Inference for Batched Bandits

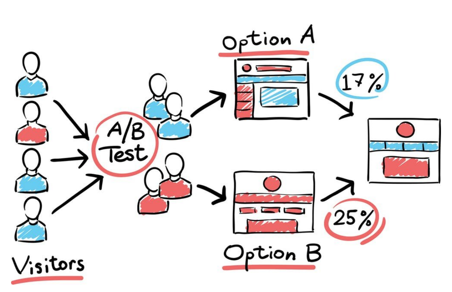
Kelly W. Zhang, Lucas Janson, Susan A. Murphy

Harvard John A. Paulson School of Engineering and Applied Sciences

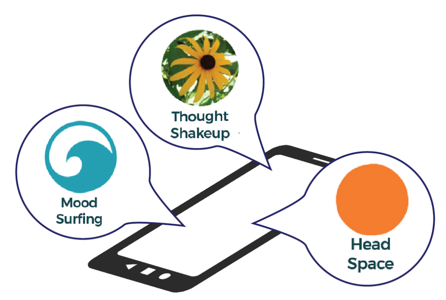
NEURAL INFORMATION PROCESSING SYSTEMS

Bandits in the Real World

Bandit algorithms are increasingly used in real world problems due to their regret minimization properties. For example in online recommendations, mobile health, and online education.



300 COURSES



Need for Uncertainty Quantification

- **Goal:** Constructing confidence intervals and perform hypothesis testing using bandit data (e.g. CI for margin, or the difference in expected rewards between two arms)
- **Uncertainty quantification is crucial for:**
  - **Scientific discovery:** sharing findings (e.g. in online education setting, we may find that one teaching strategy performs better than another)
  - **Regret minimization ``between" experiments:** inform design of future experiments (e.g. drop under-performing arms in the version of the online course)
- **Confidence intervals:** Construct using asymptotic distribution of estimators by approximating finite sample distributions of estimators with their asymptotic distribution.
  - Asymptotic approximations has a long history of success in science and leads to much narrower CI than those constructed using high probability bounds.

Batched Bandit Setting

- Bandit arms are selected in batches of size n for a total of T batches.
- We analyze asymptotics as $n \to \infty$ with T fixed.

Motivation 1: Batched setting is common in digital age
- Multiple users take the course/use the app/visit site at once

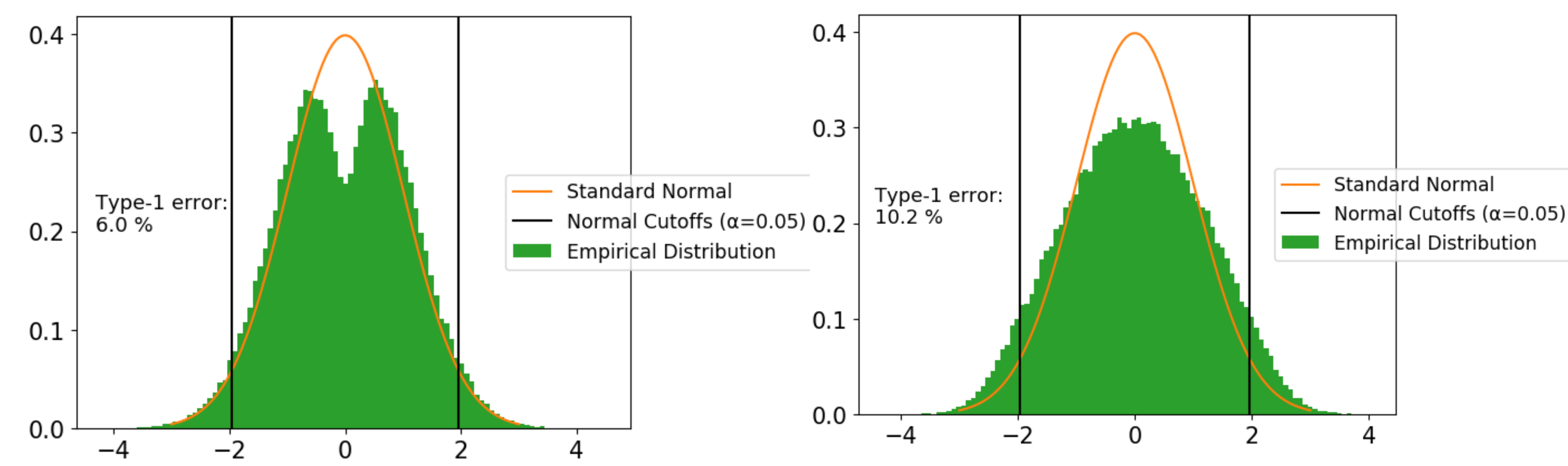
Motivation 2: Temporal non-stationarity
- Users become disengaged over time (online education, mobile health)
- News/ad popularity changes over time

Motivation 3: Often T cannot be adjusted, but n can
- Online education course cannot be made arbitrarily long
- Mobile health studies length of depends on domain science

Why is inference on bandit data challenging?

- For bandit data, $\{A_{t,i}, R_{t,i}\}_{t=1}^T$ are **not** independent, because actions $A_{t,i}$ depend the history $H_{t-1,n} = \{A_{t',i}, R_{t',i}\}_{i=1}^n$ for $t' < t$
- Estimators that are asymptotically normal for independent data can be asymptotically non-normal have inflated Type-1 Error on bandit data

Z-statistic for treatment effect using OLS estimator when the $\Delta = 0$
Testing $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$



ϵ-greedy   Thompson Sampling

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

Notation

- We focus on the two-arm bandit setting. For all $t \in [1: T]$,
- **Expected rewards:** $\beta_{0,t}, \beta_{1,t}$
- **Treatment effect:** $\Delta_t = \beta_{1,t} - \beta_{0,t}$
- **Action selection probabilities:** $\pi_t \in [0,1]$, function of history $H_{t-1,n}$
- **Actions:** $\{A_{t,i}\}_{i=1}^n \overset{iid}{\sim} Bernoulli(\pi_t)$ conditional on $H_{t-1,n}$
- **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 $E[\epsilon_{t,i} | H_{t-1}, A_{t,i}] = 0$
- **History:** $H_t = \cup_{t'<t} \{A_{t',i}, R_{t',i}\}_{i=1}^n$

Batched OLS Estimator (BOLS)

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

Standard OLS Estimator:

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

Batched OLS Estimator:

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

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

Batched OLS Test Statistic

Here we consider the following hypotheses: $H_0 : \Delta = c$ vs. $H_1 : \Delta \neq c$

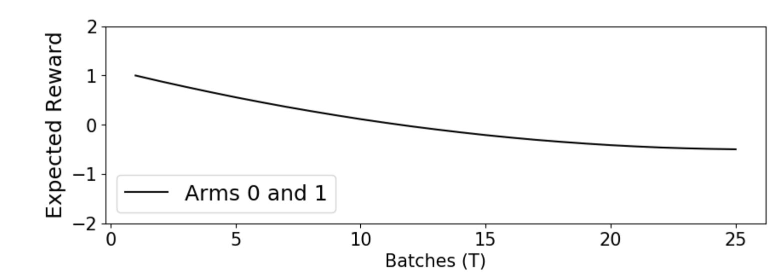
$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \sqrt{\frac{(n - N_t) N_t}{n \sigma^2}} (\hat{\Delta}_t^{BOLS} - c) \xrightarrow{D} \mathcal{N}(0,1)$$

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

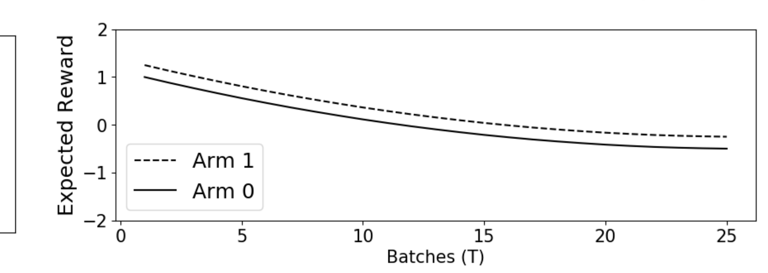
BOLS is robust to non-stationarity in the baseline reward, i.e., $\beta_{t,1}, \beta_{t,0}$ can change from batch to batch, but $\Delta_t := \beta_{t,1} - \beta_{t,0} = c$ for all t.

Non-Stationary Baseline Reward Simulations

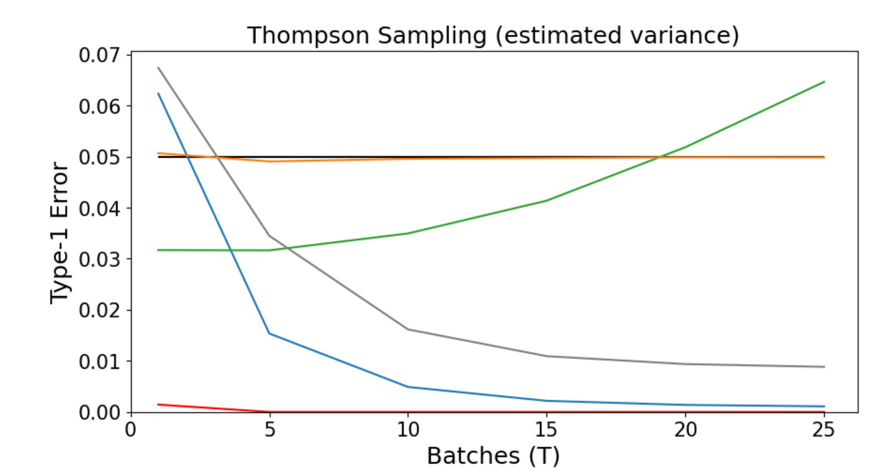
Null Arm Means $\Delta = 0$



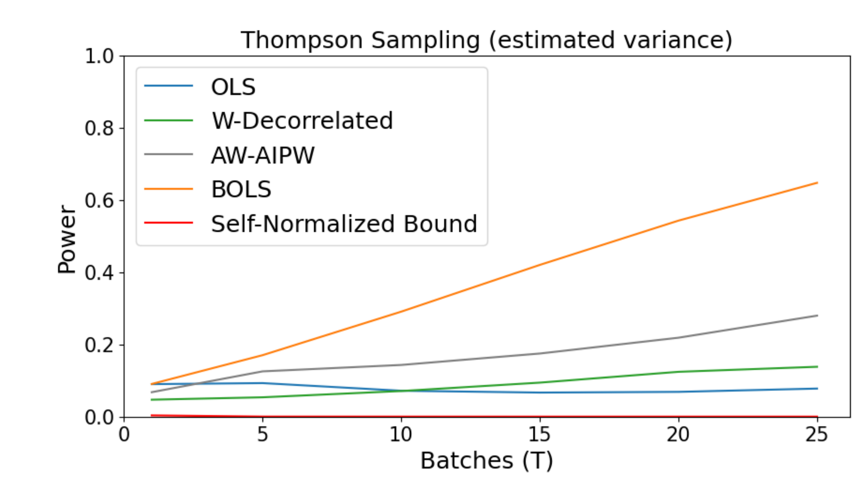
Alternative Arm Means $\Delta = 0.25$



Type-1 Error (prob. falsely rejecting null)



Power (prob. correctly rejecting null)



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.
- We develop the BOLS estimator that is asymptotically normal even when the treatment effect is zero for multi-arm and contextual bandits.
- BOLS is robust to non-stationarity over batches.

References

Deshpande, Y., Mackey, L., Syrgkanis, V. and Taddy, M., 2018, July. Accurate inference for adaptive linear models. In *International Conference on Machine Learning* (pp. 1194-1203). PMLR.

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.*

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C., 2011. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems* (pp. 2312-2320).