Summary Document: Inference for Batched Bandits

Kelly W. Zhang

1 How can bandit data be used for scientific discovery?

Bandit algorithms are strategies for regret minimization in sequential decision making problems. The regret of a bandit algorithm is how much worse it performs in terms of average cumulative reward compared to an oracle algorithm. Due to their regret minimizing properties, bandit algorithms are increasingly used in real-world sequential decision making problems, like online advertising, mobile health, and online education.

However, suppose I have run my bandit algorithm. Now from the resulting data can I infer...

- Is one bandit arm better than another in terms of expected reward?
- What is the *magnitude* of the difference in expected rewards of two bandit arms (treatment effect)?
- How can we have guarantees in answering the above questions?

Answering these questions are crucial for scientific discovery and important for industrial applications.

To answer these questions, we turn to statistical inference methods, i.e., hypothesis testing and constructing confidence intervals. In order to perform statistical inference we need an estimator of the treatment effect (difference in expected reward between two arms) that has an asymptotic distribution that well approximates it's the finite sample distribution.

2 Batched bandit setting

We focus on the setting in which bandit arms are selected in "batches". For many real world problems, it is not realistic to expect experiments to have a large number of stationary time periods in which the expected rewards for arms do not change at all over time. For example, in online advertising the effectiveness of an ad may change over time due to people seeing the ad before and general changes in society. Since non-stationarity is so prevalent in real world problems, asymptotic results for statistical estimators that rely on the number of stationary time periods T of a bandit experiment going to infinity are less applicable.

Often it is more realistic to assume a fixed number of time periods of a study T, and analyze asymptotics as the number of arm selections in each batch, n, go to infinity. For example, this corresponds to online advertising problems in which ads are sent out to many users simultaneously. Specifically, in the batched bandit setting we assume that the bandit algorithm is updated a total of T times and there are nT arm pulls in the experiment total.

Though our results extend to K-arm contextual bandits, we focus on the two-arm bandit setting throughout this document. We now introduce some notation. For each $t \in [1\colon T]$,

- **Action selection probability:** $\pi_t \in [0, 1]$
- **Actions:** $\{A_{t,i}\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi_t)$ where π_t is a function of history H_{t-1}
- **Rewards:** $\{R_{t,i}\}_{i=1}^n$
- **History:** $H_t = \cup_{t'<t}\{A_{t',i}, R_{t',i}\}_{i=1}^n$
- **Mean rewards for arms:** $\beta_{0,t}, \beta_{1,t}$ (in the stationary case $\beta_0 := \beta_{0,t} = \beta_{0,t'}$, $\beta_1 := \beta_{1,t} = \beta_{1,t'}$ for all $t \in [1\colon T]$)
- **Treatment effect:** $\Delta_t = \beta_{1,t} - \beta_{0,t}$ (in the stationary case $\Delta := \Delta_t = \Delta_{t'}$, for all $t \in [1\colon T]$)
- **Model of expected rewards:** $\mathbb{E}[R_{t,i}|H_{t-1}, A_{t,i}] = \beta_{1,t}A_{t,i} + \beta_{0,t}(1 - A_{t,i})$
- **Residuals:** $\epsilon_{t,i} = R_{t,i} - \beta_{1,t}A_{t,i} + \beta_{0,t}(1 - A_{t,i})$

3 Why not use standard statistical estimators on bandit data?

3.1 Bandit algorithms induce dependence

For data collected with bandit algorithms, action and reward tuples $(A_{t,i}, R_{t,i})$ for all $t \in [1\colon T], i \in [1\colon n]$ are **not** independent. This is because which actions are chosen, $A_{t,i}$, can depend on the history of past rewards and actions, $H_{t-1} = \cup_{t'<t-1}\{A_{t',i}, R_{t',i}\}_{i=1}^{n}$.

The induced dependence between action and reward tuples that occurs when choosing actions with bandit algorithms, violates the independence assumptions necessary for asymptotic results for many standard statistical estimators.

Several works in the last few years have shown that *standard statistical estimators that are unbiased on i.i.d. data can be biased on bandit data.* For example, on bandit data the sample mean is a biased estimator of the expected reward for a given arm [Shin et al., 2019, Nie et al., 2018]. There also have been several works trying to address this problem – estimators that reduce bias, as well as work that changes sampling algorithms to reduce bias [Deshpande et al., 2018, Dimakopoulou et al., 2019].

3.2 Asymptotic distribution of OLS estimator (sample mean) on bandit data

Lai & Wei, 1982 prove that under certain conditions, the ordinary least squares (OLS) estimator is asymptotically normal on *adaptively collected data. Adaptively collected data* refers to data in which action selections can depend on the past history of action selections and rewards, and encompasses all data collected using bandit and reinforcement learning algorithms. We prove that for two arm bandits, in the stationary case, the asymptotic normality conditions for the OLS estimator are satisfied under the following conditions:

1. Residuals $\epsilon_{t,i}$ satisfy conditional moment conditions (e.g. finite variance)
2. Action selection probabilities are constrained, so $\pi_t \in [\pi_{\min}, \pi_{\max}]$ for constants $0 < \pi_{\min} \leq \pi_{\max} < 1$
3. **The treatment effect is non-zero ($\Delta \neq 0$)**

We also prove that even with conditions 1 and 2 above, **when condition 3 is not satisfied the OLS estimator (sample mean) is asymptotically non-normal for data collected using typical bandit algorithms.** Specifically, we prove the asymptotic non-normality result for when the treatment effect is zero under Thompson Sampling and ϵ-greedy.

$$\hat{\Delta}^{\text{OLS}} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sum_{t=1}^{T}\sum_{i=1}^{n} A_{t,i}} - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sum_{t=1}^{T}\sum_{i=1}^{n}(1-A_{t,i})}$$

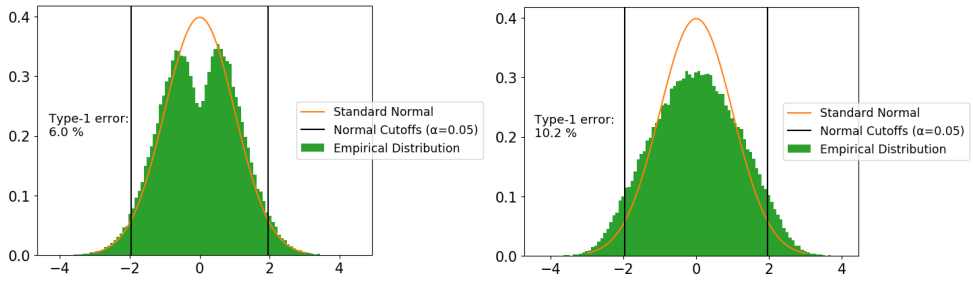

Figure 1: We plot the Z-statistic for $\hat{\Delta}^{\text{OLS}}$ under ϵ-greedy (left) and Thompson Sampling (right) under $\Delta = 0$. Note that the Z-statistic for $\hat{\Delta}^{\text{OLS}}$ is asymptotically normal when the data is collected i.i.d.

Suppose we would like to test whether the treatment effect is zero:

$$H_0 : \Delta = 0 \qquad \text{vs.} \qquad H_1 : \Delta \neq 0$$

Recall that in hypothesis testing, we set a significance level α (e.g. 0.05) which is the probability that we reject the null hypothesis when the null is true. We use the asymptotic distribution of the estimator under the null hypothesis to compute a critical value based on the significance level α, which determines what is considered an "extreme value" of the test statistic that occurs with less than probability α when the null is true. Observing a test statistic that is greater than the critical value would warrant us to reject the null hypothesis.

The above figure demonstrates that using the OLS estimator for hypothesis testing and assuming asymptotic normality would lead to *type-1 error inflation*, which means we would make a false discovery (reject the null hypothesis when the null is true) with probability greater than our significance level α.

4 Batched Ordinary Least Squares (OLS) Estimator

4.1 Introduction

Given that standard estimators like OLS (sample mean) are asymptotically non-normal when the treatment effect is zero, which prevents us from using them for hypothesis testing and constructing confidence intervals, we develop an alternative estimator that *is* asymptotically normal on bandit data for all values of Δ, even when the treatment effect is zero, and thus can be used for hypothesis testing and constructing confidence intervals.

Idea: Compute OLS estimator on each batch separately. Construct Z-statistic for each batch and show multivariate normality.

Standard OLS Estimator:

$$\hat{\Delta}^{\text{OLS}} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sum_{t=1}^{T}\sum_{i=1}^{n} A_{t,i}} - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sum_{t=1}^{T}\sum_{i=1}^{n}(1-A_{t,i})}$$

Z-statistic for $\hat{\Delta}^{\text{OLS}}$ testing $H_0 : \Delta = c$ vs. $H_1 : \Delta \neq c$:

$$\sqrt{\frac{(\sum_{t=1}^{T}\sum_{i=1}^{n} A_{t,i})(\sum_{t=1}^{T}\sum_{i=1}^{n}(1-A_{t,i}))}{nT}}(\hat{\Delta}^{\text{OLS}} - c)$$

Batched OLS Estimator: For each batch $t \in [1\colon T]$ we have the BOLS estimator:

$$\hat{\Delta}_t^{\text{BOLS}} = \frac{\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sum_{i=1}^{n} A_{t,i}} - \frac{\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sum_{i=1}^{n}(1-A_{t,i})}$$

Test statistic for BOLS estimator when testing $H_0 : \Delta = c$ vs. $H_1 : \Delta \neq c$:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{\frac{(\sum_{i=1}^{n} A_{t,i})(\sum_{i=1}^{n}(1-A_{t,i}))}{n}}(\hat{\Delta}_t^{\text{BOLS}} - c)$$

We prove that for data collected with standard bandit algorithms, the BOLS estimator will be asymptotically normal even when the treatment effect $\Delta = 0$. Moreover, BOLS is robust to non-stationary in the baseline reward, i.e., when Δ is the same for all batches but can $\beta_{t,0}$ and $\beta_{t,1}$ change over time. This kind of non-stationary might occur if we believe that one ad is more effective than another, but both ads become less effective over time.

See our full paper for simulation results and comparisons to other estimators.

4.2 Why is BOLS asymptotically normal while OLS is asymptotically non-normal when the treatment effect is zero?

We prove that the OLS estimator is asymptotically non-normal when the treatment effect is zero because the action selection probability π_t *does not concentrate* when the treatment effect is zero. When the treatment effect is non-zero, since one arm has greater expected reward than the other, under typical bandit algorithms, eventually π_t will converge to either 0 or 1. However, when the treatment effect is zero, π_t can fluctuate between 0 and 1 indefinitely, which leads to the asymptotic non-normality of the OLS estimator.

The key to proving asymptotic normality for BOLS is that the following ratio converges in probability to one:

$$\frac{(\sum_{i=1}^{n} A_{t,i})(\sum_{i=1}^{n}(1-A_{t,i}))}{n}\frac{1}{n\pi_t(1-\pi_t)} \xrightarrow{P} 1.$$

Since π_t is a function of H_{t-1}, thus $\frac{1}{n\pi_t(1-\pi_t)}$ is a constant given H_{t-1}. Thus, even if π_t does not concentrate, we are still able to apply the martingale central limit theorem to prove asymptotic normality.

References

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1194–1203, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453, 2019.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*, 2019.

Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. *International Conference on Artificial Intelligence and Statistics*, 2018.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems*, pages 7100–7109, 2019.