Inference for Bandit Algorithms with Pooling for Mobile Health

Kelly Zhang [1]

[1]Harvard University

Motivation and Problem Overview

Learning a single policy for all users by pooling the data of multiple users together could potentially allow for much faster learning, especially when the users are similar. However, techniques for inference after the study is over generally assume that the sequence of rewards from each user are independent. Pooling the users' data creates dependencies in the user reward sequences over time, since how one user behaves affects how the algorithm learns, which in turn can affect how the policy makes a decision on a different user in the future.

Our primary goal is to obtain a confidence interval that has faithful empricial coverage in finite samples for the treatment effect or margin between the mean rewards for two actions. For this project, we focus on the simplified setting in which we assume that there is no state information and all users have the same mean rewards for each action. We primarily build on the work of Lai and Wei [2], who prove a central limit theorem for the OLS estimator for dependent samples.

Pooling Problem Setup

Let there be m users and T timesteps. We assume a total of $n = mT$ samples, T from each user, and assume that all users start the study at the same time. We do not assume any state or context information. Furthermore, we assume that all users have the same mean rewards for all actions with the following model for user u and timestep t,

$$y_t^{(u)} =< \mathbf{x}_t^{(u)}, \boldsymbol{\beta} > +\epsilon_t^{(u)}$$

- $\boldsymbol{\beta} \in \mathbb{R}^2$ is a vector of mean rewards for the two arms
- $y_t^{(u)} \in \mathbb{R}$ is the observed reward
- $\mathbf{x}_t^{(u)} \in \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$ is a vector in $\mathbb{R}^2$ indicating which of the two actions was chosen
- $\epsilon_t^{(u)} \in \mathbb{R}$ is the noise parameter (form a martingale difference array)

We define the history $H_t := \{\mathbf{x}_i^{(u)}, y_i^{(u)}, \epsilon_i^{(u)} : i \leq t, \forall u \in [1, m]\}$. We also define a filtration $\{\mathcal{F}_t\}_{t=0}^T$ where $\mathcal{F}_t = \sigma(H_t)$ is a sigma-algebra of the data for all users up to timestep t. We also assume that $\mathbf{x}_t^{(u)}, y_t^{(u)}, \epsilon_t^{(u)}$ are well adapted to the filtration, meaning they are $\mathcal{F}_t$ measurable.

We assume that the data is adaptively collected so $\mathbf{x}_t^{(u)}, y_t^{(u)}, \epsilon_t^{(u)}$ can depend on data from all users from previous time periods $\{\mathbf{x}_i^{(u)}, y_i^{(u)}, \epsilon_i^{(u)}\}_{i<t, \forall u}$. Moreover, conditioned on the sigma algebra $\mathcal{F}_{t-1}$, the data collected from different users in the same timestep are independent, so for all $u \neq v$,

$$\left((\mathbf{x}_t^{(u)}, y_t^{(u)}, \epsilon_t^{(u)}) \perp\!\!\!\perp (\mathbf{x}_t^{(v)}, y_t^{(v)}, \epsilon_t^{(v)})\right) \big| \mathcal{F}_{t-1}$$

We assume that the samples $\mathbf{x}_t^{(u)}$ are chosen according to some bandit algorithm, $\mathcal{A}(\cdot)$, that takes history H_t and outputs a distribution over actions. We assume that samples for all m users in a single time period $t+1$ are drawn i.i.d. from the distribution $\mathcal{A}(H_t)$,

$$\mathbf{x}_{t+1}^{(1)}, \mathbf{x}_{t+1}^{(2)}, ..., \mathbf{x}_{t+1}^{(m)} \overset{i.i.d.}{\sim} \mathcal{A}(H_t)$$

Future Work

- Extending analysis to the setting where $\mathbf{x}_i$ not only represents the action but also state (e.g. first two dimensions represent the action and the remaining dimensions represent context)
- Analyzing local asymptotics of L & W CLT - how fast we can allow the margin to go to zero and still have the CLT hold (to better understand finite sample performance)
- Analyzing the setting in which data is pooled, but users are assumed to be heterogenous and policies are personalized

Lai and Wei [2] Central Limit Theorem

$$\mathbf{y}_n = \mathbf{X}_n \beta + \boldsymbol{\epsilon}_n$$

For linear models, when the data $\mathbf{x}_i$ (actions in our case) are sampled i.i.d. from some distribution such that $E[\mathbf{x}_i \mathbf{x}_i^T]$ is full rank, if ϵ_i i.i.d. and independent of $\mathbf{x}_i$'s are such that $E[\epsilon_i] = 0$ and $E[\epsilon_i^2] = \sigma^2$, then the OLS estimator will have a central limit theorem of the form:

$$\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\right)^{1/2} (\hat{\beta}_{\text{OLS}} - \beta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 I_p)$$

In reinforcement learning, the actions $\mathbf{x}_i$ are not independent as they can depend on previous samples $\{\mathbf{x}_i, y_i, \epsilon_i\}_{i<n}$. Lai and Wei prove that the OLS estimator for adaptively sampled data will still have the *same* central limit theorem as the independently sampled $\mathbf{x}_i$ case as long as the noise ϵ_i and the sampling algorithm satisfy certain conditions.

Conditions

- $\{\epsilon_i\}_{i=1}^n$ is a martingale difference sequence w.r.t. filtration $\{\mathcal{F}_i\}_{i=1}^n$ where $\mathcal{F}_i$ is a sigma algebra of the history up to time i, so We do not notice this severe bias for probability constrained TS (not displayed), which performed similarly to ϵ-greedy. $\mathcal{F}_i = \sigma(\{\mathbf{x}_i, y_i, \epsilon_i : i < n\})$
- $\sup_n E[|\epsilon_n|^\alpha | \mathcal{F}_{n-1}] < M < \infty$ a.s. for some $\alpha > 2$
- $\lim_{n\to\infty} E[\epsilon_n^2 | \mathcal{F}_{n-1}] = \sigma^2$ a.s. for some constant σ^2
- $\mathbf{x}_n \in \mathbb{R}^p$ is $\mathcal{F}_{n-1}$ measurable
- There exists A_n, a sequence of nonrandom positive definite symmetric matrices such that

$$A_n^{-1} (\sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T)^{1/2} \xrightarrow{P} I_p \tag{1}$$

$$\max_{1 \leq k \leq n} ||A_n^{-1} \mathbf{x}_k||_2 \xrightarrow{P} 0 \tag{2}$$

In our simple bandit setting, the covariance matrix $\sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T$ is a diagonal matrix with the number of times each arm is sampled along the diagonal. Thus (1) is a stability condition on the rate at which the arms are sampled and (2) roughly ensures that all arms are sampled infinitely often.

Sampling Algorithms

We show for that a fixed non-zero margin $\beta_1 - \beta_2$, epsilon greedy and Thompson Sampling with Gaussian priors and constrained action probabilities (the probability of selecting a particular arm cannot drop below some $\pi_{\min} > 0$ specified by domain science) will satisfy these conditions.

Adapting Lai and Wei's CLT to Pooling Setting

In the mobile health setting, studies are generally short in duration. Thus, to get a confidence interval for the treatment effect, we would like to analyze the OLS estimator's asymptotic distribution as the number of users $m \to \infty$ with T fixed.

Note that we cannot directly apply Lai and Wei's result because our pooled setting requires a triangular array version of the result. For example, suppose that $T = 2$, note how our data changes as the number of users m increases,

$\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}$
$\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \mathbf{x}_2^{(1)} \mathbf{x}_2^{(2)}$
$\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \mathbf{x}_1^{(3)}, \mathbf{x}_2^{(1)}, \mathbf{x}_2^{(2)}, \mathbf{x}_2^{(3)}$

Since the distribution of H_1 changes with m, thus our distribution of our actions $\mathbf{x}_2^{(u)} \sim \mathcal{A}(H_1)$ will change with m as well. Thus, the ϵ_i do not form a martingale difference sequence, as required by Lai and Wei's theorem. The primary contribution of this project is proving a triangular array version of Lai and Wei's CLT that is applicable to the pooled data setting.

Simulation Results

We run simulations for the non-pooling case to empirically examine Lai and Wei's CLT result for small samples. For all the simulations, arm 1 is optimal. Arm 2 has mean reward 0 and arm 1's mean reward is exactly the margin or gap between the mean rewards for the arms. The reward noise are drawn from $\mathcal{N}(0, 1)$. For Thompson Sampling (TS), we use standard normal priors on each arm. We compare results for samples collected using an adaptive strategy (ϵ-greedy or TS) and compare to drawing independent samples, using the same number of samples per arm.

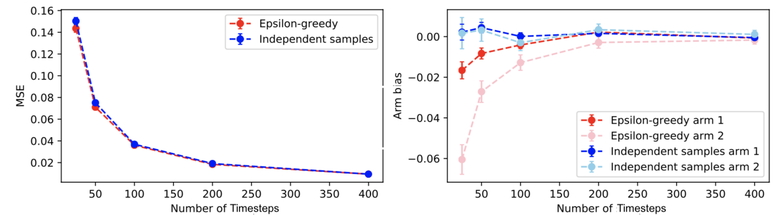

Figure 1. Epsilon Greedy (epsilon = 0.3, margin=1, n=4000)

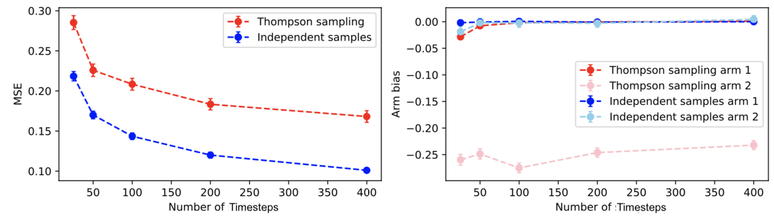

Figure 2. Thompson Sampling (pi_min=0, margin=1, n=4000)

Strategy	T	Margin	MSE	Margin bias	% margin in CI (nom 95%)	Prob sample arm 2
ϵ-greedy	100	1	0.036 ± 0.0007	0.087 ± 0.004	95.6 ± 0.3	0.15
ϵ-greedy	100	0.5	0.035 ± 0.0007	0.011 ± 0.004	95.2 ± 0.3	0.16
ϵ-greedy	100	0.25	0.035 ± 0.0007	0.013 ± 0.004	95.0 ± 0.3	0.19
TS	100	1	0.208 ± 0.007	0.273 ± 0.009	95.1 ± 0.3	0.011
TS	100	0.5	0.167 ± 0.007	0.272 ± 0.008	94.2 ± 0.4	0.037
TS	100	0.25	0.148 ± 0.006	0.196 ± 0.008	93.2 ± 0.4	0.092

We notice that the arm sample mean rewards have a negative bias when sampling using the adaptive strategies, especially for smaller sample sizes [3]. The negative bias is particularly severe for TS, for which the probability of sampling the suboptimal arm converges to zero quickly. Similar to the simulation results of [1], we find that as the margin gets smaller, for TS the empirical coverage of the confidence interval (CI) for the margin begins to drop below the nominal level. We do not notice this severe bias for probability constrained TS (not displayed), which performed similarly to ϵ-greedy.

References

[1] Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy.
Accurate inference for adaptive linear models.
International Conference on Machine Learning, 2018, 2018.

[2] Tze Leung Lai and Ching Zong Wei.
Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems.
The Annals of Statistics, 1982.

[3] 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 (AISTATS), 2018.