# Multi-Armed Bandits with Inference Considerations

Sandeep Gangarapu
University of Minnesota

Edward McFowland III
University of Minnesota

Ravi Bapna
University of Minnesota

## Abstract

Multi-armed bandits (MAB) are sequential experimentation procedures that use a combination of exploration and exploitation techniques to reduce allocations to interventions with sub-optimal outcomes. MAB's are very effective in reducing the regret of the experimentation process compared to A/B testing, especially in the presence of multiple policy levers. However, unlike A/B testing, MAB's may fail to accurately estimate the parameters of treatment effect distributions of interventions. In many Marketing, Clinical Trials, and Public Policy settings, estimating the parameters of treatment effect distributions is as crucial as that of identifying the best intervention, e.g., feedback for intervention designers. In this paper, we propose a new MAB algorithm called UCB-INF that solves the above problem. We show that UCB-INF has regret comparable to the best MAB algorithms while having the parameter estimation properties of A/B testing.

Keywords: Multi-armed bandit, A/B testing, variance estimation

# Introduction

A/B testing is an experimentation strategy used to infer the effectiveness of an intervention compared to a baseline (control). Many scientific fields like Agriculture, Medicine, Social Sciences, Online Experimentation, etc. use A/B testing. Due to the proliferation of internet technologies, A/B testing is widely used on internet platforms like Amazon, Microsoft, Google, etc. (Kohavi et al., 2009). It is used in a variety of tasks, e.g., to test various design features of a website, test multiple marketing campaigns, etc. For a given tolerance of Type 1 and Type 2 errors, A/B testing is very effective in statistically rejecting a null hypothesis (i.e. the intervention has no additional benefit compared to the baseline). During an A/B test, units are allocated randomly to the baseline condition (control group) and the interventions (treatment groups). If the mean of the treatment outcomes is statistically different compared to the mean of control outcomes, we consider treatment to be effective. Typically, during the exploration phase of an A/B test, multiple variants of a feature/ad/promotion are tested, and the best performing option is selected to be exploited for the subsequent time period.

While A/B testing is effective in online experimentation, it has many challenges. In the presence of multiple interventions, there is a loss of utility (regret) in assigning subjects to sub-optimal conditions during the experimentation process (White, 2012). Multi-armed bandit(MAB) methods that use a combination of exploration and exploitation techniques are used to overcome this challenge. Unlike A/B testing, allocation procedure in MAB methods is not random. Instead, MAB's use dynamic allocation procedure based on outcome distribution at any given point of time. Multi-armed bandits derive their name from slot machines (one-armed bandits) at a casino. A gambler has multiple slot machines with unknown payoff distributions to choose. At each time period, the gambler plays an arm of a slot machine observes a payoff. The objective of the gambler is to maximize the total payoff obtained after a series of $n$ arm plays. This is equivalent to minimizing the regret i.e. the difference between the expected payoff obtained by playing the best arm $n$ times and the observed payoffs (Lattimore and Szepesvári, 2018).

Each MAB algorithm differs in its adaptive allocation procedure. For example, in Beta-Bernoulli Thompson sampling MAB (Thompson, 1933), the outcome distribution of each arm is initialized as beta distribution with random parameters. At each round, a sample is drawn from prior distribution and the arm that has the highest value of the draw is played. The Bernoulli payoff is observed, and a posterior distribution is estimated using a Bayesian update. This process is continued for $n$ arm plays. In the experimentation process, interventions are similar to arms of MAB, and outcomes are payoffs. The adaptive procedure reduces the regret of the experimentation process as allocations to sub-optimal interventions are considerably less. Schwartz et al. (2017) used Thompson Sampling based MAB algorithm and achieved an 8% increase in customer acquisition rate on an online display advertising campaign compared to A/B testing.

In experiments that involve a large number of interventions (e.g. testing multiple ad variants), MAB's are effective as optimizing for regret solves the problem of the total number of sub-optimal allocations. However, it is possible that some interventions receive so few allocations that inference about the mean and variance of the outcome distributions are highly inaccurate. This is not a problem in contexts where the only objective is to minimize regret and find the intervention that performs the best. However, in many settings, accurate estimation of mean and variance of outcome distributions of all interventions is critical as that knowledge is used to design future experiments and helps strengthen the domain knowledge of the practitioner. For example, in clinical trials, knowing the effect size and variance of the drug is crucial for scientists who developed the drug (Nie et al., 2017). This is also true for most public policy and high-cost marketing campaign settings.

In order to solve this problem, we propose a new MAB algorithm called "UCB-INF" where allocations are made not only to minimize the overall regret but also to accurately estimate the mean and variance outcome distributions of all interventions. UCB-INF has regret comparable to the best-in-class MAB's and also has mean and variances estimation properties of A/B testing. In the next section, we discuss the literature related to MAB's and A/B testing. We then formally define the MAB problem and provide the algorithm for UCB-INF. We then compare it to A/B testing, and popular MAB methods like Upper Confidence Bound (UCB) and Epsilon-Greedy ($\epsilon$-greedy). We then describe the simulation procedure and compare the results and properties of MAB-VAR with other allocation procedures.

# Related Literature

Multi-armed bandits are sequential experimentation problems with an exploration-exploitation trade-off. The key decision in an armed bandit problem is to whether stay with the intervention with the highest payoff based on observed data and exploit it or explore other interventions that might give higher payoffs in the future. Bandit problems have been studied since the 1930s and have application in several problem domains such as clinical trials, ad placement, website optimization, packet routing (Bubeck et al., 2012).

One of the early bandit applications was in a clinical trial setting where Thompson (1933) used adaptive allocation to dynamically assign subjects to different drugs based on the observed outcomes in order to minimize allocations to the drug that did not perform well. In Thompson sampling, prior outcome distributions for each intervention are initialized with random parameters. At each round, a sample is drawn from each of these distributions, and an allocation is made to the intervention that has the highest value of the draw. An outcome is observed, and posterior distribution of that intervention is estimated using Bayesian update. This process is repeated every round. $\epsilon$-greedy is another simple algorithm where exploration and exploitation are clearly specified (Kuleshov and Precup, 2014). At each round, the intervention with the best estimated mean outcome so far is selected with probability $\epsilon$ and a random intervention is selected with the empirical probability of $1 - \epsilon$. However, the constant $\epsilon$ prevents the algorithm from asymptotically converging to the best arm (Vermorel and Mohri, 2005). Upper confidence bound (UCB) is one of the most efficient MAB algorithms. In UCB algorithm, at each round, upper confidence bound of mean is calculated for each intervention using Chernoff-Hoeffding bound (Chernoff et al., 1952) and allocation is made to the intervention with the highest value of UCB. The intuition behind UCB is that if an allocation is made to the sub-optimal intervention, then a value drawn from that distribution reduces the upper confidence bound of the mean which in-turn makes it less probable for that intervention to have higher UCB in the future.

MAB's have many practical applications. Villar (2018) use MAB's in the context of clinical trials for rare, life-threatening diseases where the priority is given to patient over hypothesis testing. Schwartz et al. (2017) use MAB's to acquire customers by adaptively showing different types of ads on various websites. They report an 8% increase in conversion rate over traditional A/B testing. Misra et al. (2019) use a combination of MAB's and microeconomic choice theory in a dynamic pricing setting and show a 43% increase in profits. Most online experimentation platforms like Google Optimize, Optimizely, Facebook Ax, use MAB's for practical applications.

# Problem Formulation

Multi Armed Bandit problem is generally framed as a repeated game played between a learner and an environment. There are a total of $N$ rounds in the game where $N$ is called the horizon. In experimentation, this is equivalent to the total number of subjects available for allocation. There are $k$ arms available which is the total number of treatment groups or interventions including one control group. We assume the outcome distribution of each arm $a$ follows a Gaussian distribution $\mathcal{N}(\mu_a, \sigma_a^2)$. At each time $t$ an arm $a_t$ is played and the corresponding outcome $x_{a_t}$ is drawn from $\mathcal{N}(\mu_{a_t}, \sigma_{a_t}^2)$. The mean reward estimate of an arm $a$ is given by $\bar{x}_a$ an the variance estimate is given by $s_a^2$. The mean reward of the best arm is given by

$$\mu^* = \max_{a=1,2..k} \bar{x}_a$$

The empirical regret at any time period $t$ is given by

$$\mathcal{R}_t = \mu^*.t - \sum_{t=1}^{t} x_{a_t}$$

The root mean square error of mean estimate and variance estimate of all arms in given by

$$RMSE_{mt} = \sqrt{\sum_{a=1}^{k} \frac{(\mu_a - \bar{x}_a)^2}{k}},\ RMSE_{vt} = \sqrt{\sum_{a=1}^{k} \frac{(\sigma_a^2 - s_a^2)^2}{k}}$$

In general, the objective of most MAB algorithms is to minimize regret. The objective of UCB-INF is not only to minimize regret but also to estimate $\mu_a, \sigma_a^2$ accurately by reducing $RMSE_m, RMSE_v$.

## UCB-INF Algorithm

$$\bar{x} + \sqrt{\frac{2 * log(N)}{n}}$$

$$\frac{2s^4}{n-1}$$

UCB-INF is motivated by Upper Confidence Bound (UCB) and $\epsilon$-greedy algorithms. In $\epsilon$-greedy, an $\epsilon$ is chosen, and a constant exploration is done with a probability of $\epsilon$ and exploitation is done with probability $1 - \epsilon$. $\epsilon$-greedy has a regret bound of $\mathcal{O}(klogN)$ and is not the most efficient of algorithms. If an optimal $\epsilon$ is not chosen, the performance degrades over time (Bubeck et al., 2012). UCB algorithm is more efficient and has sub-linear regret bound of $\mathcal{O}(logN)$. However, there is no guarantee of significant exploration across all arms, so, the inference of parameters of outcome distributions may not be accurate. UCB-INF overcomes this by making variance-based allocations.

$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$$

$$Std.Err(\bar{x}) \propto \frac{\sigma^2}{n}$$

**Algorithm 1:** UCB-INF Algorithm

```
Inputs: Probability of variance based allocation ($\epsilon \in [0,1]$), No. of interventions ($k$),
  No. of subjects ($N$), Variance change tolerance ($\tau$)
Initialize $s_a^2 = 0$ (Variance estimate), ${s'_a}^2 = 0$ (Variance change estimate)
/* At the start, we make one allocation per arm. This is required to calculate Upper
  Confidence Bound*/
for $a = 1, 2, 3...k$ do
    Choose intervention $a$
    Observe reward $x_a$
    Calculate $s_a^2$, ${s'_a}^2$
end
for $a = k+1, k+2.....N$ do
    Draw from Uniform distribution $d \sim U(0,1)$
    if $d > \epsilon$ then
        for $a = 1, 2, 3...k$ do
            $UCB_a = \bar{x_a} + \sqrt{\frac{2*log(N)}{n_a}}$
        end
        u = $\text{argmax}_a UCB_a$
        Choose intervention $u$
        Observe payoff $x_u$
        Calculate $s_u^2$, ${s'_u}^2$
    else
        if $\max_a {s'_a}^2 < \tau$ then
            Continue
        else
            for $a = 1, 2, 3...k$ do
                Draw from $\tilde{s_a}^2 \sim \mathcal{N}(\frac{s_a^2}{n-1}, \frac{s_a^4}{n-1})$
            end
```

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$Std.Err(s^2) \propto \frac{2\sigma^4}{n-1}$$

Similar to $\epsilon$-greedy, at each round, UCB-INF makes allocations according to UCB algorithm with a probability of $1-\epsilon$ and variance-based allocation with a probability of $\epsilon$. $\epsilon$ can be treated as a hyper-parameter and can be changed by the user.

Variance-based allocation is defined as follows. For a Gaussian outcome distribution, the sample mean, $\bar{x}$ is drawn from the distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ and sample variance, $s^2$ is drawn from $\mathcal{N}(\frac{\sigma^2}{n}, \frac{\sigma^4}{n})$. This means that the variance of variance estimate is proportional to the square of true variance and inversely proportional to the sample size. Variance-based allocation uses the principle of Thompson sampling. The prior distribution here is the distribution of sample variance $s^2 \sim \mathcal{N}(\frac{\sigma^2}{n}, \frac{\sigma^4}{n})$. As we do not know the true population variance, we use sample variance $s^2$ in order to estimate sample variance distribution $s^2 \sim \mathcal{N}(\frac{s^2}{n-1}, \frac{s^4}{n-1})$. For each arm, a draw is made from sample variance distribution of that arm and the arm with the highest value of draw is made an allocation. A reward $x_{a_t}$ is observed and is used to estimate the posterior distribution. This process constitutes the variance-based allocation.

There is another hyper-parameter $\tau$ called variance change tolerance which dictates the stopping point of variance-based allocation. After each round, we track the change in the variance estimate of an arm after observing the reward. If the maximum variance change across all arms is less than variance change tolerance, the allocation shifts to UCB. This process makes it more efficient.

The intuition behind UCB-INF is to make the most of the efficiency provided by UCB while using selective exploration for estimating mean and variance of the arms effectively. The algorithm 1 shows the step by step procedure for UCB-INF.

# Simulation and Results

## Simulation setup

We ran simulations and compared the performances of A/B testing, UCB, $\epsilon$-greedy and UCB-INF in order to compare their properties. We expect UCB-INF to have regret minimization properties of UCB and variance estimation properties of A/B testing. The experimental setup has one control condition and nine other treatment conditions. We assume the outcome distribution to be Gaussian. We set the control mean at 0 and randomly draw treatment group means from a uniform distribution between 0 and 5. Similarly, we draw all outcome variances from a uniform distribution of 0 and 5. We chose 2000 subjects available for allocation and sequentially allocate the subjects based on algorithmic suggestions. If an algorithm suggests group 3 to be allocated, we draw from $\mathcal{N}(\mu_3, \sigma_3^2)$ and set it as the payoff. For A/B testing, the sample size is decided based on the smallest effect size of all arms. This is because the sample size in A/B testing is the same across all arms and should have sufficient power to detect the minimum effect size. We chose a significance level of 0.05 and a power of 0.8. Exploration is done until each group is uniformly allocated with the number of subjects decided from power analysis with the above parameters. After exploration, the intervention with the highest mean of the observed outcome is allocated the remaining subjects. Sample size need not be set for MAB's as they are sequential allocation algorithms. For both $\epsilon$-greedy and UCB-INF, the $\epsilon$ i.e. probability of exploration is set at 0.1. The variance change tolerance for UCB-INF is set at 0.05.

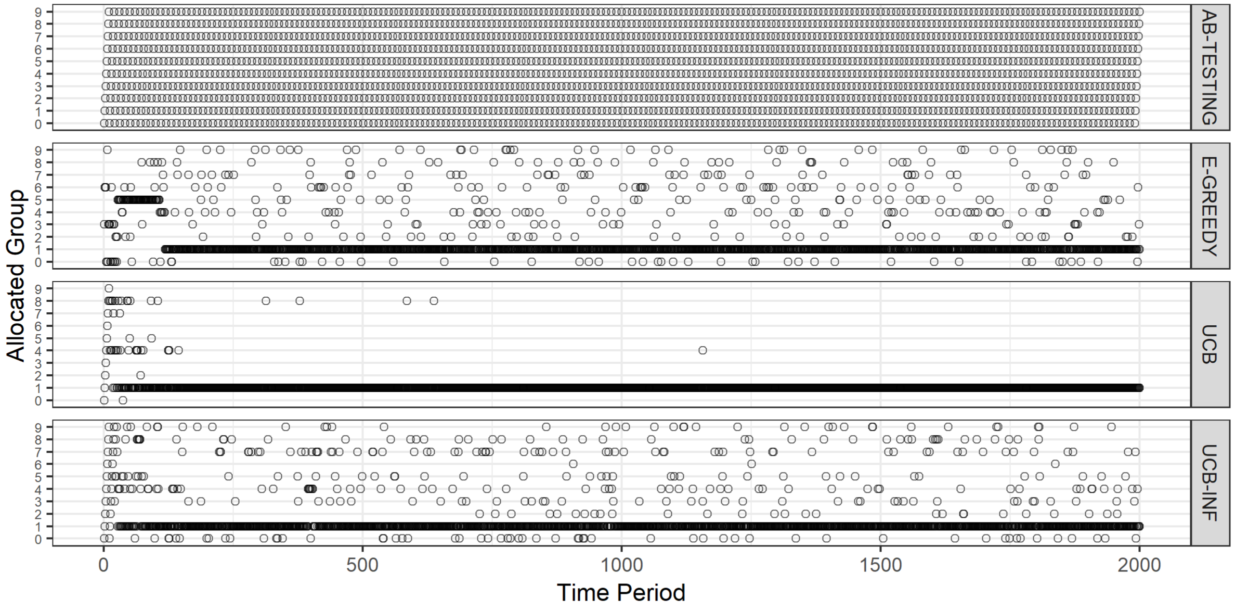

Figure 1: This figure shows the allocations of subjects to treatment groups over time for different allocation procedures

## Results

We ran the simulation for A/B testing, UCB, $\epsilon$-greedy and UCB-INF by allocating subjects to different treatment conditions as per algorithmic suggestions and drawing a value from the outcome distribution of allocated intervention. Fig:1 shows how subjects are allocated to different interventions by different algorithms. It can be seen that all groups in A/B testing are sequentially allocated until the smallest effect is detected with sufficient power. Once the exploration is done, the rest of the subjects are allocated to group 5 as it has the highest

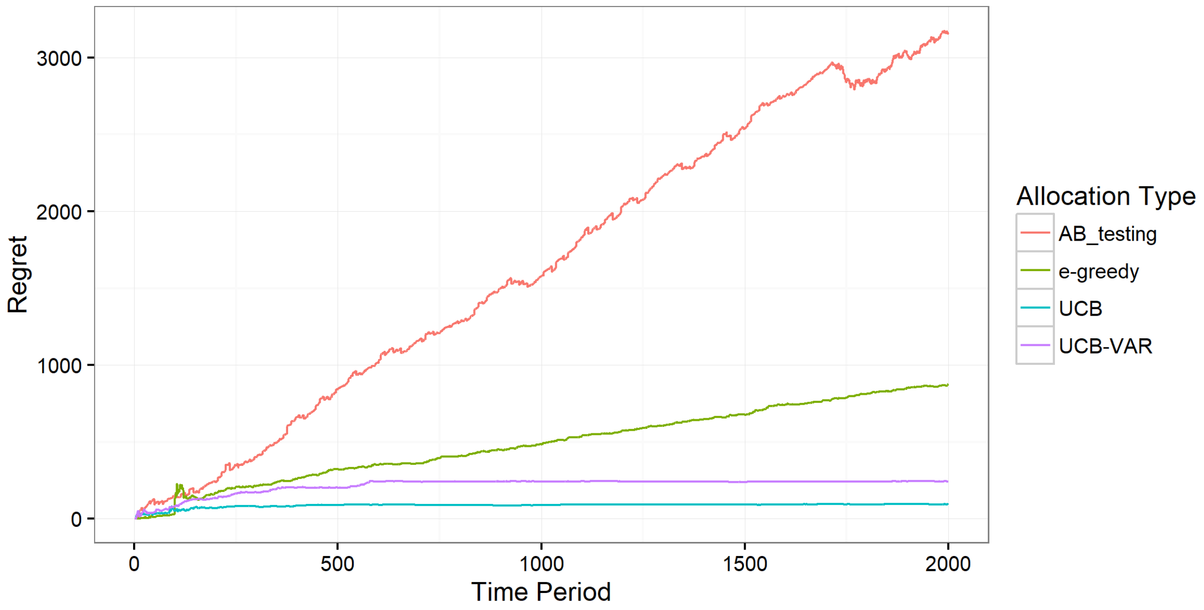

Figure 2: This figure shows the regret of different allocation procedures over time

mean outcome. Allocations in UCB are much more efficient and they swiftly converge to group 5. However, most treatment groups barely receive any allocations. For example, group 0 only has two subjects allocated and group 3 only has one subject allocated. Owing to this, the estimation of mean and variance is inaccurate. This can be seen from Fig:3 where UCB has the highest RMSE of variance estimation and does not reduce even after all subjects are allocated. This is because the primary objective of UCB is to reduce regret and has one of the best finite time regret bounds among MAB algorithms. Fig:4 shows the variance estimate provided by different allocation procedures at different time periods. The black solid line represents the true variances drawn from the uniform distribution of 0 and 5.

Both $\epsilon$-greedy and UCB-INF converge to the best intervention, however, exploration in $\epsilon$-greedy is random while exploration in UCB-INF is selective. In UCB-INF, allocations are made to those interventions where there is high uncertainty in the variance estimate.

Also, the variance change tolerance parameter helps in further reducing the regret. Once there is certainty in the variance estimate, UCB-INF switches to UCB more frequently. This can be seen in Fig:2 where after initial exploration, the regret line of UCB and UCB-INF stay parallel. In contrast, the regret line of $\epsilon$-greedy keeps increasing as more allocations are made. The variance estimation properties of A/B testing, UCB-INF and $\epsilon$-greedy are equally good because of continuous exploration across all arms by A/B testing and $\epsilon$-greedy and selective exploration by UCB-INF. Fig:4 shows that, for all arms, the estimate of the variance converges to true variance for every algorithm except UCB. This shows that the variance estimation properties of UCB-INF are as good as A/B testing.

Overall, we see that UCB-INF does sacrifice the overall utility but the regret is still comparable to that of UCB which has one of the best regret bounds among MAB algorithms. At the same time, the variance estimation of UCB-INF is as good as A/B testing.

## Conclusion and Future Work

In this paper, we present a new MAB algorithm called UCB-INF, where the objective of the allocation procedure is not only to minimize regret but also to estimate the mean and variance of outcome distributions accurately. We conducted simulations and compared the performance of UCB-INF with A/B testing and other MAB algorithms like UCB and $\epsilon$-greedy. We show that UCB-INF has regret comparable to that of UCB algorithm while estimation of mean and variance of outcome distributions is comparable to that of A/B testing. This algorithm is very useful in multiple domains like online experimentation, marketing, clinical trials, and public policy. It is especially effective in scenarios where there

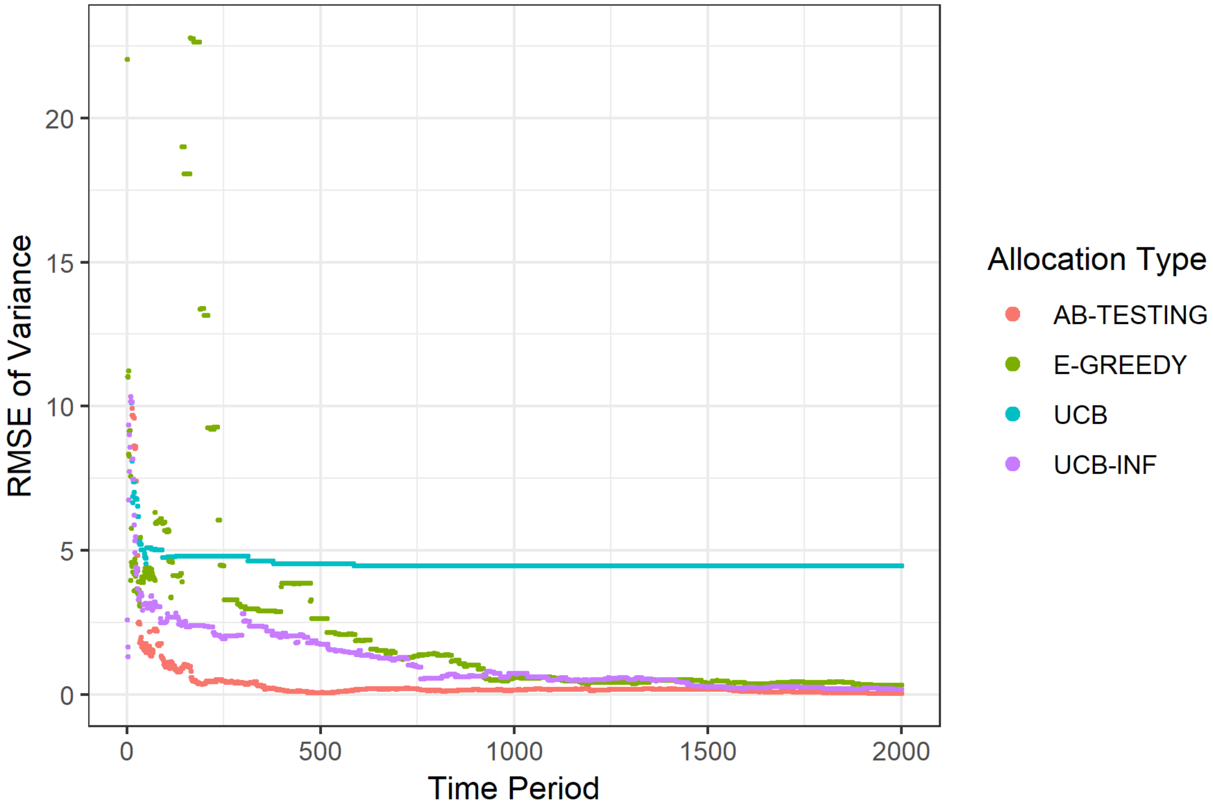

Figure 3: This figure shows the root mean squared error of variance estimation by different allocation procedures over time

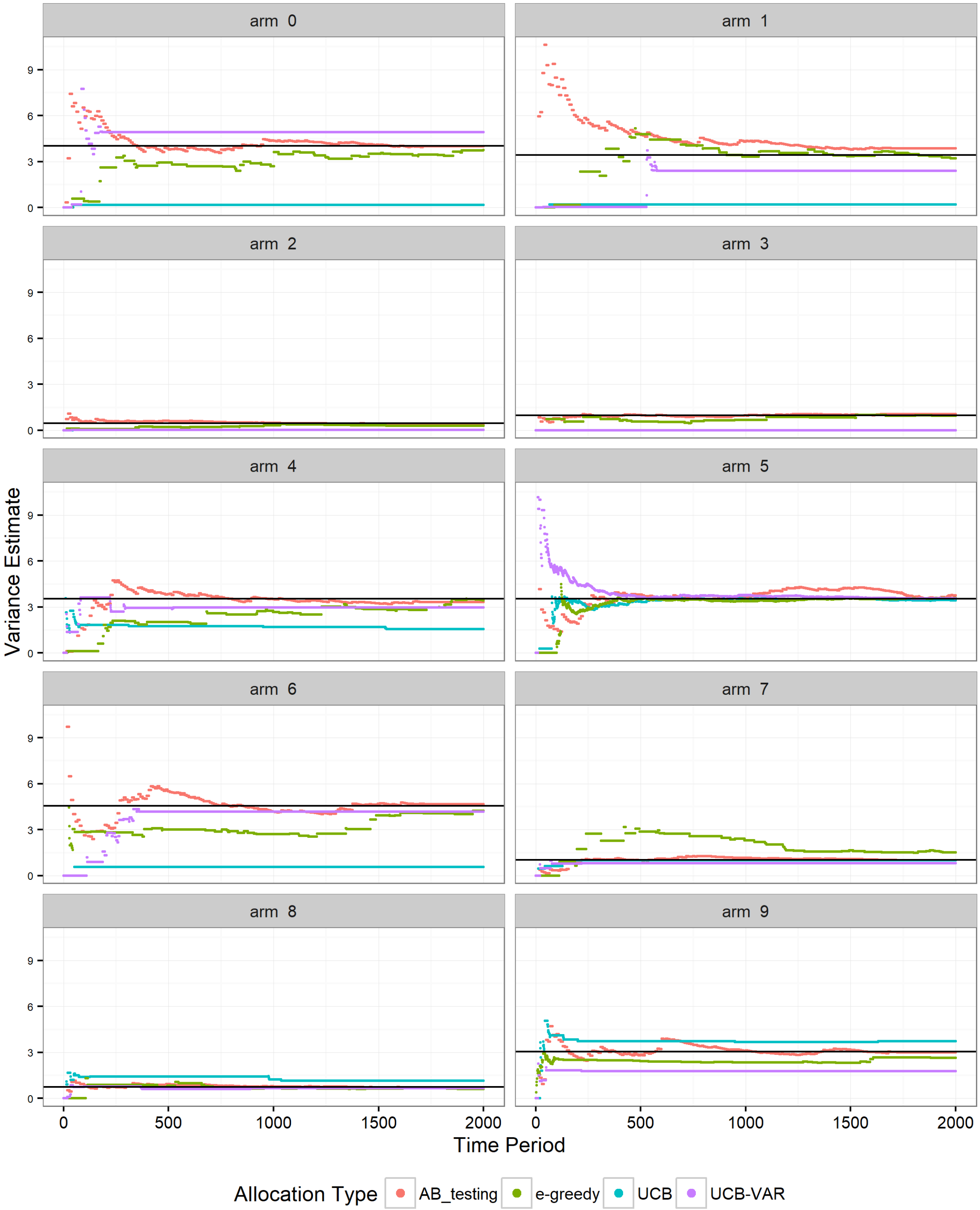

Figure 4: This figure shows the variance estimate of different algorithms for multiple interventions

are a limited number of subjects available for allocation and the utility maximization is as important as estimating outcome distributions accurately.

In future, we want to use an algorithm similar to UCB-INF and investigate how the allocations change by optimizing for treatment effect distributions instead of outcome distribution. We also want to derive the asymptotic and finite time regret and variance bounds using statistical theory for UCB-INF.

# References

Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Chernoff, H. et al. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.

Kuleshov, V. and Precup, D. (2014). Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*.

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. *preprint*.

Misra, K., Schwartz, E. M., and Abernethy, J. (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252.

Nie, X., Tian, X., Taylor, J., and Zou, J. (2017). Why adaptively collected data have negative bias and how to correct for it. *arXiv preprint arXiv:1708.01977*.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522.

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

Vermorel, J. and Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer.

Villar, S. S. (2018). Bandit strategies evaluated in the context of clinical trials in rare life-threatening diseases. *Probability in the engineering and informational sciences*, 32(2):229–245.

White, J. (2012). *Bandit algorithms for website optimization.* " O'Reilly Media, Inc.".