E_k^{\theta}(t)} | \mathcal{F}_{t-1}; E_m^{ts}(t)) \leq \frac{1 - p_{k,t}}{p_{k,t}} \cdot \mathbb{P}(I_t = 1; \overline{E_k^{\theta}(t)} | \mathcal{F}_{t-1}; E_m^{ts}(t)) \qquad (22)$$

$$
\begin{aligned}
& p_{k,t} \cdot \mathbb{P}(I'_t = k|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \cdot \mathbb{P}(\overline{E_k^\theta(t)}|\mathcal{F}_{t-1}; E_m^{ts}(t)) \\
\leq \quad & (1-p_{k,t}) \cdot \mathbb{P}(I'_t = 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \cdot \mathbb{P}(\overline{E_k^\theta(t)}|\mathcal{F}_{t-1}; E_m^{ts}(t)),
\end{aligned} \tag{23}
$$

and

$$
p_{k,t} \cdot \mathbb{P}(I_t = k|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \leq (1-p_{k,t}) \cdot \mathbb{P}(I_t = 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)). \tag{24}
$$

The inequality (24) can be derived based on the following two inequalities:

$$
\mathbb{P}(I_t = 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \geq p_{k,t} \cdot \mathbb{P}(E_{kj}^\theta(t), \forall j \neq 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)), \tag{25}
$$

and

$$
\mathbb{P}(I_t = k|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \leq (1-p_{k,t}) \cdot \mathbb{P}(E_{kj}^\theta(t), \forall j \neq 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)). \tag{26}
$$

To prove (25), note that

$$
\begin{aligned}
& \mathbb{P}(I_t = 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \geq \mathbb{P}(I_t = 1; E_{kj}^\theta(t), \forall j \neq 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \\
= \quad & \mathbb{P}(I_t = 1|E_{kj}^\theta(t), \forall j \neq 1; \overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \cdot \mathbb{P}(E_{kj}^\theta(t), \forall j \neq 1|\overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)).
\end{aligned} \tag{27}
$$

Now given the event: $E_{kj}^\theta(t)$: $\theta_k(t) \geq \theta_j(t), \forall j \neq 1; \theta_k(t) \leq y_k$, it holds that for all $j \neq k$, $j \neq 1$,

$$
\theta_j(t) \leq \theta_k(t) \leq y_k
$$

Then

$$
\begin{aligned}
& \mathbb{P}(I_t = 1|\theta_k(t) \geq \theta_j(t), \forall j \neq 1; \overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)]) \\
\geq & \mathbb{P}(\theta_1(t) > y_k|\theta_k(t) \geq \theta_j(t), \forall j \neq 1; \overline{E_k^\theta(t)}; \mathcal{F}_{t-1}; E_m^{ts}(t)) \\
\geq & \mathbb{P}(\theta_1(t) > y_k|\mathcal{F}_{t-1}; E_m^{ts}(t)) = p_{k,t}.
\end{aligned} \tag{28}
$$

The second last equality in (28) follows because given $\mathcal{F}_{t-1}$, $\theta_1(t)$ is independent of all the other $\theta_j(t), j \neq 1$ and hence independent of these events $\theta_k(t) \geq \theta_j(t), \forall j \neq 1$ and $\theta_1(t) > y_k$. This together with (27) gives(25).

For (26), note that

$$
\begin{aligned}
& \mathbb{P}(I_t = k|E_1^\theta(t); \mathcal{F}_{t-1}; E_m^{ts}(t)) = \mathbb{P}(E_{kj}^\theta(t), \forall j \neq k|E_1^\theta(t); \mathcal{F}_{t-1}; E_m^{ts}(t)) \\
\leq \quad & \mathbb{P}[E_{kj}^\theta(t), \forall j \neq 1; \overline{E_1^\theta(t)}; |E_1^\theta(t); \mathcal{F}_{t-1}; E_m^{ts}(t)] \\
= \quad & \mathbb{P}(\overline{E_1^\theta(t)}|E_1^\theta(t); \mathcal{F}_{t-1}; E_m^{ts}(t)) \cdot \mathbb{P}(E_{kj}^\theta(t), \forall j \neq 1|E_k^\theta(t); \mathcal{F}_{t-1}; E_m^{ts}(t)) \\
= \quad & (1-p_{k,t}) \cdot \mathbb{P}(E_{kj}^\theta(t), \forall j \neq 1|E_1^\theta(t); \mathcal{F}_{t-1}; E_m^{ts}(t)).
\end{aligned} \tag{29}
$$

Combine (25) and (26) and we complete the proof.

Proof of Lemma 6.2.

The proof relies on the relationship between the Beta distribution and the cumulative probability

distribution of the Binomial distribution. Note that $F^{Beta}_{\alpha,\beta}(y) = 1 - F^{Bin}_{\alpha+\beta-1,y}(\alpha-1)$, for all integers α, β. Then

$$\begin{aligned} & p_{k,\tau_q} = \mathbb{P}(\theta_1(\tau_q) > y_k | \mathcal{F}_{\tau_q - 1}; E^{ts}_m(\tau_q)) = \mathbb{P}(E^{\theta}_k(\tau_q) | \mathcal{F}_{\tau_q-1}; E^{ts}_m(\tau_q)) \\ = \quad & 1 - F^{Beta}_{s_1(\tau_q), f_1(\tau_q)}(y_k) = F^{Bin}_{q_1(\tau_q), y_k}(s_1(\tau_q)). \end{aligned} \tag{30}$$

For simplicity, we denote $y_k = y$ and $S_1(t) = S_{Bin}$ is a random variable which follows $Bin(k, \mu_1)$. In addition, $P(S_{Bin} > s)$ is a fixed value with the given s. Therefore:,

$$\begin{aligned} & \mathbb{E}[\frac{1}{p_{k,\tau_q}} | E^{ts}_m(\tau_q)] = \mathbb{E}[\frac{1}{\mathbb{P}(\theta_1(\tau_q) > y_k | \mathcal{F}_{\tau_q-1}; E^{ts}_m(\tau_q)}) \\ = \quad & \mathbb{E}[\frac{1}{1 - F^{Beta}_{s_1(\tau_q), f_1(\tau_q)}(y_k)}] = \mathbb{E}[\frac{1}{F^{Bin}_{q(\tau_q), y_k}(s_1(\tau_q))}] \\ = \quad & \mathbb{E}[\frac{1}{\mathbb{P}(S_{Bin} \le s_1(\tau_q))}] = \sum_{s=0}^{q} \frac{\mathbb{P}(S = s)}{\mathbb{P}(S_{Bin} \le s_1(\tau_q))} \\ = \quad & \sum_{s=0}^{q} \frac{\mathbb{P}(S = s)}{F^{Bin}_{q+1,y}(s)} = \sum_{s=0}^{q} \frac{f^{Bin}_{q,\mu_1}(s)}{F^{Bin}_{q+1,y}(s)}. \end{aligned} \tag{31}$$

According to Agrawal and Gpyal [2013], we know that

$$\sum_{s=0}^{q} \frac{f^{Bin}_{q,\mu_1}(s)}{F^{Bin}_{q+1,y}(s)} \le \begin{cases} \frac{3}{\Delta'_k}, & q < \frac{8}{\Delta'_k} \\ 1 + \Theta\big(e^{-{\Delta'_k}^2 q/2} + \dfrac{e^{-D_k q}}{(q+1){\Delta'_k}^2} + \dfrac{1}{e^{{\Delta'_k}^2 q/4} - 1}\big), & q \ge \frac{8}{\Delta'_k}. \end{cases}$$

6.2 More numerical results

In this section, we will present additional comprehensive numerical results to illustrate the advantages of randomized MAB algorithms. Due to its computational inefficiency, the cMLE has been omitted from our analysis.

6.2.1 Bias.

Figures 7-14 provide the bias of various methods under different parameter settings.

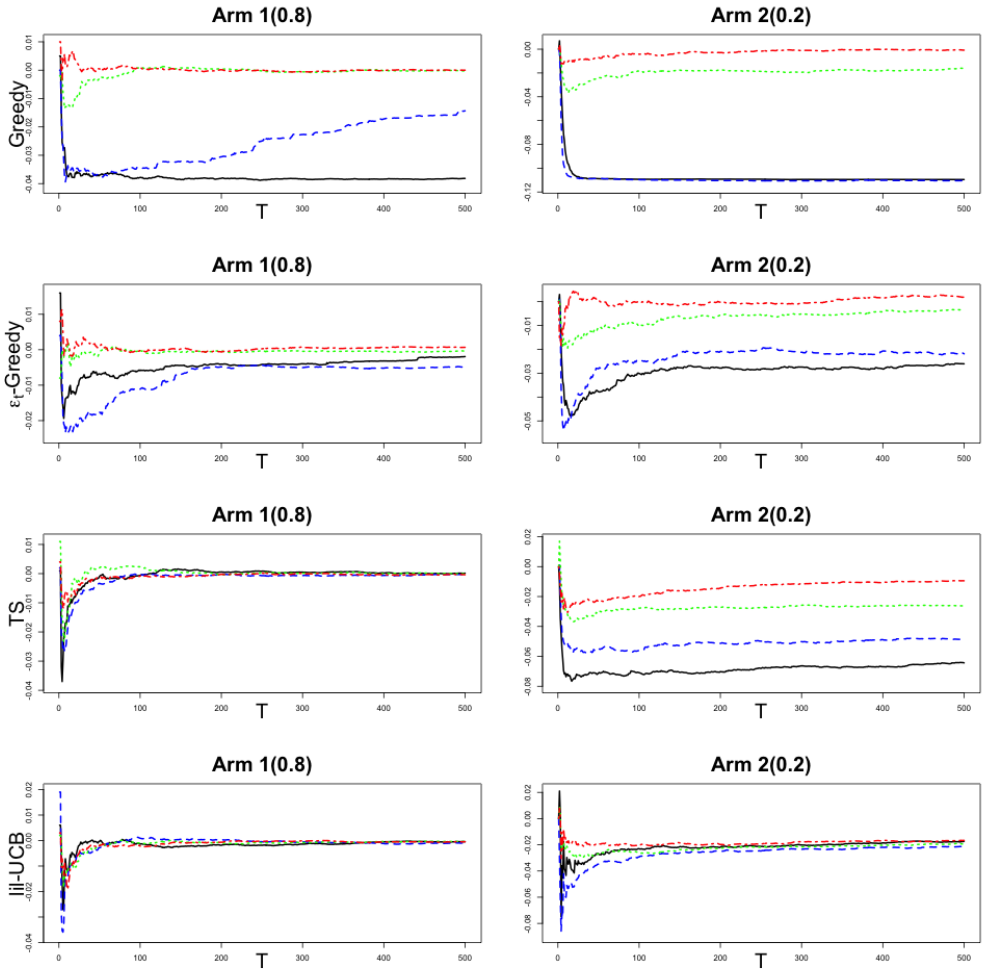

Figure 7: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 2-arm Bernoulli case with $T = 500$ and 1,000 replications. The true parameters of the arms are set as 0.8 and 0.2. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dot-dash).

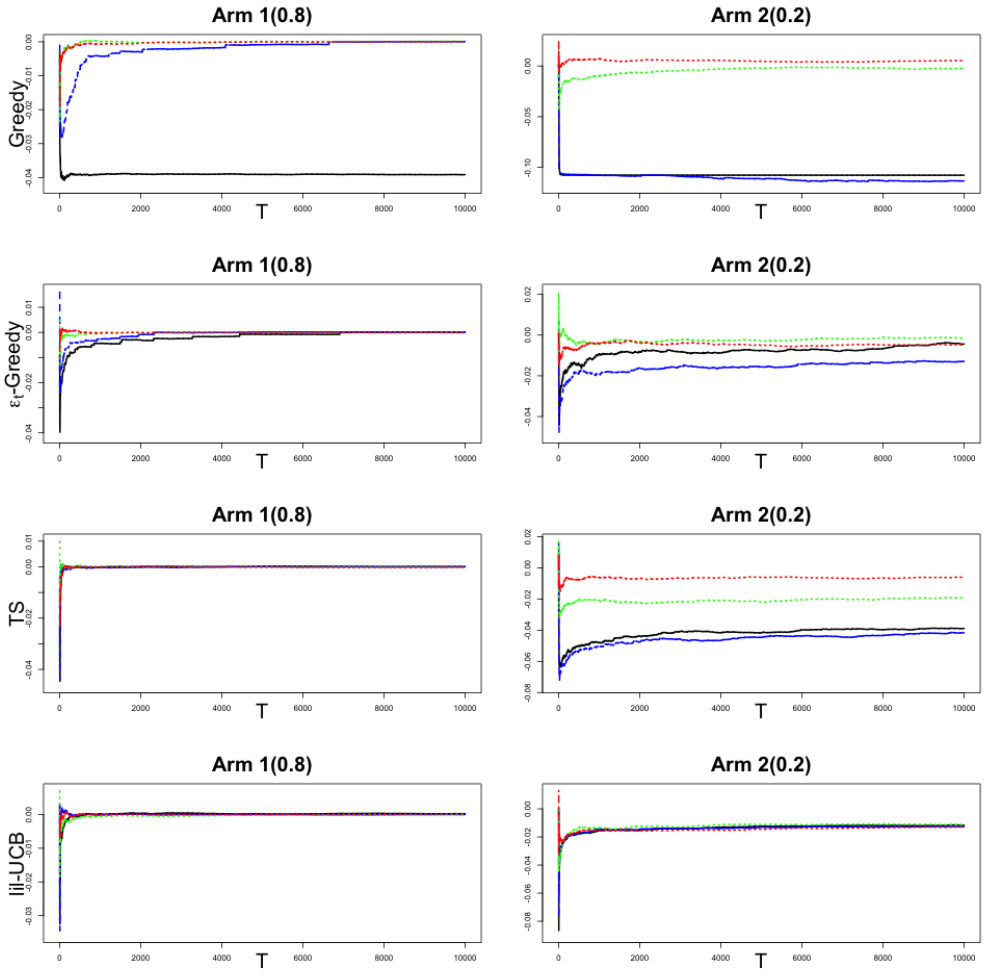

Figure 8: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 2-arm Bernoulli case with $T = 10,000$ and 1,000 replications. The true parameters of the arms are set as 0.8 and 0.2. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dot-dash).

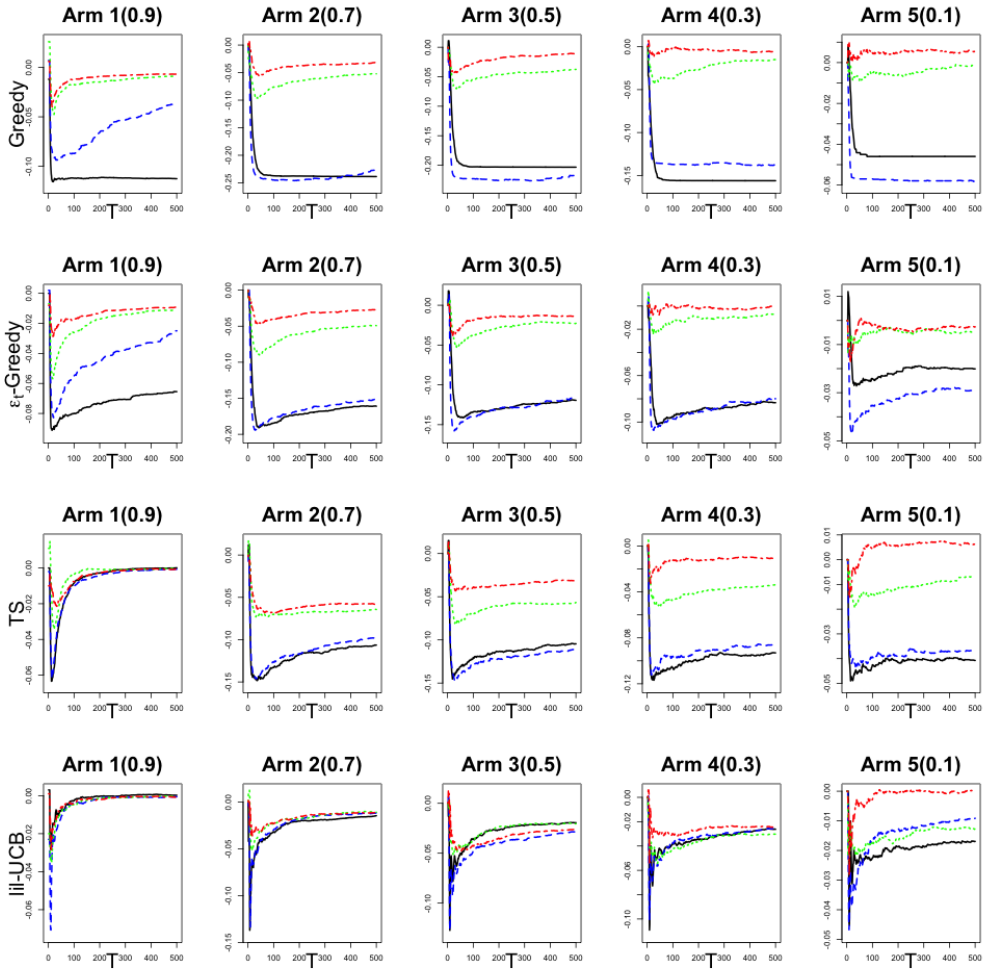

Figure 9: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Bernoulli case with $T = 500$ and 1,000 replications. The true parameters of the arms are set as 0.9, 0.7, 0.5, 0.3, and 0.1. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dotdash).

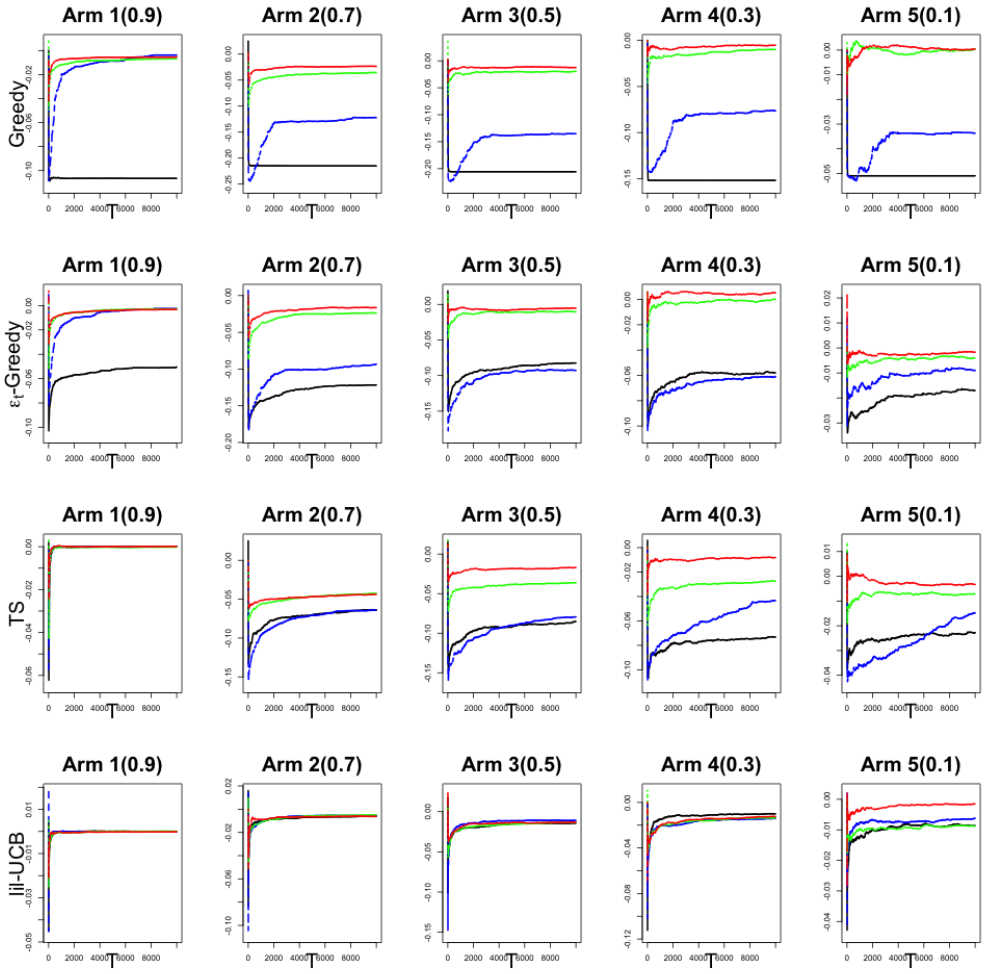

Figure 10: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Bernoulli case with $T = 10,000$ and 1,000 replications. The true parameters of the arms are set as 0.9, 0.7, 0.5, 0.3, and 0.1. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dotdash).

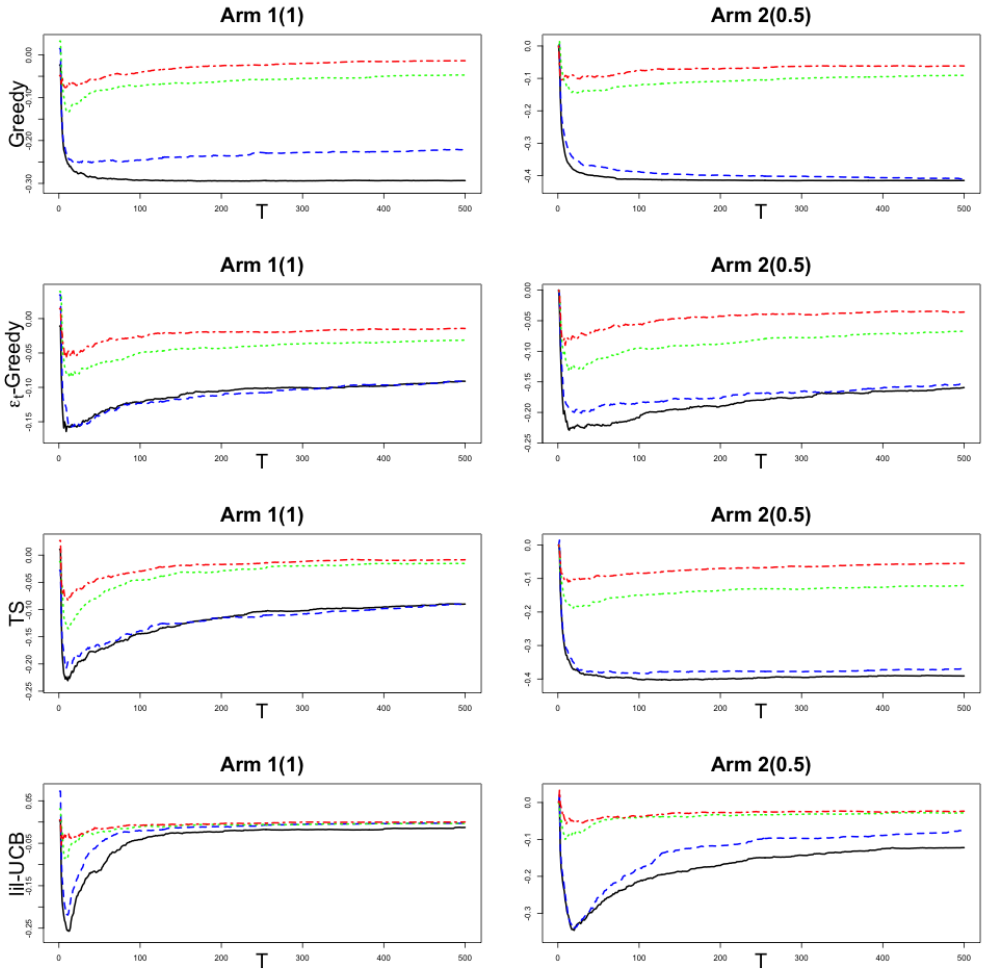

Figure 11: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Gaussian case with $T = 500$ and 1,000 replications. The true parameters of the arms are set as 1.0 and 0.5. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dot-dash).

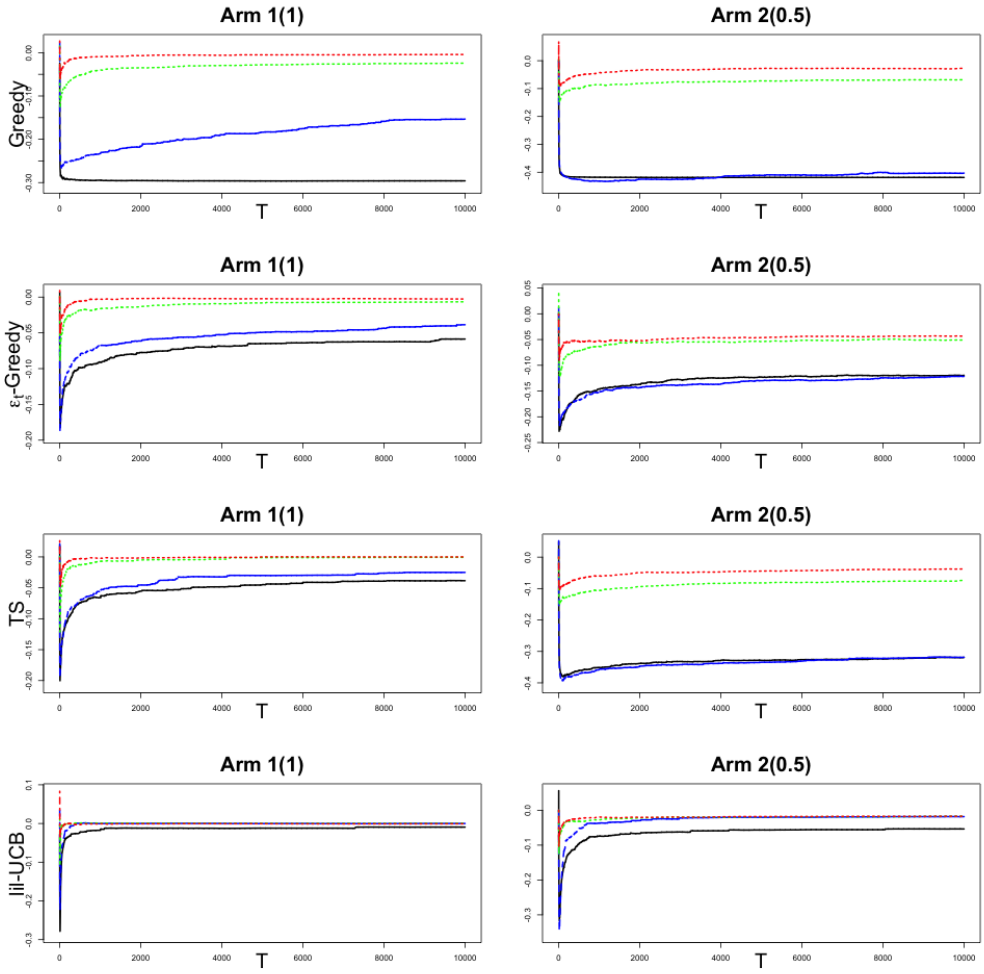

Figure 12: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Gaussian case with $T = 10,000$ and 1,000 replications. The true parameters of the arms are set as 1.0 and 0.5. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dot-dash).

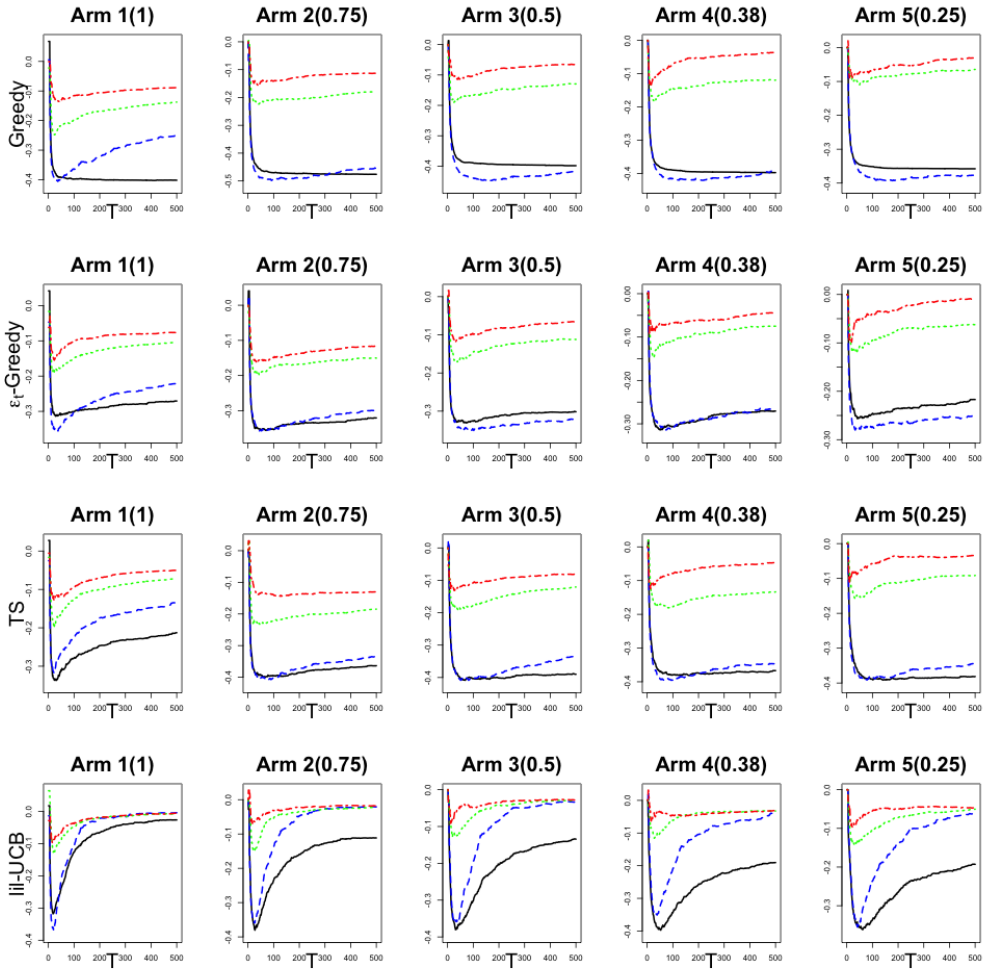

Figure 13: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Gaussian case with $T = 500$ and 1,000 replications. The true parameters of the arms are set as 1.0, 0.75, 0.5, 0.38, and 0.25. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dotdash).

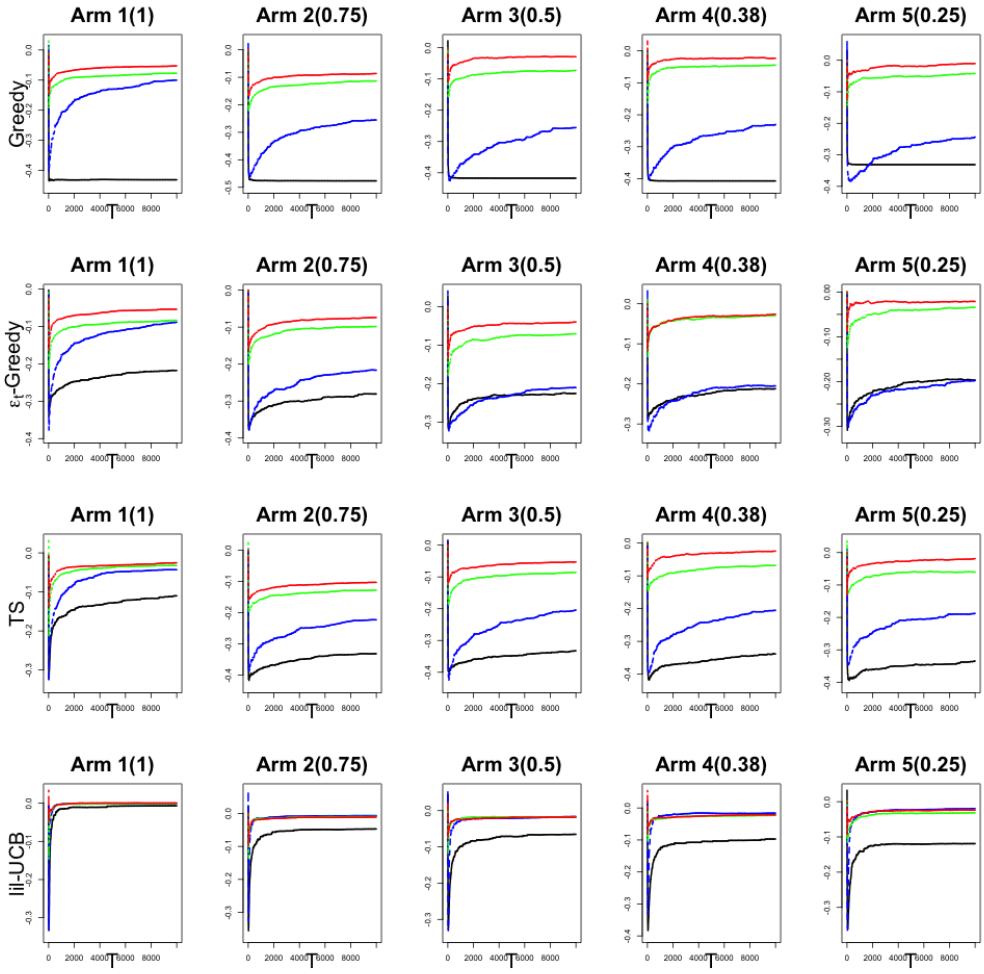

Figure 14: The biases of various methods (lilUCB, DP, rMAB-lilUCB(US), and rMAB-lilUCB(WF)) were assessed for the 5-arm Bernoulli case with $T = 10000$ and 1,000 replications. The true parameters of the arms are set as 1.0, 0.75, 0.5, 0.38, and 0.25. Four curves correspond to lilUCB (Black solid), DP (blue dashed), rMAB-lilUCB(US) (green dotted), and rMAB-lilUCB(WF) (red, dotdash).

6.2.2 Regret.

In this section, we provide the numerical result to compare the regret of various methods from Figure 15 - 22.

6.2.3 Statistical Inference.

In this section, we plot the result on type I errors and coverage probabilities based various methods in Figure 23-30.

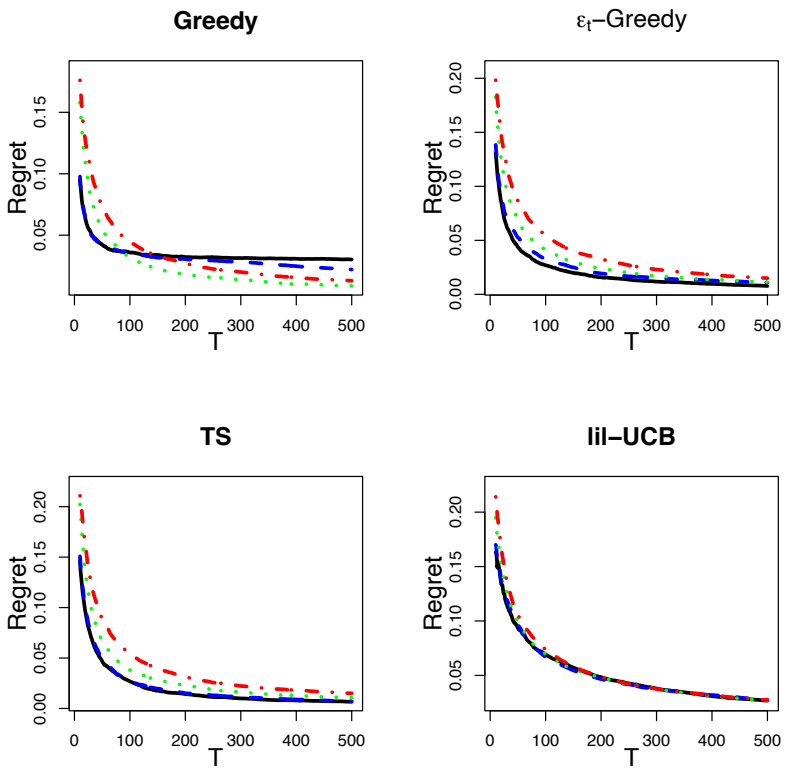

Figure 15: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 2-arm Bernoulli case with $T = 500$ and 1,000 replications. The true parameters of the arms are set as 0.8 and 0.2 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

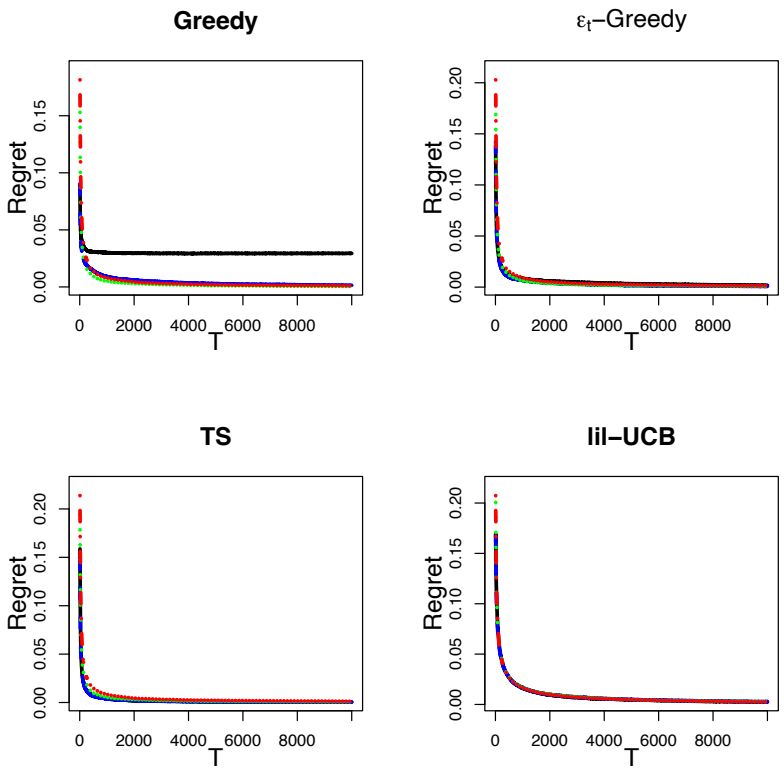

Figure 16: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 2-arm Bernoulli case with $T = 10,000$ and 1,000 replications. The true parameters of the arms are set as 0.8 and 0.2 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

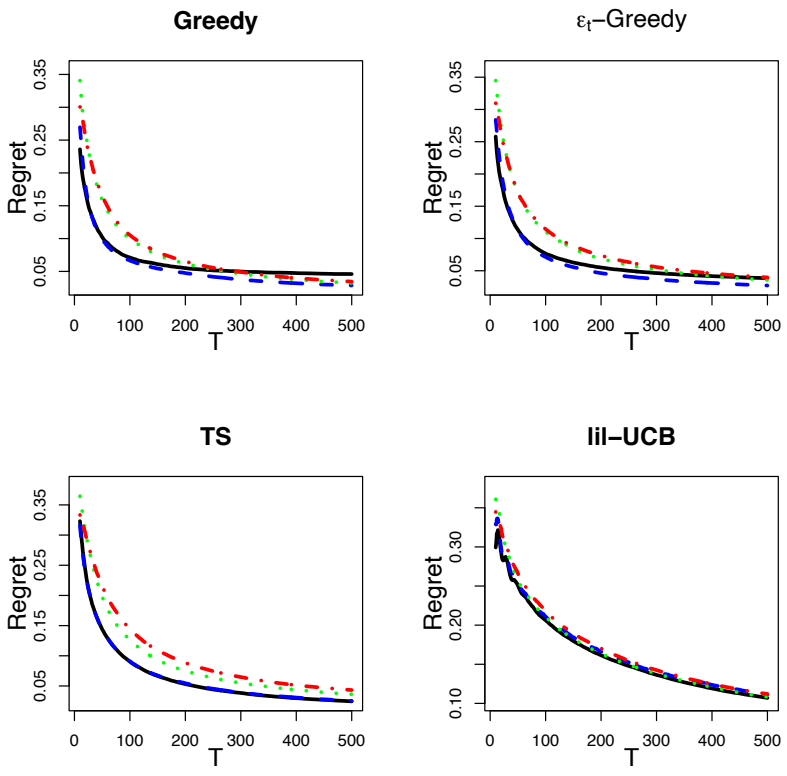

Figure 17: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 5-arm Bernoulli case with $T = 500$ and 1,000 replications. The true parameters of the arms are set as 0.9, 0.7, 0.5, 0.3, and 0.1 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

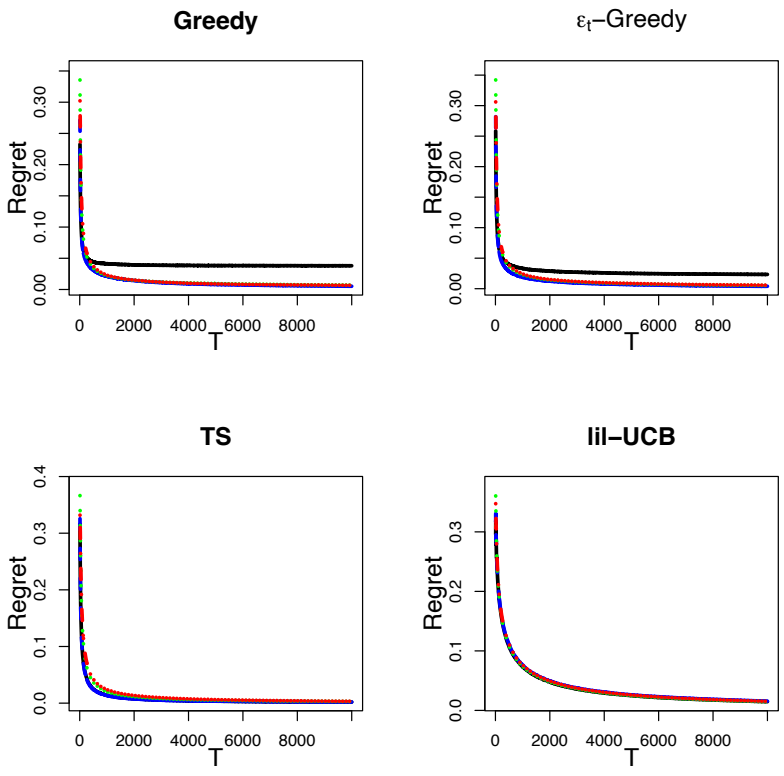

Figure 18: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 5-arm Bernoulli case with $T = 10,000$ and 1,000 replications. The true parameters of the arms are set as 0.9, 0.7, 0.5, 0.3, and 0.1 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

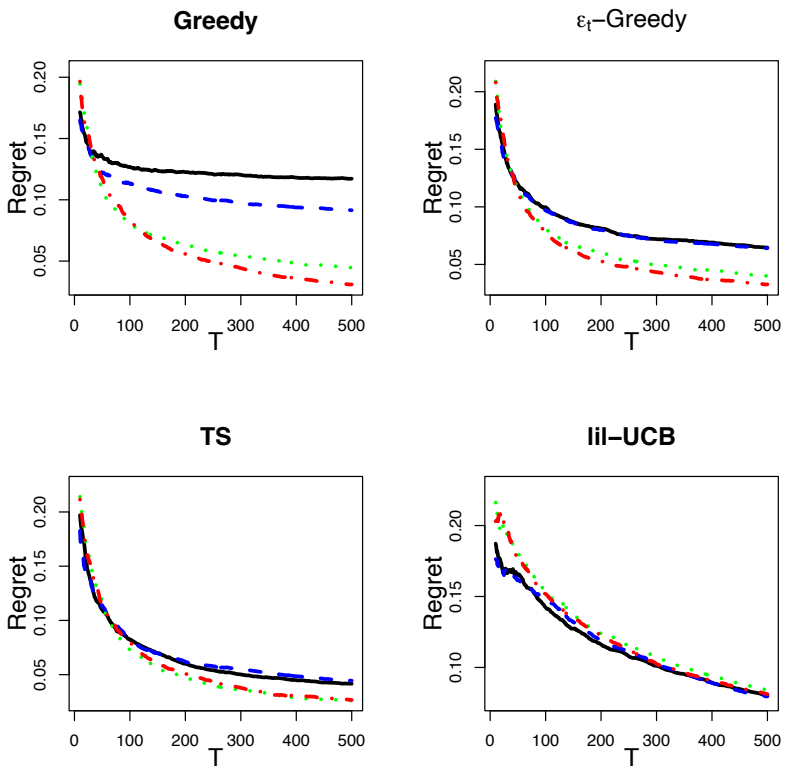

Figure 19: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 2-arm Gaussian case with $T = 500$ and 1,000 replications. The means of the five arms are set as 1 and 0.5 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

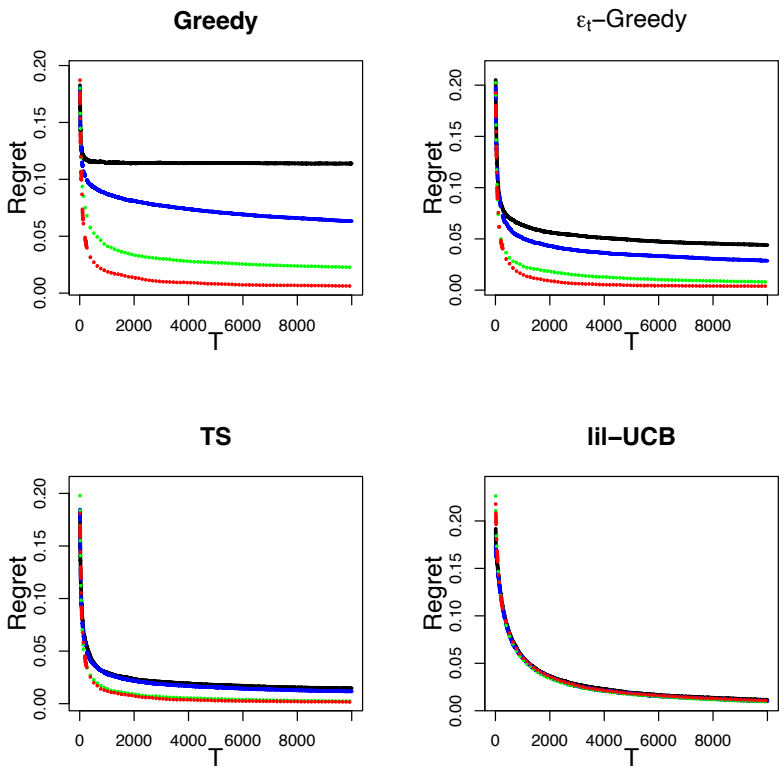

Figure 20: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 2-arm Gaussian case with $T = 10,000$ and 1,000 replications. The means of the five arms are set as 1 and 0.5 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

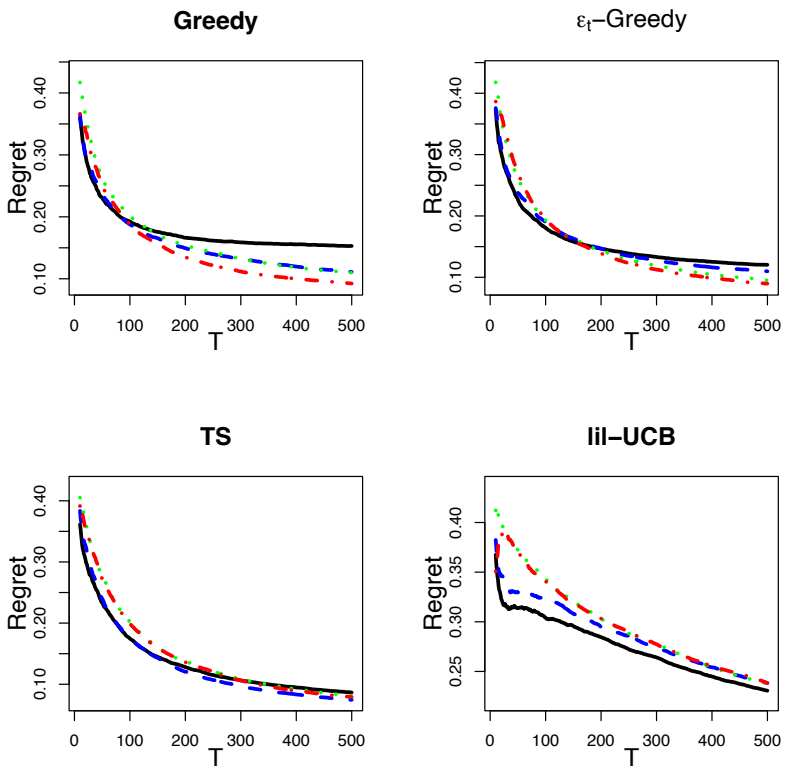

Figure 21: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 5-arm Gaussian case with $T = 500$ and 1,000 replications. The means of the five arms are set as 1, 0.75, 0.5, 0.38, and 0.25 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

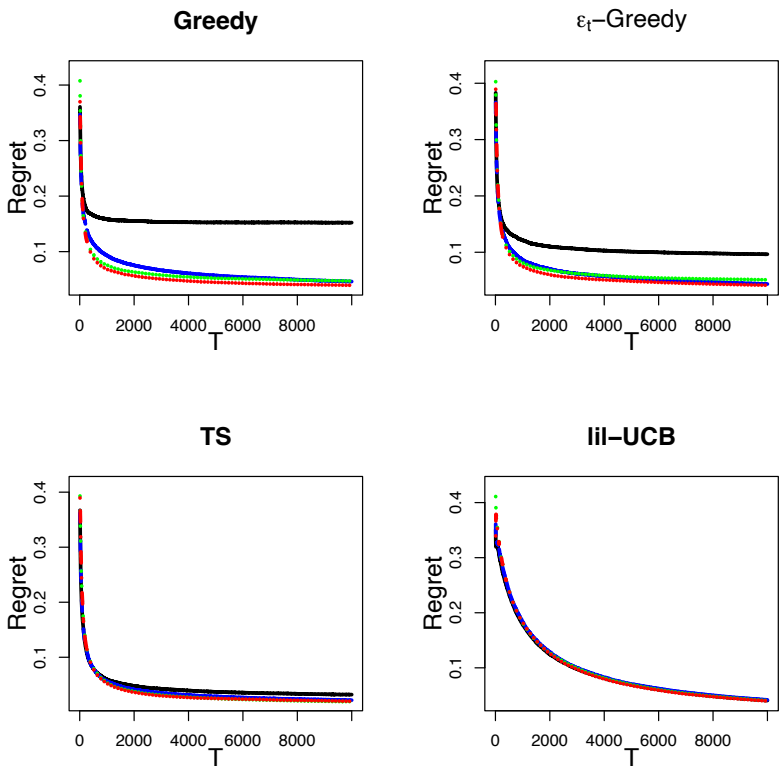

Figure 22: The regret of various methods (MAB, DP, rMAB(US), and rMAB(WF)) was evaluated for the 5-arm Gaussian case with $T = 10,000$ and 1,000 replications. The means of the five arms are set as 1, 0.75, 0.5, 0.38, and 0.25 respectively. There are four panels corresponding to four different MAB algorithms from top-left to bottom-right: greedy, ϵ_t-greedy, TS, and lil-UCB. In each panel, four curves represent MAB (Black solid), DP (blue dashed), rMAB(US) (green dotted), and rMAB(WF) (red, dotdash).

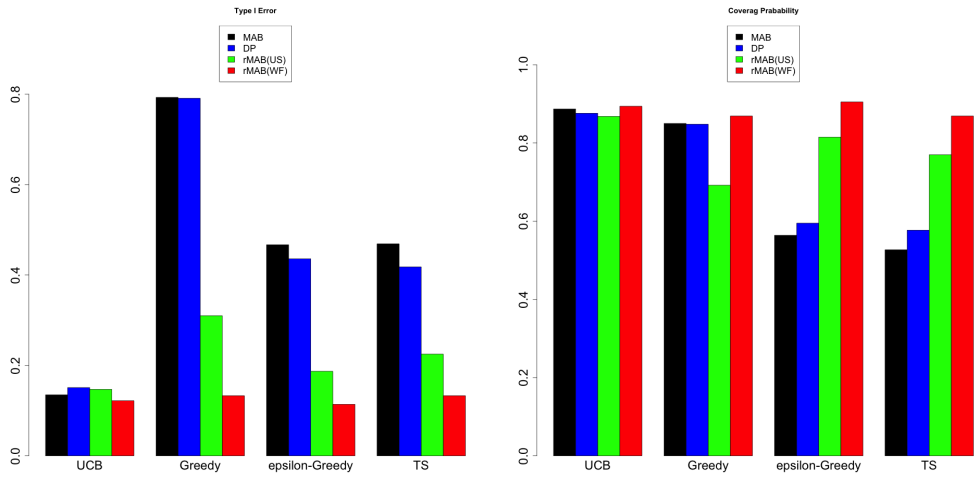

Figure 23: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $p_2 - p_1$ in the Bernoulli design with $K = 2$. We set T to 500 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

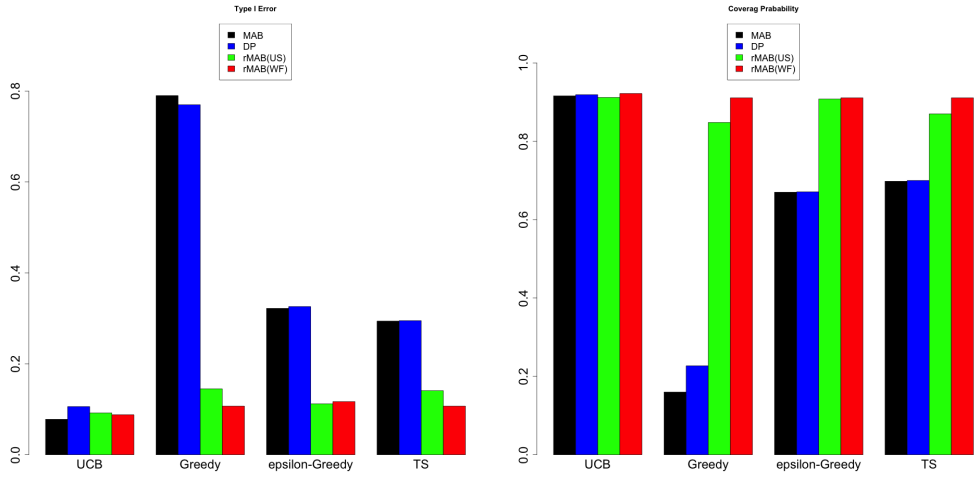

Figure 24: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $p_2 - p_1$ in the Bernoulli design with $K = 2$. We set T to 10,000 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

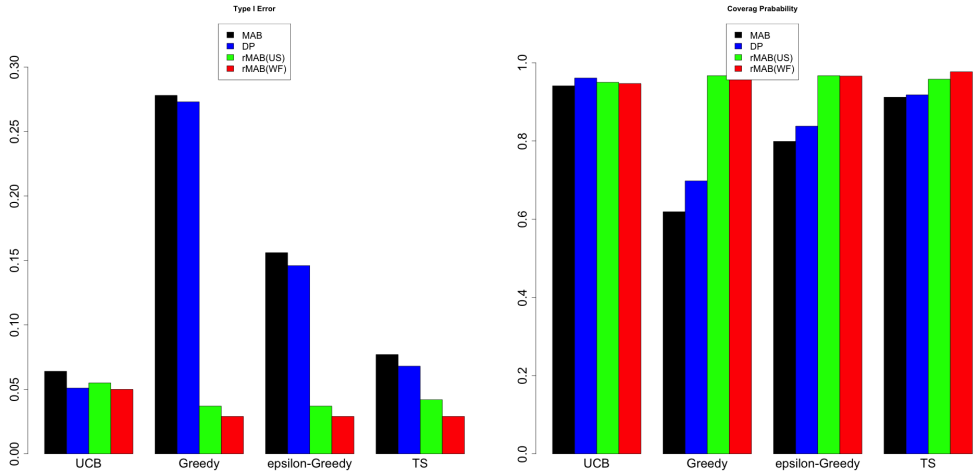

Figure 25: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $p_2 - p_1$ in the Bernoulli design with $K = 5$. We set T to 500 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

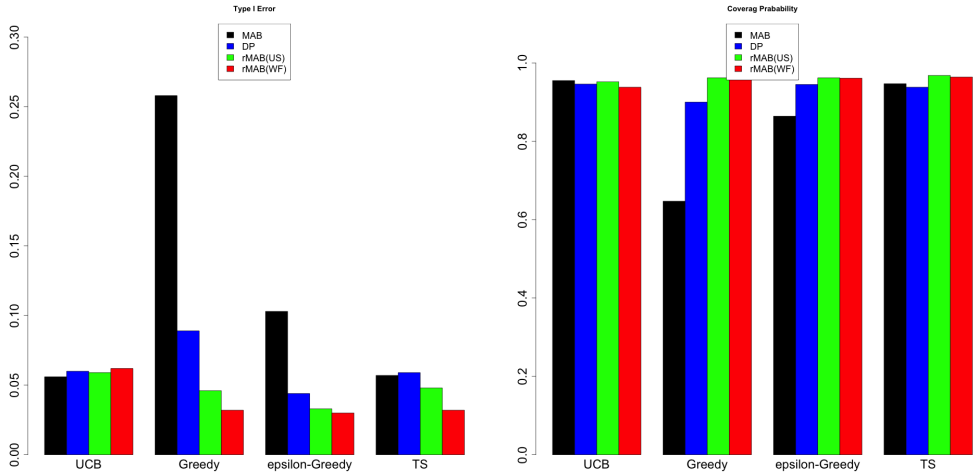

Figure 26: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $p_2 - p_1$ in the Bernoulli design with $K = 5$. We set T to 10,000 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

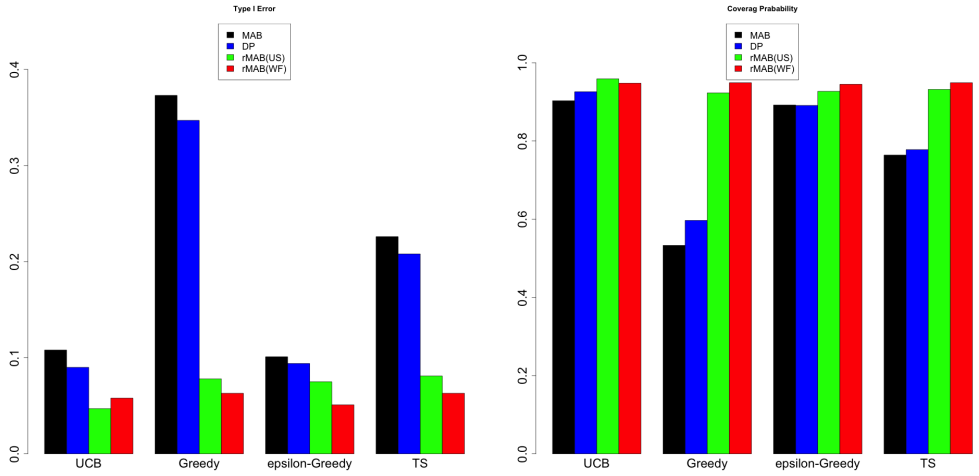

Figure 27: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $\mu_2 - \mu_1$ in the Gaussian design with $K = 2$. We set T to 500 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

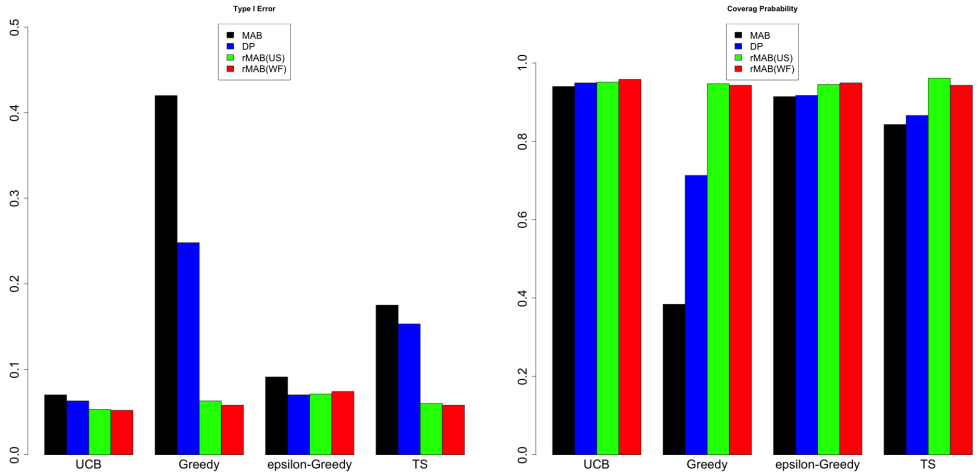

Figure 28: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $\mu_2 - \mu_1$ in the Gaussian design with $K = 2$. We set T to 10,000 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

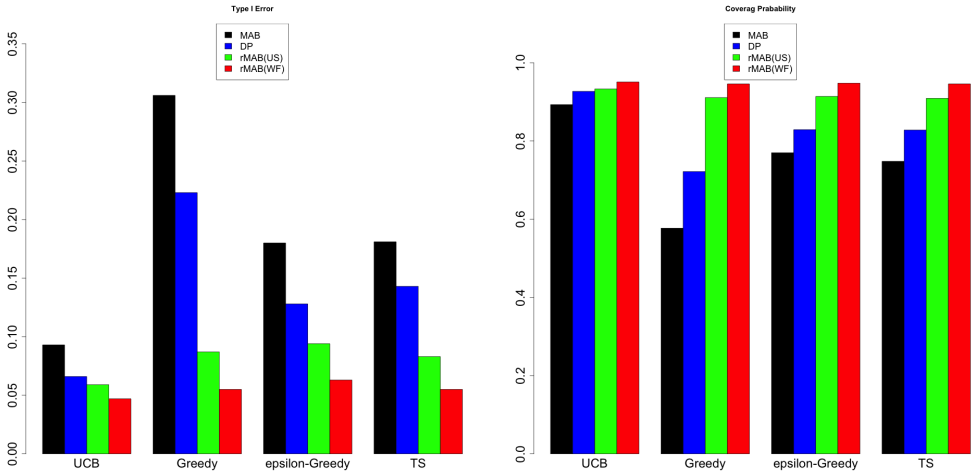

Figure 29: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $\mu_2 - \mu_1$ in the Gaussian design with $K = 5$. We set T to 500 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.

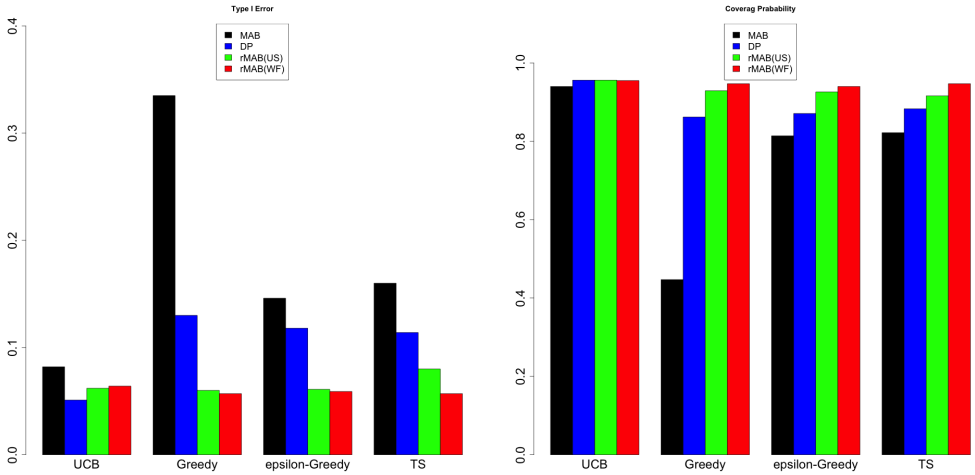

Figure 30: We compared the Type I Error rate (left panel) and coverage probability (right panel) for the parameter $\mu_2 - \mu_1$ in the Gaussian design with $K = 5$. We set T to 10,000 and conducted 1,000 replications. The targeted Type I error rate is 0.05, and the nominal coverage probability is 0.95.