Inference for Batched Bandits

Kelly W. Zhang, Lucas Janson, and Susan A. Murphy

Harvard University

Table of contents

Introduction to inference on bandit data

What are bandits?

Bandit algorithms are strategies for regret minimization in sequential decision making problems.

	t=1	t=2	t=3	t=4	t=5
Treatment arm 0	0.2			0.3	?
Treatment arm 1		0.4	0.3		?

What are bandits?

Bandit algorithms are strategies for regret minimization in sequential decision making problems.

	t=1	t=2	t=3	t=4	t=5
Treatment arm 0	0.2			0.3	?
Treatment arm 1		0.4	0.3		?

- The regret of an bandit algorithm is how much worse it performs in terms of average cumulative reward compared to an oracle algorithm.

What are bandits?

Bandit algorithms are strategies for regret minimization in sequential decision making problems.

	t=1	t=2	t=3	t=4	t=5
Treatment arm 0	0.2			0.3	?
Treatment arm 1		0.4	0.3		?

- The regret of an bandit algorithm is how much worse it performs in terms of average cumulative reward compared to an oracle algorithm.
- Bandit literature primarily focused on developing algorithms that will minimize regret.

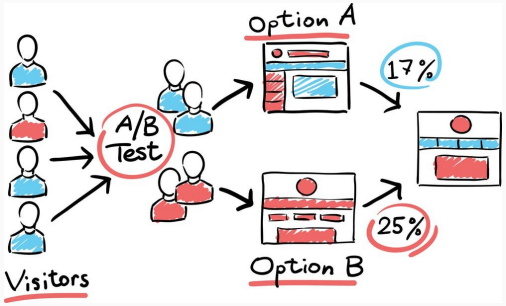
Bandits in the real world

Taking advantage of regret minimization

Bandits in the real world

Taking advantage of regret minimization

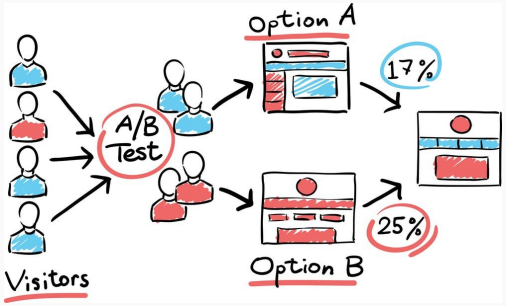
- Advertisements
  - Learn to show ads that are more interesting or relevant to users

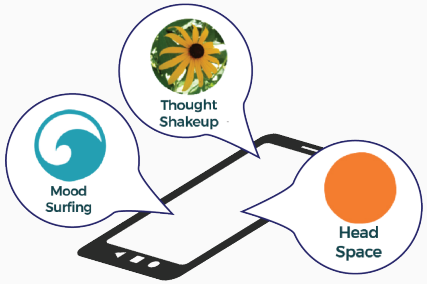


Bandits in the real world

Taking advantage of regret minimization

- Advertisements
  - Learn to show ads that are more interesting or relevant to users
- Mobile health
  - Learn when to send suggestions to users to best help them engage in healthy behaviors

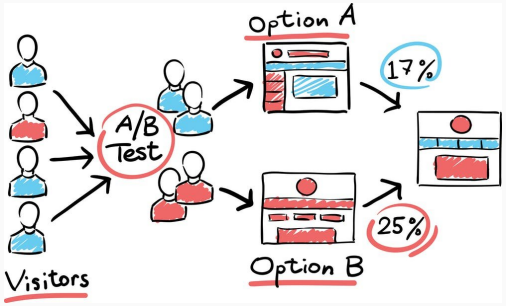


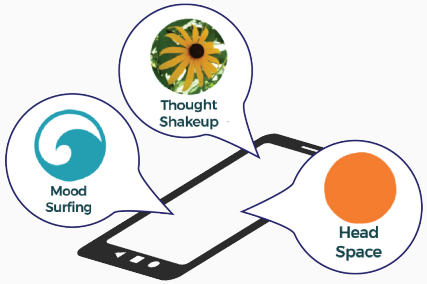


Bandits in the real world

Taking advantage of regret minimization

- Advertisements
  - Learn to show ads that are more interesting or relevant to users
- Mobile health
  - Learn when to send suggestions to users to best help them engage in healthy behaviors
- Online education
  - Learn to use more effective teaching strategies





Need for uncertainty quantification!

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

- Is one treatment arm better than another?

Need for uncertainty quantification!

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

- Is one treatment arm better than another?
- What is the *magnitude* of the difference in effectiveness of given treatments?

Need for uncertainty quantification!

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

- Is one treatment arm better than another?
- What is the *magnitude* of the difference in effectiveness of given treatments?

Note that bandit algorithms themselves do not give us any way to answer these questions.

Need for uncertainty quantification!

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

- Is one treatment arm better than another?
- What is the *magnitude* of the difference in effectiveness of given treatments?

Note that bandit algorithms themselves do not give us any way to answer these questions.

Regret minimization vs. Uncertainty quantification

Importance of Estimating Treatment Effect

Suppose we ran an online education experiment using a bandit algorithm to test different teaching strategies.

If we could infer the treatment effect between bandit arms...

Importance of Estimating Treatment Effect

Suppose we ran an online education experiment using a bandit algorithm to test different teaching strategies.

If we could infer the treatment effect between bandit arms...

- When designing a new online course...
  - Under-performing arms could be eliminated or modified
  - High-performing arms could be studied further

Importance of Estimating Treatment Effect

Suppose we ran an online education experiment using a bandit algorithm to test different teaching strategies.

If we could infer the treatment effect between bandit arms...

- When designing a new online course...
  - Under-performing arms could be eliminated or modified
  - High-performing arms could be studied further
- Could potentially publish findings, e.g., that one teaching strategy is better than another

Importance of Estimating Treatment Effect

Suppose we ran an online education experiment using a bandit algorithm to test different teaching strategies.

If we could infer the treatment effect between bandit arms...

- When designing a new online course...
  - Under-performing arms could be eliminated or modified
  - High-performing arms could be studied further
- Could potentially publish findings, e.g., that one teaching strategy is better than another
- Identify new research directions

Importance of Estimating Treatment Effect

Suppose we ran an online education experiment using a bandit algorithm to test different teaching strategies.

If we could infer the treatment effect between bandit arms...

- When designing a new online course...
  - Under-performing arms could be eliminated or modified
  - High-performing arms could be studied further
- Could potentially publish findings, e.g., that one teaching strategy is better than another
- Identify new research directions
- Make more informed high-level decisions regarding what new courses to design or fund

Confidence intervals for treatment effect

- We don’t use high probability concentration bounds
  - Confidence intervals too wide for many applications

Confidence intervals for treatment effect

- We don’t use high probability concentration bounds
  - Confidence intervals too wide for many applications
- We use asymptotic distribution of estimator of treatment effect to approximate small sample
  - Long history of being effect in classical statistics

Batched bandit setting

- Fix number of batches T
- Select n arms in each batch
- nT arm pulls total. Update bandit algorithm T times

Batched bandit setting

- Fix number of batches T
- Select n arms in each batch
- nT arm pulls total. Update bandit algorithm T times

For example, this corresponds to online advertising problems in which ads are sent out to many users simultaneously.

Batched bandit setting

- Fix number of batches T
- Select n arms in each batch
- nT arm pulls total. Update bandit algorithm T times

For example, this corresponds to online advertising problems in which ads are sent out to many users simultaneously.

- We analyze asymptotics as n (batch size) goes to infinity with T (number of batches) fixed
  - We do not need n to go to infinity in real experiments
  - We analyze asymptotics to get good approximation of finite sample behavior of estimators

Notation and Assumptions

We focus on the two-arm bandit setting.

- **Expected rewards:** β_0, β_1
- **Treatment effect:** $\Delta = \beta_1 - \beta_0$

Notation and Assumptions

We focus on the two-arm bandit setting.

- **Expected rewards:** β_0, β_1
- **Treatment effect:** $\Delta = \beta_1 - \beta_0$

- **Action selection probabilities:** $\pi_t \in [0, 1]$, function of history H_{t-1}
- **Actions:** $\{A_{t,i}\}_{i=1}^{n} \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi_t)$

Notation and Assumptions

We focus on the two-arm bandit setting.

- **Expected rewards:** β_0, β_1
- **Treatment effect:** $\Delta = \beta_1 - \beta_0$

- **Action selection probabilities:** $\pi_t \in [0, 1]$, function of history H_{t-1}
- **Actions:** $\{A_{t,i}\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi_t)$

- **Rewards:** $\{R_{t,i}\}_{i=1}^n$ with $R_{t,i} = \beta_{1,t} A_{t,i} + \beta_{0,t}(1 - A_{t,i}) + \epsilon_{t,i}$ and $\mathbb{E}[\epsilon_{t,i} | H_{t-1}, A_{t,i}] = 0$

Notation and Assumptions

We focus on the two-arm bandit setting.

- **Expected rewards:** β_0, β_1
- **Treatment effect:** $\Delta = \beta_1 - \beta_0$

- **Action selection probabilities:** $\pi_t \in [0, 1]$, function of history H_{t-1}
- **Actions:** $\{A_{t,i}\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi_t)$

- **Rewards:** $\{R_{t,i}\}_{i=1}^n$ with $R_{t,i} = \beta_{1,t} A_{t,i} + \beta_{0,t}(1 - A_{t,i}) + \epsilon_{t,i}$ and $\mathbb{E}[\epsilon_{t,i} | H_{t-1}, A_{t,i}] = 0$

- **History:** $H_t = \cup_{t'<t} \{A_{t',i}, R_{t',i}\}_{i=1}^n$
- **Batch Action Selection Count:** $N_t = \sum_{i=1}^n A_{t,i}$

Contributions

- Proving that OLS estimator does not converge uniformly on bandit data
  - Assuming the OLS estimator is asymptotically normal can lead to inflated Type-1 errors and unreliable confidence intervals

Contributions

- Proving that OLS estimator does not converge uniformly on bandit data
  - Assuming the OLS estimator is asymptotically normal can lead to inflated Type-1 errors and unreliable confidence intervals
- Prove that the Batched OLS estimator is asymptotically normal
  - Can construct confidence intervals for treatment effect on bandit data
  - Not specific to a particular bandit algorithm—works for a variety of algorithms
  - Robust to non-stationarity in the baseline reward

Why not use standard statistical estimators on bandit data?

Induced dependence

- For bandit data, $\{A_{t,i}, R_{t,i}\}_{t=1}^{T}$ are **not** independent.

Induced dependence

- For bandit data, $\{A_{t,i}, R_{t,i}\}_{t=1}^{T}$ are **not** independent.
- Actions $A_{t,i}$ depend the history of past actions and rewards H_{t-1}, i.e., $\{A_{t',i}, R_{t',i}\}_{i=1}^{n}$ for $t' < t$.

Induced dependence

- For bandit data, $\{A_{t,i}, R_{t,i}\}_{t=1}^{T}$ are **not** independent.
- Actions $A_{t,i}$ depend the history of past actions and rewards H_{t-1}, i.e., $\{A_{t',i}, R_{t',i}\}_{i=1}^{n}$ for $t' < t$.

However, most asymptotic results for statistical estimators assume independence!

Induced dependence

- For bandit data, $\{A_{t,i}, R_{t,i}\}_{t=1}^{T}$ are **not** independent.
- Actions $A_{t,i}$ depend the history of past actions and rewards H_{t-1}, i.e., $\{A_{t',i}, R_{t',i}\}_{i=1}^{n}$ for $t' < t$.

However, most asymptotic results for statistical estimators assume independence!

We now discuss what can go wrong...

Bias: Example of what can go wrong?

How is bandit data different from i.i.d data?

- *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].

3 sample illustration of bias

Suppose we have two arms that both have zero mean reward and we follow a greedy policy.

- Rewards: $R_1, R_2, R_3 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$
- Actions: $A_1, A_2, A_3 \in \{0, 1\}$

Suppose $A_1 = 1$, $A_2 = 0$ and $A_3 = \mathbb{I}_{(R_1 > R_2)}$ (greedy strategy).

3 sample illustration of bias

Suppose we have two arms that both have zero mean reward and we follow a greedy policy.

- Rewards: $R_1, R_2, R_3 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$
- Actions: $A_1, A_2, A_3 \in \{0, 1\}$

Suppose $A_1 = 1$, $A_2 = 0$ and $A_3 = \mathbb{I}_{(R_1 > R_2)}$ (greedy strategy).

Sample means:

$$\hat{\beta}_1 = \frac{R_1 + A_3 R_3}{1 + A_3} \qquad \hat{\beta}_0 = \frac{R_2 + (1 - A_3) R_3}{1 + (1 - A_3)}$$

3 sample illustration of bias

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{R_1 + A_3 R_3}{1 + A_3}\right]$$

3 sample illustration of bias

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{R_1 + A_3 R_3}{1 + A_3}\right]$$

$$= \mathbb{P}(R_1 > R_2)\mathbb{E}\left[\frac{R_1 + R_3}{2} \,\middle|\, A_3 = 1\right] + \mathbb{P}(R_1 \leq R_2)\mathbb{E}[R_1 \mid A_3 = 0]$$

3 sample illustration of bias

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{R_1 + A_3 R_3}{1 + A_3}\right]$$

$$= \mathbb{P}(R_1 > R_2)\mathbb{E}\left[\frac{R_1 + R_3}{2} \;\middle|\; A_3 = 1\right] + \mathbb{P}(R_1 \leq R_2)\mathbb{E}[R_1 \mid A_3 = 0]$$

$$= \frac{1}{2}\mathbb{E}\left[\frac{R_1 + R_3}{2} \;\middle|\; R_1 > R_2\right] + \frac{1}{2}\mathbb{E}[\, R_1 \mid R_1 \leq R_2]$$

3 sample illustration of bias

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{R_1 + A_3 R_3}{1 + A_3}\right]$$

$$= \mathbb{P}(R_1 > R_2)\mathbb{E}\left[\frac{R_1 + R_3}{2} \;\middle|\; A_3 = 1\right] + \mathbb{P}(R_1 \leq R_2)\mathbb{E}[R_1 \mid A_3 = 0]$$

$$= \frac{1}{2}\mathbb{E}\left[\frac{R_1 + R_3}{2} \;\middle|\; R_1 > R_2\right] + \frac{1}{2}\mathbb{E}[\, R_1 \mid R_1 \leq R_2]$$

Let $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$; $Z_{\mathsf{max}} := \mathsf{max}(Z_1, Z_2)$ and $Z_{\mathsf{min}} := \mathsf{min}(Z_1, Z_2)$.

$$= \frac{1}{4}\mathbb{E}[\, Z_{\mathsf{max}}\,] + \frac{1}{2}\mathbb{E}[\, Z_{\mathsf{min}}\,]$$

3 sample illustration of bias

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{R_1 + A_3 R_3}{1 + A_3}\right]$$

$$= \mathbb{P}(R_1 > R_2)\mathbb{E}\left[\frac{R_1 + R_3}{2} \;\middle|\; A_3 = 1\right] + \mathbb{P}(R_1 \leq R_2)\mathbb{E}[R_1 \mid A_3 = 0]$$

$$= \frac{1}{2}\mathbb{E}\left[\frac{R_1 + R_3}{2} \;\middle|\; R_1 > R_2\right] + \frac{1}{2}\mathbb{E}[\, R_1 \mid R_1 \leq R_2]$$

Let $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$; $Z_{\mathsf{max}} := \mathsf{max}(Z_1, Z_2)$ and $Z_{\mathsf{min}} := \mathsf{min}(Z_1, Z_2)$.

$$= \frac{1}{4}\mathbb{E}[\, Z_{\mathsf{max}}\,] + \frac{1}{2}\mathbb{E}[\, Z_{\mathsf{min}}\,]$$

Since $\mathbb{E}[Z_{\mathsf{max}}] = -\mathbb{E}[Z_{\mathsf{min}}]$ by symmetry,

$$= -\frac{1}{4}\mathbb{E}[\, Z_{\mathsf{max}}\,] < 0$$

Asymptotic Distribution of the OLS Estimator

OLS Estimator of Δ:

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

where $N_t = \sum_{i=1}^{n} A_{t,i}$

Asymptotic Distribution of the OLS Estimator

OLS Estimator of Δ:

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

where $N_t = \sum_{i=1}^{n} A_{t,i}$

On i.i.d data, we have asymptotic normality:

$$\sqrt{\frac{(\sum_{t=1}^{T} N_t)(\sum_{t=1}^{T} n - N_t)}{nT}}(\hat{\Delta}^{\text{OLS}} - \Delta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Asymptotic Distribution of the OLS Estimator

OLS Estimator of Δ:

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

where $N_t = \sum_{i=1}^{n} A_{t,i}$

On i.i.d data, we have asymptotic normality:

$$\sqrt{\frac{(\sum_{t=1}^{T} N_t)(\sum_{t=1}^{T} n - N_t)}{nT}}(\hat{\Delta}^{\text{OLS}} - \Delta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

On bandit data, the same asymptotic normality result holds if (result by Lai & Wei, 1982):

For some non-random sequence of scalars $\{a_n\}_{n=1}^{\infty}$, as $n \to \infty$,

$$a_n \cdot \frac{1}{nT}\sum_{t=1}^{T} N_t \xrightarrow{P} 1.$$

Asymptotic Distribution of the OLS Estimator (cont.)

1. $\epsilon_{t,i}$ satisfy moment conditions
2. $\pi_t \in [\pi_{\mathsf{min}}, \pi_{\mathsf{max}}]$ for constants $0 < \pi_{\mathsf{min}} \leq \pi_{\mathsf{max}} < 1$

Asymptotic Distribution of the OLS Estimator (cont.)

1. $\epsilon_{t,i}$ satisfy moment conditions
2. $\pi_t \in [\pi_{\mathbf{min}}, \pi_{\mathbf{max}}]$ for constants $0 < \pi_{\mathbf{min}} \leq \pi_{\mathbf{max}} < 1$

If $\Delta \neq 0$:

- The conditions of Lai & Wei's central limit theorem hold.
- $\hat{\Delta}^{\text{OLS}}$ is asymptotically Normal

Asymptotic Distribution of the OLS Estimator (cont.)

1. $\epsilon_{t,i}$ satisfy moment conditions
2. $\pi_t \in [\pi_{\min}, \pi_{\max}]$ for constants $0 < \pi_{\min} \leq \pi_{\max} < 1$

If $\Delta \neq 0$:

- The conditions of Lai & Wei's central limit theorem hold.
- $\hat{\Delta}^{\text{OLS}}$ is asymptotically Normal

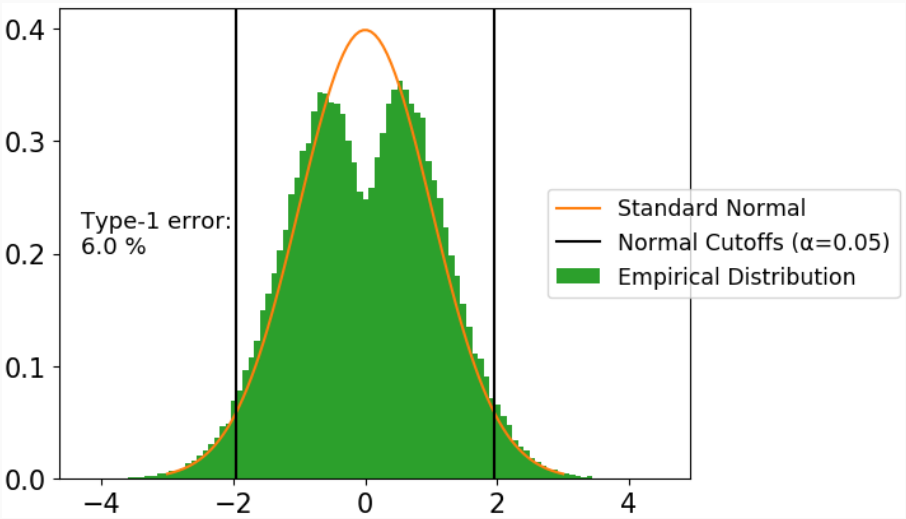
If $\Delta = 0$:

- The conditions of Lai & Wei's CLT do not to hold.
- $\hat{\Delta}^{\text{OLS}}$ is asymptotically non-Normal
- For common bandit algorithms, including Thompson Sampling and ϵ-greedy

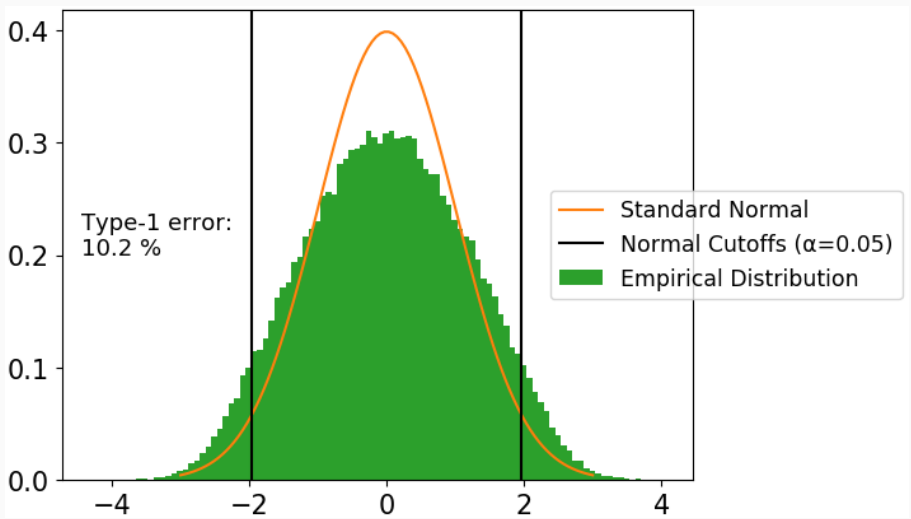
Inflated Type-1 Error!

$H_0 : \Delta = 0$ **vs.** $H_1 : \Delta \neq 0$

Z-statistic for treatment effect when the null hypothesis is true



ϵ-greedy



Thompson Sampling

$\mathcal{N}(0, 1)$ rewards, $T = 25$, $n = 100$, $\beta_1 = \beta_0 = 0$

Why? Non-concentration of π_t

Why? Non-concentration of π_t

Non-Zero Treatment Effect Case

π_t will converge to the optimal policy (π_{min} or π_{max})

Why? Non-concentration of π_t

Non-Zero Treatment Effect Case

π_t will converge to the optimal policy (π_{min} or π_{max})

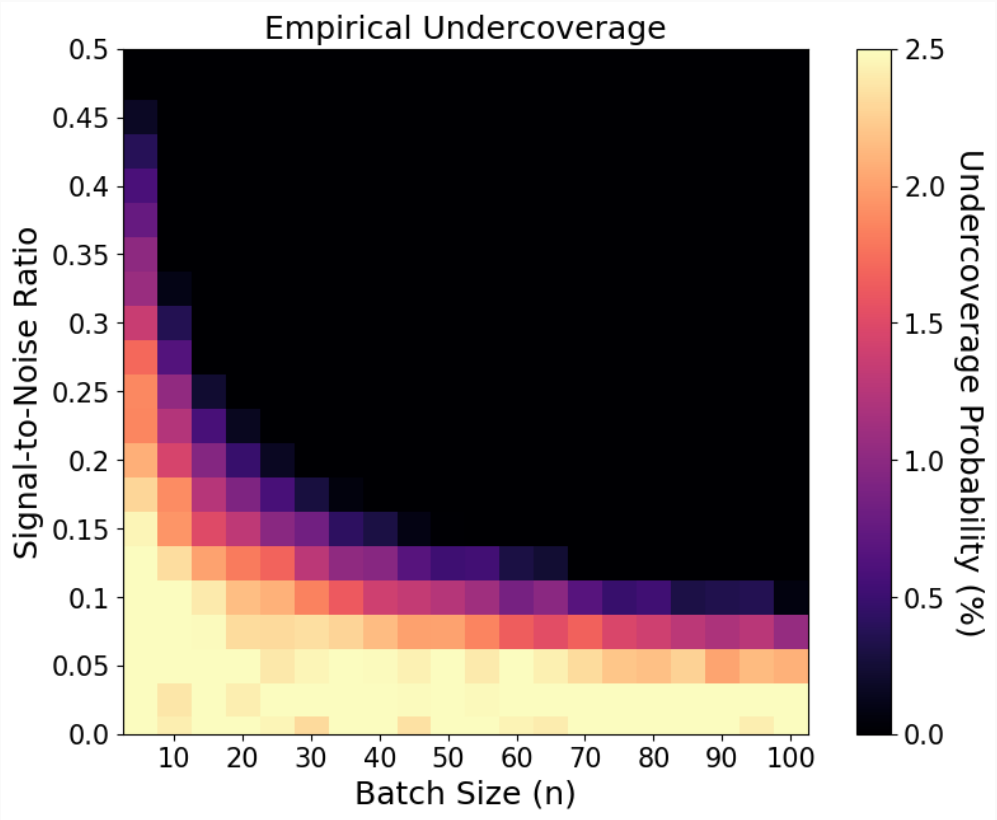
Zero Treatment Effect Case

No unique optimal policy, so π_t does not concentrate as $n \to \infty$

Non-uniform convergence & unreliable confidence intervals!

Signal-to-noise ratio: $\frac{|\Delta|}{\sigma}$

Thompson Sampling,
$\mathcal{N}(0, 1)$ rewards, $T = 25$



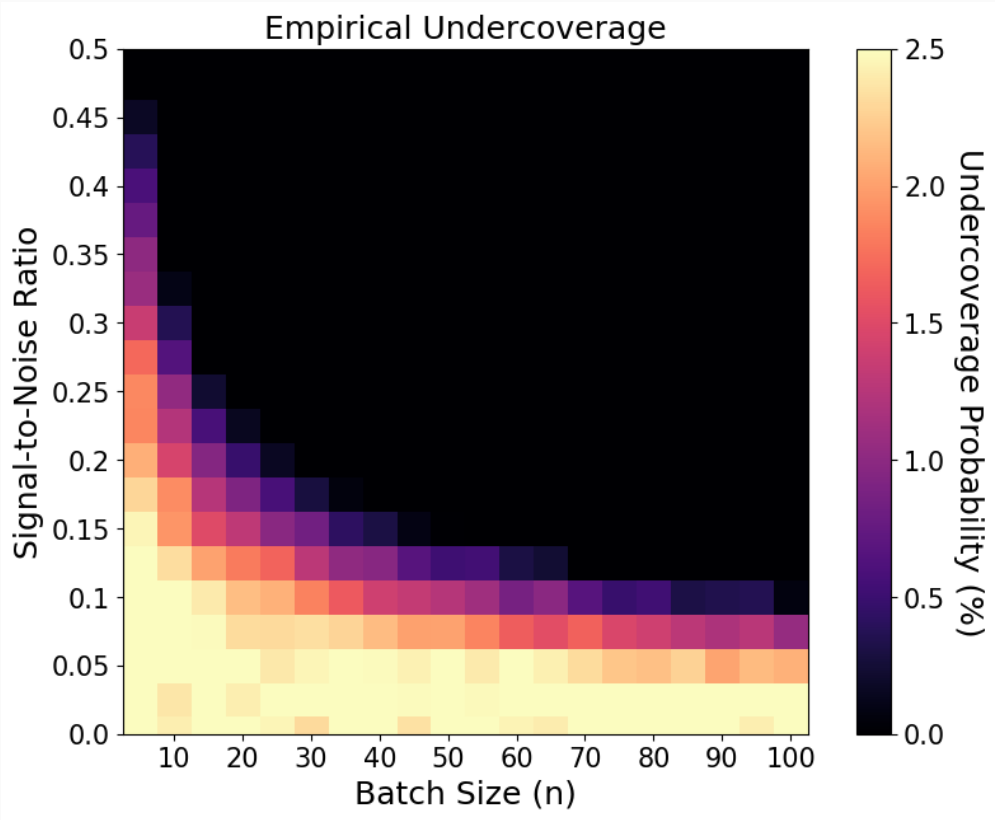
We construct 95% confidence intervals assuming the OLS estimator is approximately Normal.

Non-uniform convergence & unreliable confidence intervals!

Signal-to-noise ratio: $\frac{|\Delta|}{\sigma}$

Thompson Sampling,
$\mathcal{N}(0,1)$ rewards, $T = 25$

For any batch size n, we can find a treatment effect size Δ such would lead to confidence intervals that undercover.



We construct 95% confidence intervals assuming the OLS estimator is approximately Normal.

Recap of findings

On bandit data, the OLS estimator **does not converge uniformly over a range of different treatment effect sizes**

- Problems not limited to bias
- Type-1 error inflation and unreliable confidence intervals

Introduce the Batched OLS estimator

Non-stationarity in real world problems

Online advertisements

Effectiveness of ads may change over time due to...

- Previous exposure to the same ad make it less effective
- Introduction of competing ads
- General societal changes

Non-stationarity in real world problems

Online advertisements

Effectiveness of ads may change over time due to...

- Previous exposure to the same ad make it less effective
- Introduction of competing ads
- General societal changes

Mobile health

Effectiveness of messages may change over time due to...

- Habituation to notifications over time
- General changes in a person's routine

Introducing the Batched OLS estimator

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

Introducing the Batched OLS estimator

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

Standard OLS Estimator:

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

Introducing the Batched OLS estimator

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

Standard OLS Estimator:

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

Batched OLS Estimator:

For each batch $t \in [1:T]$,

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

Batched OLS (BOLS) Multivariate CLT

Asymptotic Normality of BOLS

1. $\mathbb{E}[\epsilon_{t,i}^2 | H_{t-1}, A_{t,i}] = \sigma^2$ and $\mathbb{E}\left[\epsilon_{t,i}^4 \middle| H_{t-1}, A_{t,i}\right] < M < \infty$ for all t, n, i.
2. $\mathbb{P}(\pi_t \in [f(n), 1 - f(n)]) \to 1$ for non-random $f(n) = \omega(\frac{1}{n})$.[1]

If the above two conditions hold then as $n \to \infty$,

$$\begin{bmatrix} \sqrt{\frac{(n-N_1)N_1}{n}}(\hat{\Delta}_1^{\text{BOLS}} - \Delta_1) \\ \sqrt{\frac{(n-N_2)N_2}{n}}(\hat{\Delta}_2^{\text{BOLS}} - \Delta_2) \\ \vdots \\ \sqrt{\frac{(n-N_T)N_T}{n}}(\hat{\Delta}_T^{\text{BOLS}} - \Delta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_T)$$

[1] $f(n)n \to \infty$

Batched OLS (BOLS) Multivariate CLT

Asymptotic Normality of BOLS

1. $\mathbb{E}[\epsilon_{t,i}^2 | H_{t-1}, A_{t,i}] = \sigma^2$ and $\mathbb{E}\left[\epsilon_{t,i}^4 \middle| H_{t-1}, A_{t,i}\right] < M < \infty$ for all t, n, i.
2. $\mathbb{P}(\pi_t \in [f(n), 1 - f(n)]) \to 1$ for non-random $f(n) = \omega(\frac{1}{n})$.[1]

If the above two conditions hold then as $n \to \infty$,

$$\begin{bmatrix} \sqrt{\frac{(n-N_1)N_1}{n}}(\hat{\Delta}_1^{\text{BOLS}} - \Delta_1) \\ \sqrt{\frac{(n-N_2)N_2}{n}}(\hat{\Delta}_2^{\text{BOLS}} - \Delta_2) \\ \vdots \\ \sqrt{\frac{(n-N_T)N_T}{n}}(\hat{\Delta}_T^{\text{BOLS}} - \Delta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_T)$$

We show a similar asymptotic normality result for BOLS for K-armed linear contextual bandits. See our paper for more details!

[1] $f(n)n \to \infty$

Key to Batched OLS

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

$$\frac{N_t}{n\pi_t} \xrightarrow{P} 1.$$

Key to Batched OLS

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

$$\frac{N_t}{n\pi_t} \xrightarrow{P} 1.$$

Since π_t is constant given H_{t-1}, even if π_t does not concentrate, we are still able to apply the martingale CLT to prove asymptotic normality.

BOLS Test statistic

There are many hypotheses we could test using the BOLS multivariate Normality result. Here we consider the following hypotheses:

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

BOLS Test statistic

There are many hypotheses we could test using the BOLS multivariate Normality result. Here we consider the following hypotheses:

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

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{\frac{(n - N_t)N_t}{n\sigma^2}} (\hat{\Delta}_t^{\text{BOLS}} - c)$$

By our multivariate CLT for BOLS, the above will be asymptotically normal under the null.

BOLS is robust to non-stationarity in the baseline reward, i.e., $\beta_{t,1}, \beta_{t,0}$ can change from batch to batch, but $\Delta_t := \beta_{t,1} - \beta_{t,0} = c$ for all t.

Other estimators we compare to

W-Decorrelated [Deshpande et al., 2018]

- Adjusted version of the OLS estimator
- Requires choosing a tuning parameter λ, which allows practitioners to trade off bias for variance

Adaptively-Weighted Augmented Inverse Probability Weighted Estimator (AW-AIPW) [Hadad et al., 2019]

- Reweights the samples of a regular AIPW estimator with adaptive weights that are non-anticipating

Note that neither of these estimators have guarantees in non-stationary settings.

Simulations: Stationary setting Type-1 error

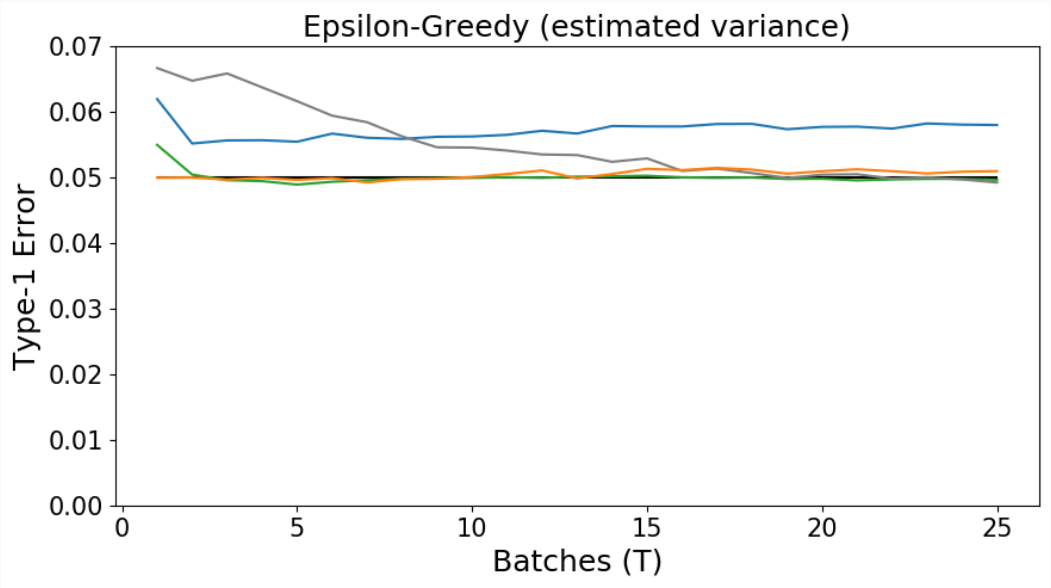
$H_0 : \Delta = 0$

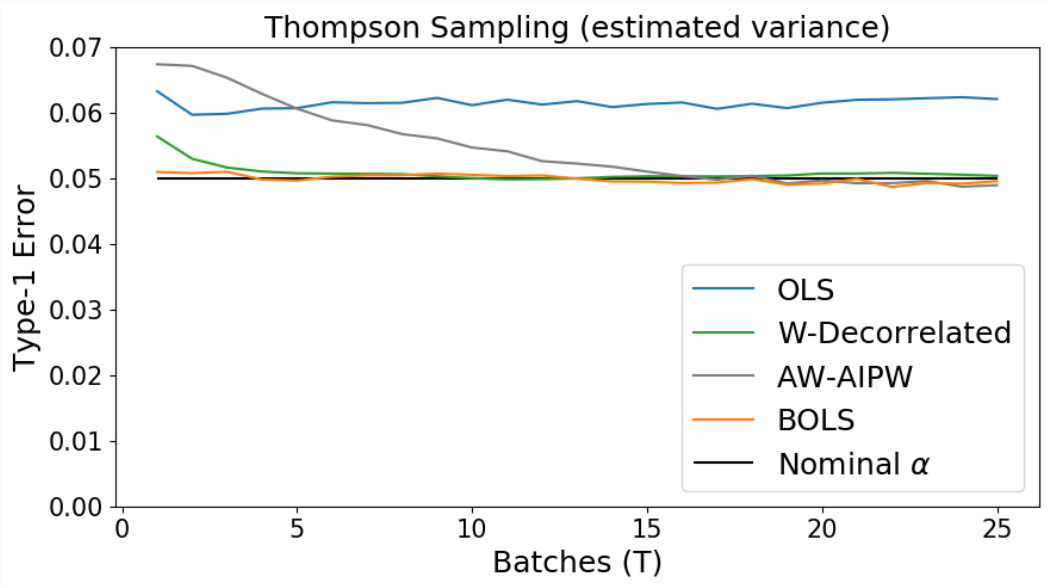
$H_1 : \Delta \neq 0$

Type-1 error

Probability of incorrectly rejecting null hypothesis (constrained ≤ 0.05).

$\mathcal{N}(0, 1)$ rewards, $n = 25$, $\beta_1 = \beta_0 = 0$



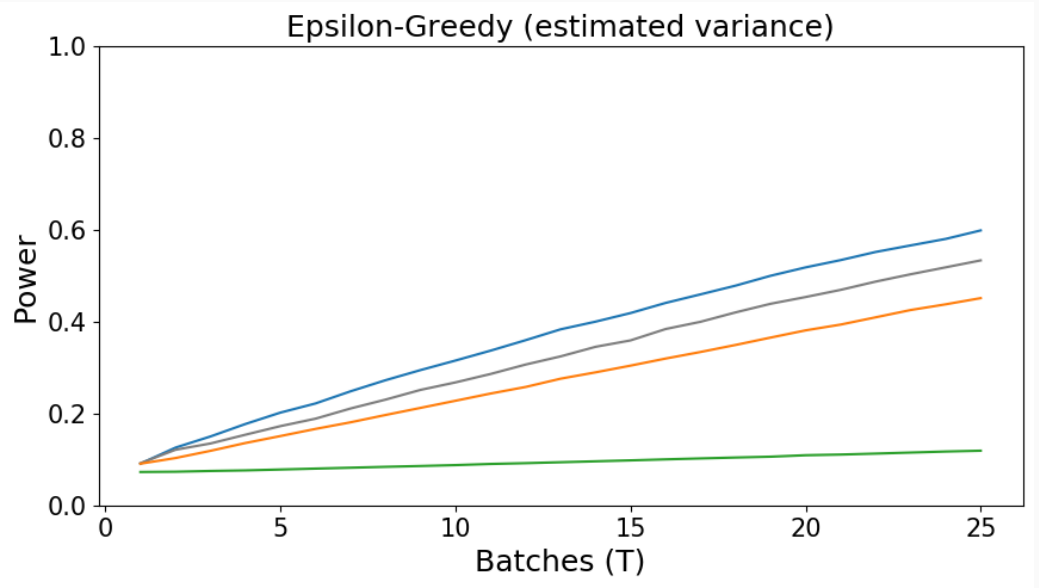


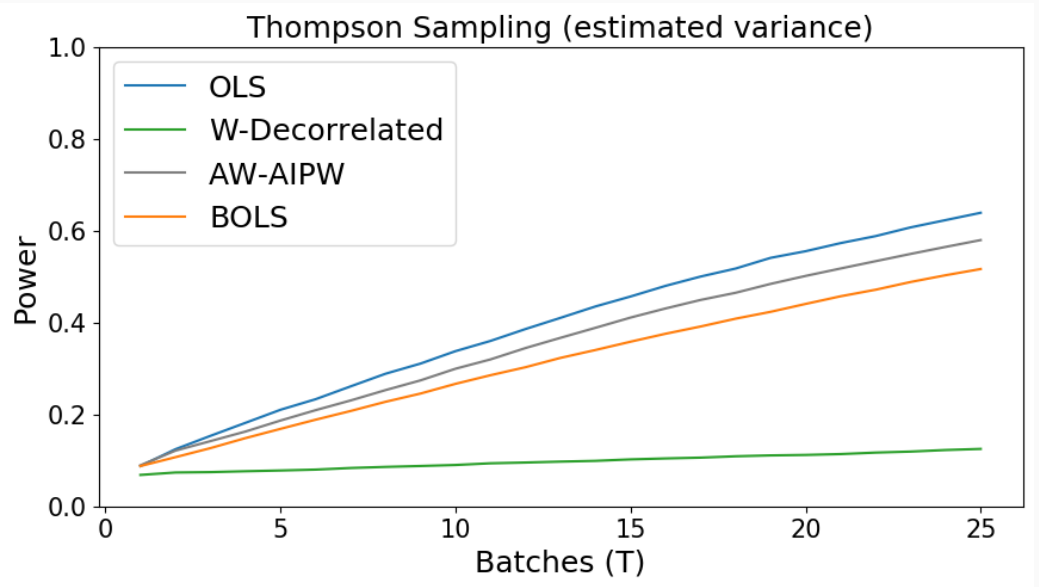
Simulations: Stationary setting power

Power

Probability of correctly rejecting null hypothesis.

$\mathcal{N}(0, 1)$ rewards, $n = 25$,
$\beta_1 = 0.25$, $\beta_0 = 0$





Simulations: Non-stationary baseline reward

Fixed treatment effect

$H_0 : \Delta_t = 0, \forall t$

$H_1 : \Delta_t = c, \forall t$ for $c \neq 0$

Figure 1: Zero treatment effect (H_0)

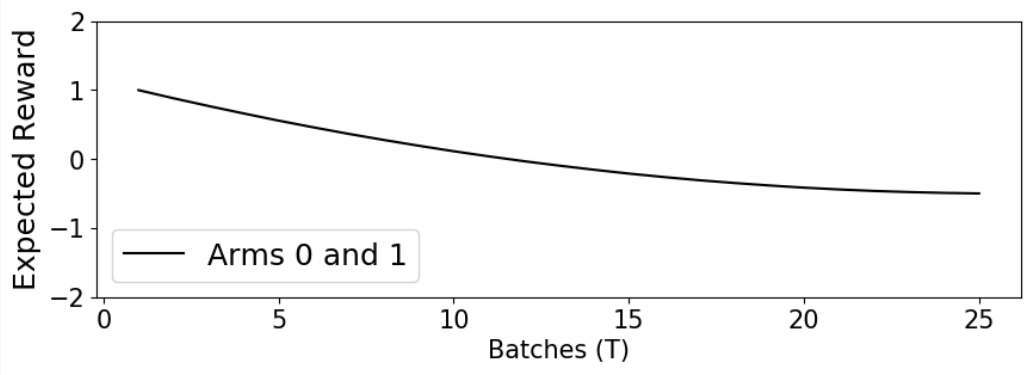
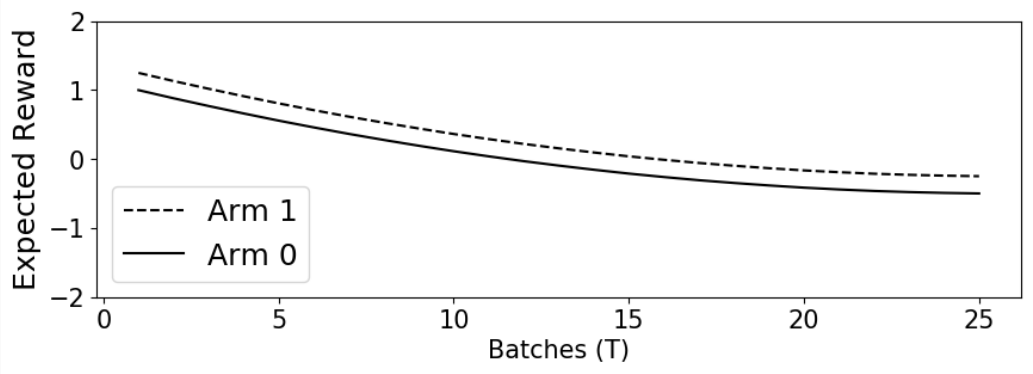


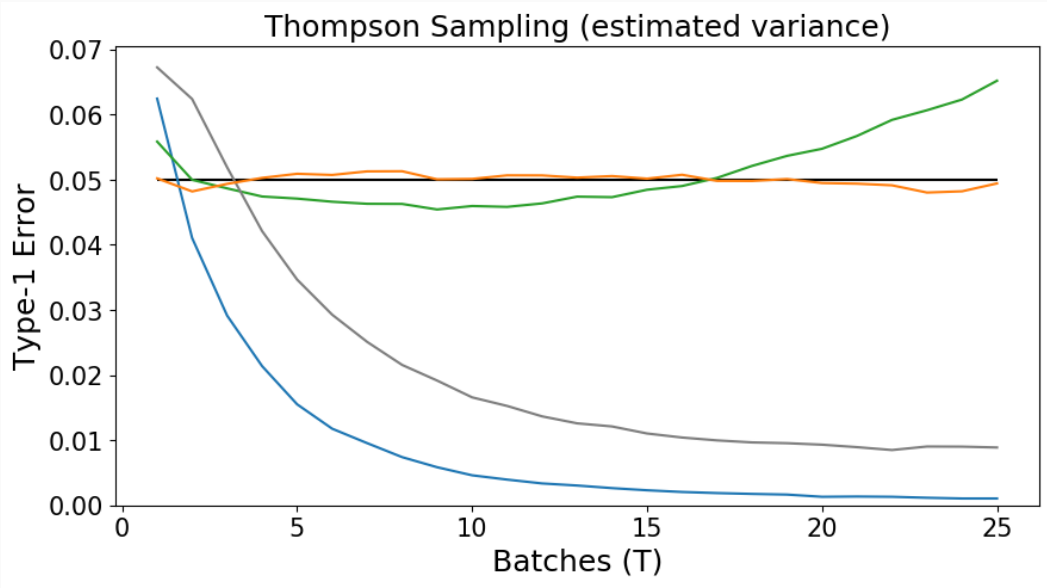
Figure 2: Non-zero treatment effect (H_1)

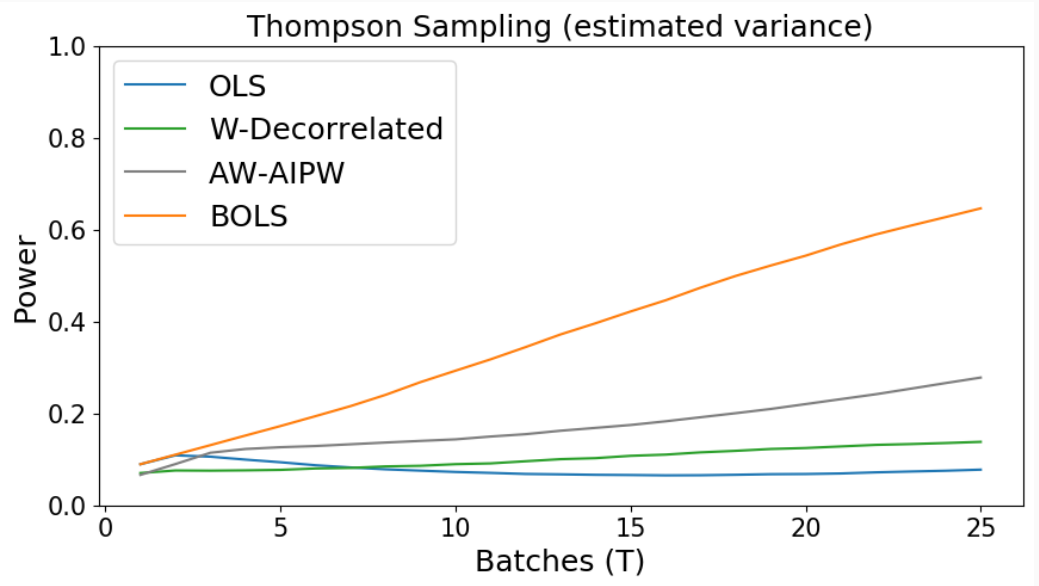


Simulations: Non-stationary baseline reward

BOLS still has proper Type-1 error control and high power.

Other estimators have no guarantees in the non-stationary setting.





Conclusion

- We demonstrate that that standard statistical estimators can converge *non-uniformly* on bandit data.

Conclusion

- We demonstrate that that standard statistical estimators can converge *non-uniformly* on bandit data.
- Assuming asymptotic normality of the OLS estimator can lead to inflated Type-1 error and unreliable confidence intervals on bandit data.

Conclusion

- We demonstrate that that standard statistical estimators can converge *non-uniformly* on bandit data.
- Assuming asymptotic normality of the OLS estimator can lead to inflated Type-1 error and unreliable confidence intervals on bandit data.
- We develop the BOLS estimator that is asymptotically normal even when the treatment effect is zero.

Conclusion

- We demonstrate that that standard statistical estimators can converge *non-uniformly* on bandit data.
- Assuming asymptotic normality of the OLS estimator can lead to inflated Type-1 error and unreliable confidence intervals on bandit data.
- We develop the BOLS estimator that is asymptotically normal even when the treatment effect is zero.
- BOLS is robust to non-stationarity over batches.

Directions for Future Work

- Batched version of method-of-moments estimators
- Allowing for correlation between rewards over batches
- Trade-off between regret minimization and power of statistical tests

References

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. (2018).

Accurate inference for adaptive linear models.

In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1194–1203, Stockholmsmässan, Stockholm Sweden. PMLR.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. (2019).

Balanced linear contextual bandits.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2019).

Confidence intervals for policy evaluation in adaptive experiments.

arXiv preprint arXiv:1911.02768.

Nie, X., Tian, X., Taylor, J., and Zou, J. (2018).

Why adaptively collected data have negative bias and how to correct for it.

International Conference on Artificial Intelligence and Statistics.

Romano, J. P., Shaikh, A. M., et al. (2012).

On the uniform asymptotic validity of subsampling and the bootstrap.

The Annals of Statistics, 40(6):2798–2822.

Shin, J., Ramdas, A., and Rinaldo, A. (2019).

Are sample means in multi-armed bandits positively or negatively biased?

In *Advances in Neural Information Processing Systems*, pages 7100–7109.