Online Statistical Inference in Decision-Making with Matrix Context



Qiyu Han   Will Wei Sun   Yichen Zhang

Purdue University

Annals of Statistics accepted paper, *Journal-to-Conference* Track

NeurIPS 2025

Objective in Sequential Decision-Making

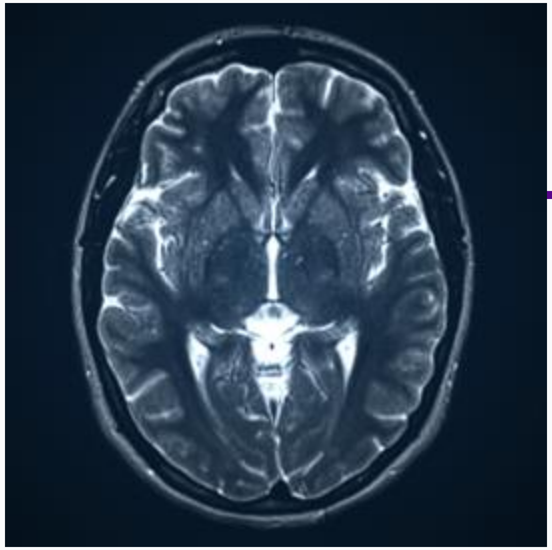

- Regret Minimization
  - Bandit algorithms are designed for regret minimization: how much worse it performs compared to an offline oracle

- In some cases, we also care about statistical inference or uncertainty quantification
  - Infer the effect of an arm / difference of effect between two bandit arms
    - Under-performing arms could be eliminated or modified, High-performing arms could be studied further, and One arm might be better than another
  - Gaining generalizable knowledge; Identifying new directions
  - Crucial for scientific discovery

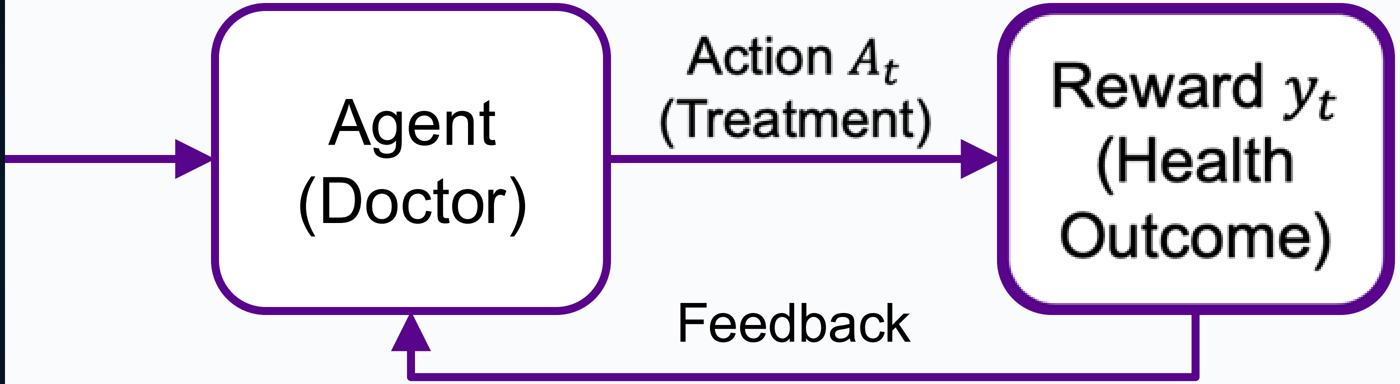
Decision-Making with Matrix Context

Some applications require the context information stored in a matrix

Matrix Context: X_t



Binary action: $A_t \in \{0,1\}$, depend on the history



- Reward $y_t = A_t \langle M_1^*, X_t \rangle + (1 - A_t) \langle M_0^*, X_t \rangle + \xi_t, \qquad A_t \in \{0,1\}$
- $\widehat{M}_{i,t-1}$ is the estimated parameter for selecting the action
- ε-greedy policy, for $\varepsilon \in (0,1)$, exploration-exploitation tradeoff.

$$A_t \sim Ber(\pi_t), \quad \pi_t = (1 - \varepsilon) I\{\langle \widehat{M}_{1,t-1}, X_t \rangle > \langle \widehat{M}_{0,t-1}, X_t \rangle\} + \frac{\varepsilon}{2}$$

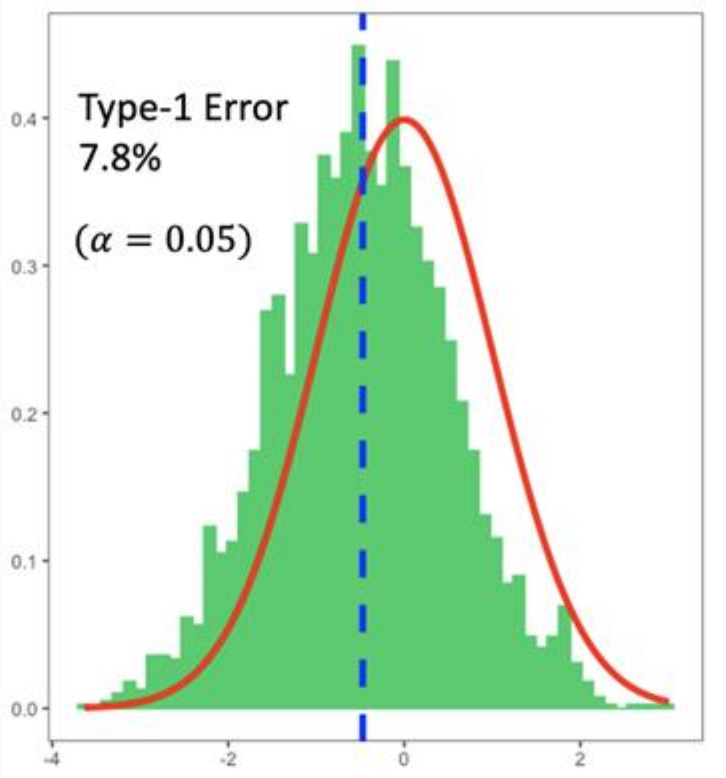
Inference - Bandit Algorithms Induce Dependence

- Observations (X_t, A_t, y_t) are not independent over $t \in \{1,2,\dots,n\}$
  - We use past history to select the action A_t in the new context X_t
  - Bandit data is *adaptively collected*

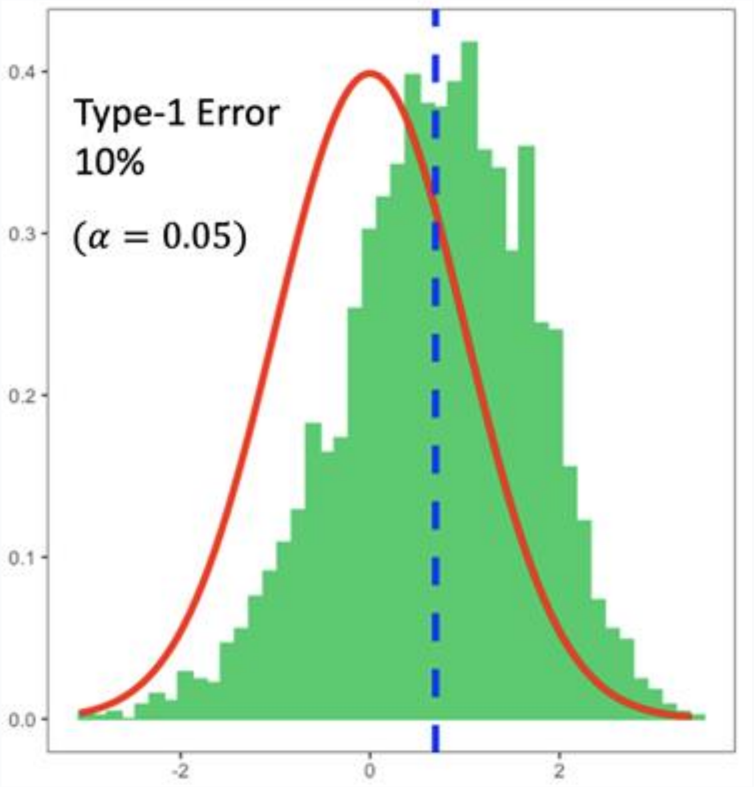
- Consequences for Statistical Inference e.g., test whether $(1,2)$ -entry of matrix M_1 is 0. $H_0 : M_1^*(1,2) = 0$
  - Violates independence assumptions of standard inference
  - Introduce a bias, or potentially non-normality
  - Villar, Bowden, Wason (2015, Stat. Sci.); Deshpande et al (2021, JASA); Khamaru, Mackey, Wainwright (2021); Zhang, Janson, Murphy (2021, 2022); Chen, Lu, Song (2021ab, JASA)
    - Not for matrix context, not fully-online algorithms.

Illustration and Comparison of Different Sources of Bias

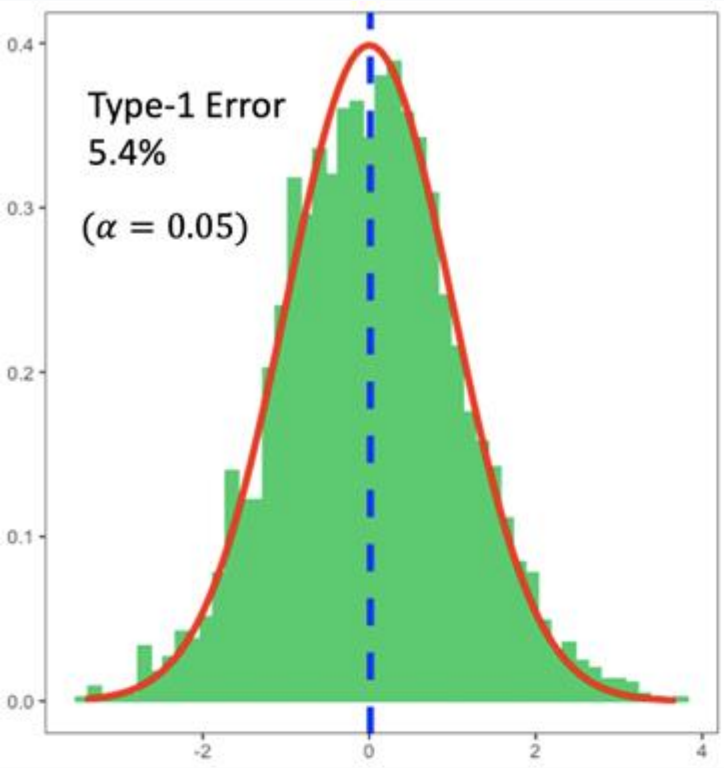
- In our problem, Bias in $\widehat{M}$ comes from two sources:



Bias caused by adaptivity of data collection.



Bias caused by nonconvex optimization.



Our doubly-debiased estimator handling both sources of bias.

Can we design estimators for M_1^*, M_0^* that are simultaneously:

sample-efficient, computation-efficient, & unbiased?

Low-rank matrix

- Sequential/Fully online.
- Non-convexity.
- Noisy samples.
- Desire sample-efficient convergence rate, $O_P(\sqrt{dr/t})$

Fully-Online

- Fast computation
- Minimal storage

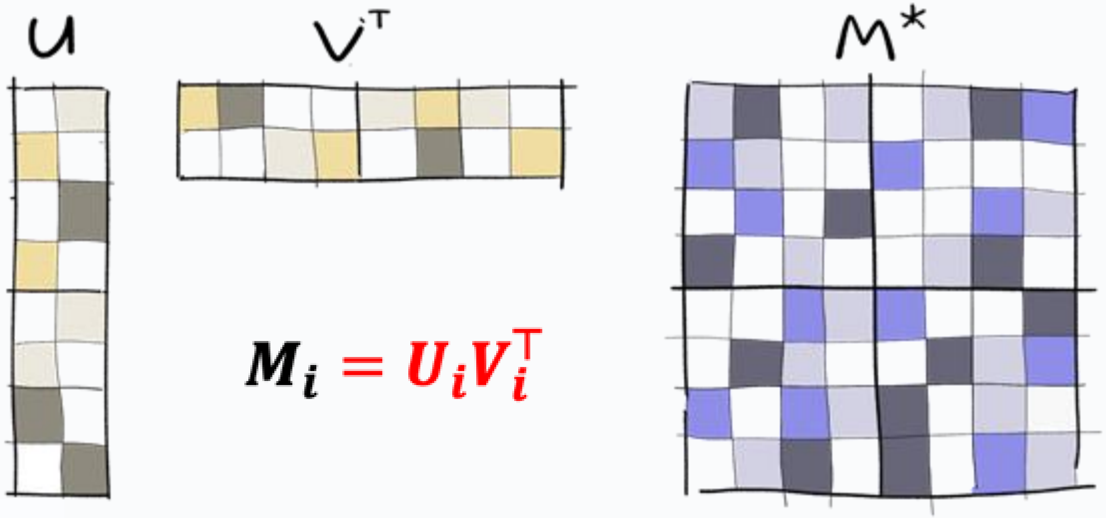
Inference

- Doubly-debiasing.
- Traditional analysis not applicable
- Consistent variance estimator

Low-rank Matrix (Trace) Regression for Each Arm

- Arm i: Reward $y_t = \langle M_i^*, X_t \rangle + \xi_t$
- Minimize the squared loss

$$\min_{\mathcal{U}_i \in \mathbb{R}^{d_1 \times r}, \mathcal{V}_i \in \mathbb{R}^{d_2 \times r}} \frac{1}{2} \sum_{t=1}^{n} (y_t - \langle \mathcal{U}_i \mathcal{V}_i^\top, X_t \rangle)^2$$



- Matrix Regression: Population Version

$$\min_{\mathcal{U}_i \in \mathbb{R}^{d_1 \times r}, \mathcal{V}_i \in \mathbb{R}^{d_2 \times r}} \mathbb{E}_{(X,y) \sim \mathsf{P}_{X,y}} \underline{f(\mathcal{U}_i, \mathcal{V}_i; (X, y))}$$

Not convex in (U_i, V_i)!

$$f(\mathcal{U}_i, \mathcal{V}_i; (X, y)) := \frac{1}{2}(y - \langle \mathcal{U}_i \mathcal{V}_i^\top, X \rangle)^2$$

- Given a local initialization (e.g., Candès and Plan, 2010), an SGD Update for Arm 1:

$$\begin{pmatrix} \mathcal{U}_{i,t} \\ \mathcal{V}_{i,t} \end{pmatrix} = \begin{pmatrix} \mathcal{U}_{i,t-1} \\ \mathcal{V}_{i,t-1} \end{pmatrix} - \eta_t \begin{pmatrix} (\langle \mathcal{U}_{i,t-1} \mathcal{V}_{i,t-1}^\top, X_t \rangle - y_t) X_t \mathcal{V}_{i,t-1} \\ (\langle \mathcal{U}_{i,t-1} \mathcal{V}_{i,t-1}^\top, X_t \rangle - y_t) X_t^\top \mathcal{U}_{i,t-1} \end{pmatrix}.$$

$\eta_t = t^{-\alpha}$ step-size

SGD Update: we only update U_i, V_i for arm i with $i = a_t$

Bandit SGD Update

Renormalized SGD Update:

$$\tilde{\mathcal{U}}_{a_t,t-1} = W_{\mathcal{U}} D^{\frac{1}{2}} \text{ and } \tilde{\mathcal{V}}_{a_t,t-1} = W_{\mathcal{V}} D^{\frac{1}{2}}, \text{ where } W_{\mathcal{U}} D W_{\mathcal{V}}^{\top} \text{ is the top-}r \text{ SVD of } \mathcal{U}_{a_t,t-1}\mathcal{V}_{a_t,t-1}^{\top}.$$

$$\begin{pmatrix} \mathcal{U}_{i,t} \\ \mathcal{V}_{i,t} \end{pmatrix} = \begin{pmatrix} \tilde{\mathcal{U}}_{i,t-1} \\ \tilde{\mathcal{V}}_{i,t-1} \end{pmatrix} - \eta_t \frac{I\{a_t = i\}}{i\pi_t + (1-i)(1-\pi_t)} \nabla f(\tilde{\mathcal{U}}_{i,t-1}, \tilde{\mathcal{V}}_{i,t-1}; \{X_t, y_t\}).$$

Theorem 1

For any constant $\alpha \in (0.5, 1)$, under some mild conditions and $n = o(d^{\gamma})$ when $d \to \infty$, with high probability we have

$$\left\| \widehat{M}_{i,t}^{\mathrm{sgd}} - M_i^* \right\|_{\mathrm{F}} \le C\gamma\sigma_i \sqrt{\frac{dr\log^2 d}{t^{\alpha}}}.$$

for any $1 \le t \le n$ and some positive constant C.

- Where $d = \max\{d_1, d_2\}$ for d_1 and d_2 are the dimensions of M_i^*.
- r is the rank of M_i^* assumed known, and $r \ll \min\{d_1, d_2\}$.

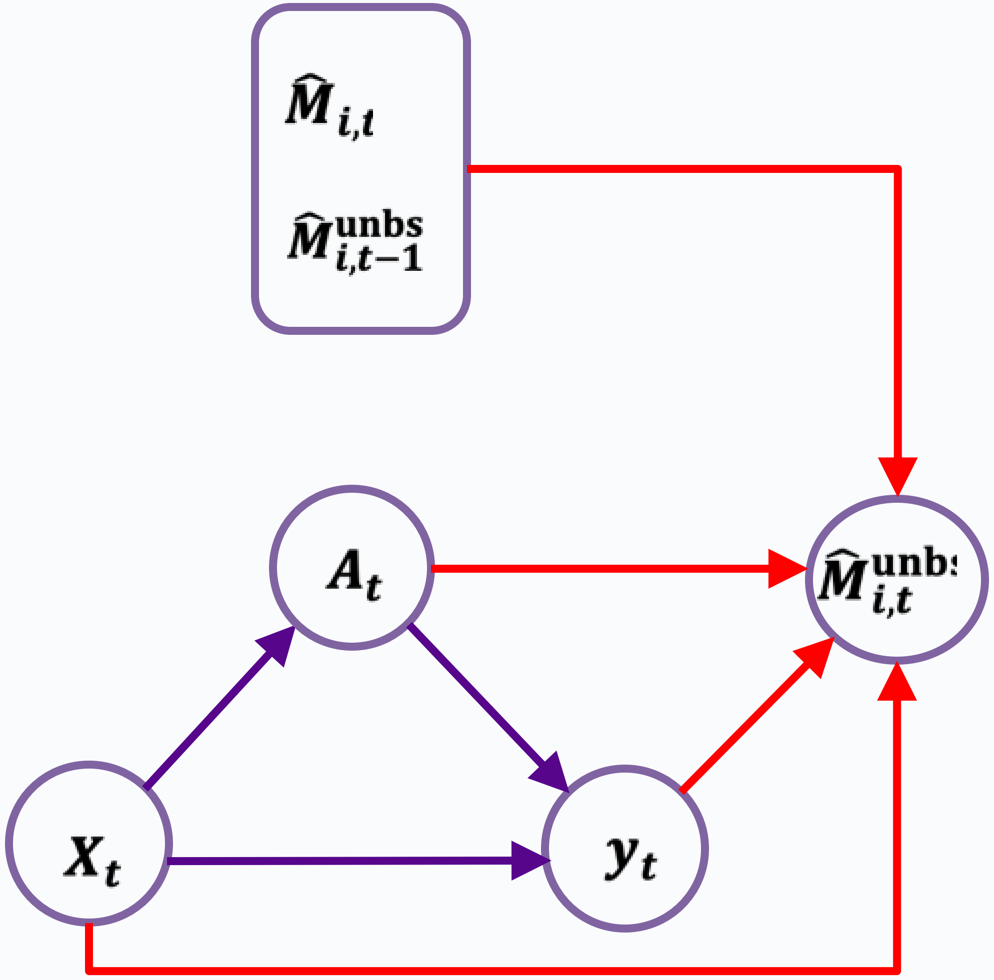
Our Method: Step 2 (Sequential Debiasing)

- At time t, taking $i = 1$ for example, we debias

$$\widetilde{M}_{1,t} = \widehat{M}_{1,t-1} - \underbrace{\frac{I\{a_t = 1\}}{\pi_t}}_{\text{Adaptive data collection}} \underbrace{\left(\langle \widehat{M}_{1,t-1}, X_t \rangle - y_t\right) X_t}_{\text{Low-Rankness}}$$

- Running average: $\widehat{M}^{\text{unbs}}_{1,t} = (\widetilde{M}_{1,t} + (t-1)\widehat{M}^{\text{unbs}}_{1,t-1})/t$
- Spectral Projection:

  Project $\widehat{M}^{\text{unbs}}_{i,t}$ to their leading singular subspaces (at the cost of negligible biases)



Estimator with Asymptotic Normality

- Parameter Inference

  inference for the $(1,2)$ entry of M_i

  $\widehat{m}_T^{(i)} = \langle M_i, T \rangle$, e.g. $T = e_1 e_2^\top$

- Optimal Policy-Value Inference

  for optimal value $V^* = E_X\left[\langle M_{\pi^*(X)}, X \rangle\right]$

Theorem 1: Asymptotic Normality (For i = 1)

Under mild conditions, with the presented decision-making model, as online sample size $n \to \infty$. Then

$$\sqrt{n}\left(\widehat{m}_T^{(1)} - m_T^{(1)}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 S_1^2\right), \text{ where}$$

$$S_1^2 = \int \frac{\left(\left\langle U_\perp U_\perp^\top X V V^\top, T \right\rangle + \left\langle U U^\top X V_\perp V_\perp^\top, T \right\rangle\right)^2}{(1-\varepsilon_\infty) I\left\{\langle M_1 - M_0, X \rangle > 0\right\} + \frac{\varepsilon_\infty}{2}} dP_X.$$

Theorem 2: Asymptotic Normality for $\widehat{V}_n$

Under mild conditions, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{V}_n - V^*\right) \xrightarrow{d} \mathcal{N}(0, s^2), \text{ where}$$

$$s^2 = \int \sum_{i=0}^{1} \sigma_i I\{\langle M_i, X \rangle > \langle M_{1-i}, X \rangle\} dP_X + \mathbb{E}\left[\langle M_{\pi^*(X)}, X \rangle^2\right].$$

Summary and Future Work

- We propose a **sample-efficient**, **computational efficient**, & **unbiased** fully-online inference procedure for low-rank matrix contextual bandit with **adaptive data collection**.
- Two inference procedures for **parameter inference**, and **optimal policy value** inference.
- Qiyu Han, Will Wei Sun, and Yichen Zhang. *Online Statistical Inference in Decision-Making with Matrix Context*. *Annals of Statistics*, to appear. arXiv: 2212.11385.

Thank You !