

# **Online Policy Learning and Inference by Matrix Completion**

**Duan, Li and Xia (2024)**

**Jun 21 2024**

# Contents

- Motivation Examples
- Methodology
- Theoretical Results
- Simulation Studies
- Real Data Analysis

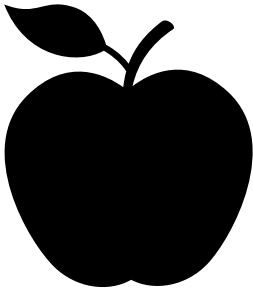


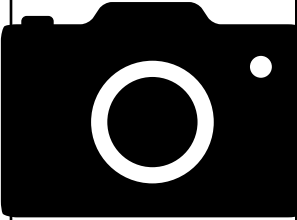
# A Walmart Discount Story

- A manager at Walmart aims to decide whether to offer discounts on  $d_1$  products at  $d_2$  time point.
- A discount: reduce per-unit profit v.s. increase overall sales.
- Goal: a policy can maximize total profits.
- Question: **When** to put discounts on **what** product?

# A Walmart Discount Story

## Mathematical Formulation

- Consider two profit matrices:
- $M_1$ : Profit matrix with discount

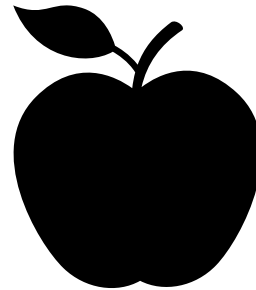


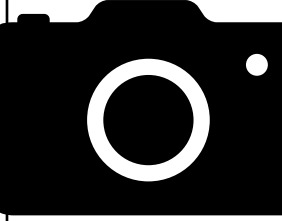
	Mon	Tue	Wed	Thr	Fri	Sat	Sun
							
							
							
							

# A Walmart Discount Story

## Mathematical Formulation

- Consider two profit matrices:
- $M_1$ : Profit matrix with discount

On Wednesday,  
The total profit from selling guitars with  
discount is \$4000

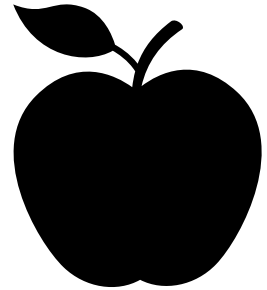
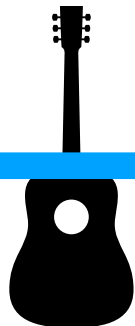

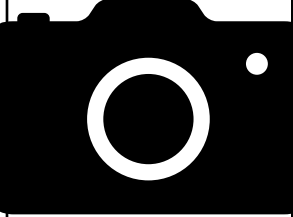
	Mon	Tue	Wed	Thr	Fri	Sat	Sun
							
			\$4000				
							
							

# A Walmart Discount Story

## Mathematical Formulation

- Consider two profit matrices:
- $M_0$ : Profit matrix without discount

On Wednesday,  
The total profit from selling guitars without  
discount is \$2000

	Mon	Tue	Wed	Thr	Fri	Sat	Sun
							
			\$2000				
							
							

# A Walmart Discount Story

## Mathematical Formulation

- A bandit problem with two arms (w/o discount)
- Suppose a customer arrives on Wednesday to buy a guitar:
- Push arm 1 (have a discount) if  $M_1(\text{Wed, Guitar}) > M_0(\text{Wed, Guitar})$ .
- Approaches for estimating two matrices are needed.

# A San Francisco Parking Story

- SF government plans to implement dynamic pricing for parking lots across  $d_1$  blocks during  $d_2$  time periods.
- High price: reduce parking on overcrowded blocks.
- Low price: attract parking on undercrowded blocks.
- Goal: Achieve moderate occupancy rates across more blocks throughout major time periods.
- Question: **when** and **where** to put the high parking price?



# A San Francisco Parking Story

## Mathematical Formulation

- A matrix describing the target occupancy rate that is given:
- - The government hopes to Control the occupancy rate at The Castro To be around 70%

	7am	8am	9am	10am	11am	...	7pm
Haight Street							
The Castro		70%					
Union Square							
Mission Street							

# A San Francisco Parking Story

## Mathematical Formulation

- Consider two deviation matrices.
- $M_1$ : the deviation matrix under high pricing strategy.

The occupancy rate at The Castro  
Under high parking prices  
Deviates the 'ideal' 70% by 10% (in abs value)


	7am	8am	9am	10am	11am	...	7pm
Haight Street							
The Castro		10%					
Union Square							
Mission Street							

# A San Francisco Parking Story

## Mathematical Formulation

- Consider two deviation matrices.
- $M_2$ : the deviation matrix under low pricing strategy.

The occupancy rate at The Castro  
Under low parking prices  
Deviates the 'ideal' 70% by 5% (in abs value)



	7am	8am	9am	10am	11am	...	7pm
Haight Street							
The Castro		5%					
Union Square							
Mission Street							

# A San Francisco Parking Story

## Mathematical Formulation

- A bandit problem with two arms: (high/low parking price.)
- Suppose a car wants to park at the Castro at 8 am:
- Push arm0 if  $M_0(\text{Castro}, 8\text{am}) < M_0(\text{Castro}, 8\text{am})$ .
- Approaches for estimating two matrices are needed.
-

# Matrix Completion Bandit: Problem Formulation

# Problem Formulation

- Consider a sequence of random pairs  $\{r_\tau, X_\tau\}_{\tau=1}^t$ , where  $r_\tau \in \mathbb{R}$  and  $X_\tau$  uniformly sampled from  $\mathcal{E} = \{e_j e_k^T, j \in [d_1], k \in [d_2]\}$ .
- $r_\tau$ : time  $\tau$  reward;  $X_\tau$ : time  $\tau$  request.
- $K$  arms associated with a matrix  $M_k \in \mathbb{R}^{d_1 \times d_2}$ .
- $r_\tau = \text{tr}(M_{a_\tau}^T X_\tau) + \xi_\tau := \langle M_{a_\tau}, X_\tau \rangle + \xi_\tau$ .
- At each time  $\tau$ , only a noisy entry of  $M$  can be observed.
- Optimal policy:  $\arg\max_{k \in 1, 2 \dots K} [M_k]_{j_1, j_2}$ .
- $d_1 d_2 \gg T$
- $\text{rank}(M) \ll T$ .

# Problem formulation

- A two-armed matrix completion bandit problem.
- Online algorithm for matrix estimation.
- $\epsilon$ -greedy policy and regret analysis.
- Policy Inference procedure.

# Methodology



# Methodology

- Two armed bandit with  $M_0$  and  $M_1 \in \mathbb{R}^{d_1 \times d_2}$ , both are of rank  $r$ .
- SVD:  $M_i = L_i \Lambda_i R_i^T$ . ( $i = 0, 1$ )
- $L_i$  and  $R_i$ :  $d_1$  by  $r$  and  $d_2$  by  $r$  orthogonal matrices.
- $\Lambda_i$ :  $r$  by  $r$  diagonal matrix.
- $U_i := L_i \Lambda_i^{1/2}$ ,  $V_i := R_i \Lambda_i^{1/2}$ .
- $M_i = U_i V_i^T$ .

# Methodology

## An offline approach

- Let  $\pi_\tau = P(a_\tau = 1)$ .
- $M_1$  estimated by:
- $$\min_{U,V} \mathcal{L}_{1,t}^\pi(U, V) = \sum_{\tau=1}^t \frac{1\{a_\tau = 1\}}{\pi_\tau} (r_\tau - \langle X_\tau, UV^T \rangle)^2,$$
- Subject to  $U^T U = V^T V$ .
- $M_0$  estimated similarly.

# Methodology

## An $\epsilon$ -greedy online approach

- At the time  $\tau - 1$ , some key elements are needed:
- The previous estimators  $\hat{M}_{0,\tau-1}$  and  $\hat{M}_{1,\tau-1}$ .
- The exploration probability  $\epsilon_\tau$  at time  $\tau$ .
- The updating steplength  $\eta_\tau$  at time  $\tau$ .
- $\pi_\tau = P(a_\tau = 1 \mid \mathcal{H}_{\tau-1}, \hat{M}_{i,\tau-1}, i = 0,1) = (1 - \epsilon_\tau)1\{\langle \hat{M}_{1,\tau-1} - \hat{M}_{0,\tau-1}, X_\tau \rangle > 0\} + \frac{\epsilon_\tau}{2}$

# Methodology

## An $\epsilon$ -greedy online approach

- Let  $l_{1,\tau}^\pi(U, V) = \frac{1\{a_\tau = 1\}}{\pi_\tau} (r_\tau - \langle X_\tau, UV^T \rangle)^2$ .
- $\tilde{U}_{1,\tau} = \hat{U}_{1,\tau-1} - \eta_t \nabla_U l_{1,\tau}^\pi(\hat{U}_{1,\tau-1}, \hat{V}_{1,\tau-1})$
- $\tilde{V}_{1,\tau} = \hat{V}_{1,\tau-1} - \eta_t \nabla_V l_{1,\tau}^\pi(\hat{V}_{1,\tau-1}, \hat{V}_{1,\tau-1})$
- $\hat{M}_{1,\tau} = \tilde{U}_{1,\tau} \tilde{V}_{1,\tau}^T$
- Obtain  $\hat{U}_{1,\tau}, \hat{V}_{1,\tau}$  by doing SVD for  $\hat{M}_{1,\tau}$ .

# Methodology

## An $\epsilon$ -greedy online approach

---

**Algorithm 1**  $\epsilon$ -greedy two-arm MCB with online gradient descent

---

**Input:** exploration probabilities  $\{\epsilon_t\}_{t \geq 1}$ ; step sizes  $\{\eta_t\}_{t \geq 1}$ ; initializations with balanced factorization  $\widehat{M}_{0,0} = \widehat{U}_{0,0}\widehat{V}_{0,0}^\top$ ,  $\widehat{M}_{1,0} = \widehat{U}_{1,0}\widehat{V}_{1,0}^\top$

**Output:**  $\widehat{M}_{0,T}$ ,  $\widehat{M}_{1,T}$ .

**for**  $t = 1, 2, \dots, T$  **do**

    Observe a new request  $X_t$

    Calculate  $\pi_t = (1 - \epsilon_t)\mathbb{1}(\langle \widehat{M}_{1,t-1} - \widehat{M}_{0,t-1}, X_t \rangle > 0) + \frac{\epsilon_t}{2}$

    Sample an action  $a_t \sim \text{Bernoulli}(\pi_t)$  and get a reward  $r_t$

**if**  $a_t = 1$  **then**

        Update by

$$\begin{pmatrix} \widetilde{U}_{1,t} \\ \widetilde{V}_{1,t} \end{pmatrix} = \begin{pmatrix} \widehat{U}_{1,t-1} \\ \widehat{V}_{1,t-1} \end{pmatrix} - \frac{\eta_t}{\pi_t} \cdot \begin{pmatrix} (\langle \widehat{U}_{1,t-1}\widehat{V}_{1,t-1}^\top, X_t \rangle - r_t)X_t\widehat{V}_{1,t-1} \\ (\langle \widehat{U}_{1,t-1}\widehat{V}_{1,t-1}^\top, X_t \rangle - t_t)X_t^\top\widehat{U}_{1,t-1} \end{pmatrix},$$

        Set  $\widehat{U}_{1,t} = \widehat{L}_{1,t}\widehat{\Lambda}_{1,t}^{1/2}$  and  $\widehat{V}_{1,t} = \widehat{R}_{1,t}\widehat{\Lambda}_{1,t}^{1/2}$ , where  $\widehat{L}_{1,t}\widehat{\Lambda}_{1,t}\widehat{R}_{1,t}^\top$  is the thin SVD of  $\widehat{M}_{1,t} = \widetilde{U}_{1,t}\widetilde{V}_{1,t}^\top$ .

**else**

        Update by

$$\begin{pmatrix} \widetilde{U}_{0,t} \\ \widetilde{V}_{0,t} \end{pmatrix} = \begin{pmatrix} \widehat{U}_{0,t-1} \\ \widehat{V}_{0,t-1} \end{pmatrix} - \frac{\eta_t}{1 - \pi_t} \cdot \begin{pmatrix} (\langle \widehat{U}_{0,t-1}\widehat{V}_{0,t-1}^\top, X_t \rangle - r_t)X_t\widehat{V}_{0,t-1} \\ (\langle \widehat{U}_{0,t-1}\widehat{V}_{0,t-1}^\top, X_t \rangle - t_t)X_t^\top\widehat{U}_{0,t-1} \end{pmatrix},$$

        Set  $\widehat{U}_{0,t} = \widehat{L}_{0,t}\widehat{\Lambda}_{0,t}^{1/2}$  and  $\widehat{V}_{0,t} = \widehat{R}_{0,t}\widehat{\Lambda}_{0,t}^{1/2}$ , where  $\widehat{L}_{0,t}\widehat{\Lambda}_{0,t}\widehat{R}_{0,t}^\top$  is the thin SVD of  $\widehat{M}_{0,t} = \widetilde{U}_{0,t}\widetilde{V}_{0,t}^\top$ .

**end if**

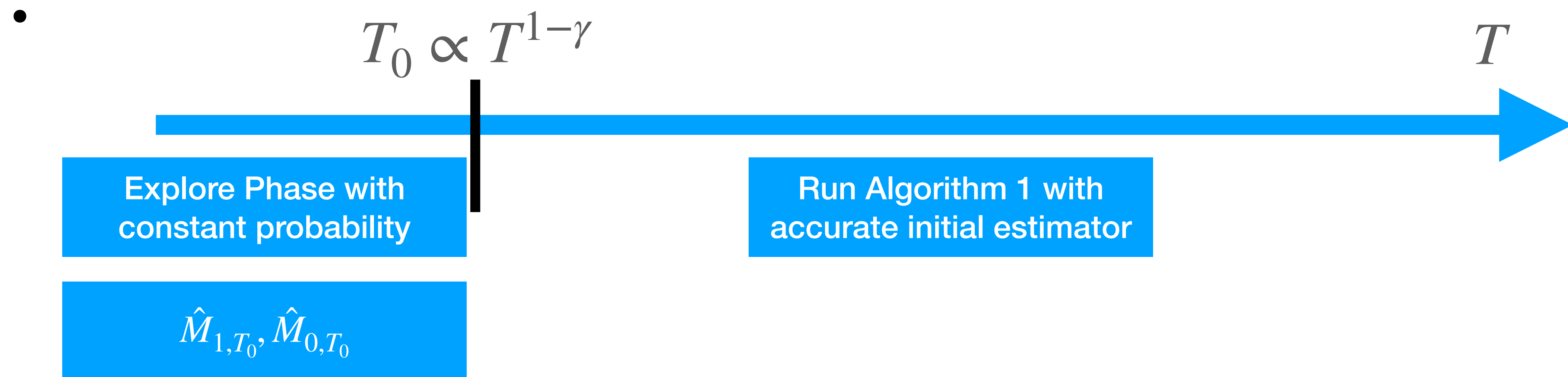
**end for**

---

# Methodology

## An $\epsilon$ -greedy online approach

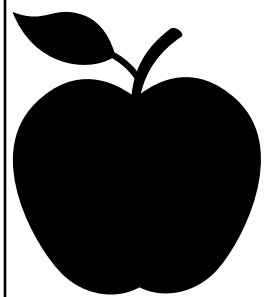
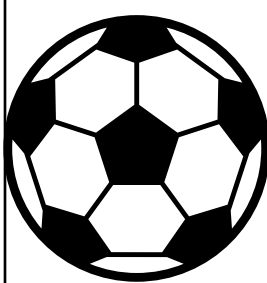
- How to obtain initial estimators?
- Explore-Then-Commit Scheme.
- Time horizon:  $T$ .



# Methodology

## Policy Inference

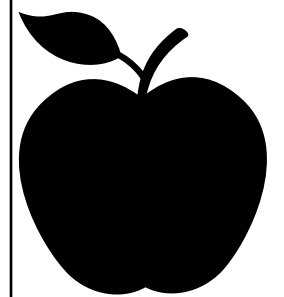
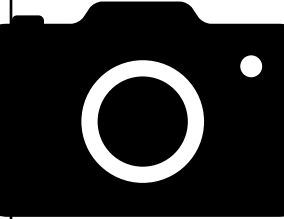
- A manager wants to decide on a marketing strategy for a group of requests.
- Confidence in the decision's correctness.
- 

	Mon	Tue	Wed	Thr	Fri	Sat	Sun
	2						
			3				
							
							

# Methodology

## Policy Inference

- A manager wants to decide on a marketing strategy for a group of requests.
- Confidence in the decision's correctness.
- $Q = 2e_1e_1^T + 3e_2e_3^T$ .
- $H_0 : \langle M_1 - M_0, Q \rangle = 0$ .
- $H_1 : \langle M_1 - M_0, Q \rangle > 0$ .

	Mon	Tue	Wed	Thr	Fri	Sat	Sun
	2						
			3				
							
							



# Methodology

## Policy Inference: Debiasing

$$\bullet \hat{M}_0^{IPW} = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{M}_{0,t-1} + \frac{d_1 d_2}{T - T_0} \sum_{t=T_0+1}^T \frac{1\{a_t = 0\}}{1 - \pi_t} (r_t - \langle \hat{M}_{0,t-1}, X_t \rangle) X_t.$$

$$\bullet \hat{M}_1^{IPW} = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{M}_{1,t-1} + \frac{d_1 d_2}{T - T_0} \sum_{t=T_0+1}^T \frac{1\{a_t = 1\}}{\pi_t} (r_t - \langle \hat{M}_{1,t-1}, X_t \rangle) X_t.$$

# Methodology

## Policy Inference: Debiasing

- $$\hat{M}_0^{IPW} = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{M}_{0,t-1} + \frac{d_1 d_2}{T - T_0} \sum_{t=T_0+1}^T \frac{1\{a_t = 0\}}{1 - \pi_t} (r_t - \langle \hat{M}_{0,t-1}, X_t \rangle) X_t.$$

Base estimator

Bias correction term: a rotated score of the

- Loss function: 
$$\frac{1}{T - T_0} \sum_{t=1}^T \frac{1\{a_t = 0\}}{1 - \pi_t} (r_t - \langle \hat{M}_{0,t-1}, X_t \rangle)^2.$$

# Methodology

## Policy Inference: Debiasing

- Consider the rank constrain of  $M_i$  :
- $\hat{L}_1, \hat{R}_1$ : matrices formed by top-r left(right) singular vectors of  $\hat{M}_1^{IPW}$ .
- $\hat{M}_1 = \hat{L}_1 \hat{L}_1^T \hat{M}_1^{IPW} \hat{R}_1 \hat{R}_1^T$ .
- Asymptotic distribution of  $\langle \hat{M}_1 - \hat{M}_0, Q \rangle$  TBD.

# Theoretical Results

# Theoretical Results

## Notation

- $\|\cdot\|$ :  $\ell_2$  norm for vectors and spectral norm for matrices.
- $\|\cdot\|_F$ : Frobenius norm.
- $\|\cdot\|_{\max}$ : maximum absolute entry value.
- $\|\cdot\|_{2, \max}$ : maximum row-wise  $\ell_2$  norm.
- $\mathbb{O}^{d \times r} := \{U \in \mathbb{R}^{d \times r}; U^T U = I_r\}$ .
- $U_{\perp}$ : Orthogonal complement of  $U$ .

# Theoretical Results

## Notation

- $\kappa := \frac{\lambda_{\max}}{\lambda_{\min}}.$
- Incoherence condition: Recall  $M_i = L_i \Lambda_i R_i$ ,  $i = 0, 1$
- $\mu(M_i) := \max \left\{ \sqrt{\frac{d_1}{r}} \|L_i\|_{2,\max}, \sqrt{\frac{d_2}{r}} \|R_i\|_{2,\max} \right\}.$
- $\max \{ \mu(M_0), \mu(M_1) \} \leq \mu_0.$
- $\mu_0, \kappa$  are bounded constants.
- $\sigma_0^2 = \text{Var}(\xi_{0,t}), \sigma_1^2 = \text{Var}(\xi_{1,t}).$

# Theorem 1

## Matrix Estimation

- Assume the following conditions ( $c_0, C_1, \dots, C_4$  are some constants):
- $T \leq d_1^{100}$ .
- $\|\hat{M}_{0,0} - M_0\|_F + \|\hat{M}_{1,0} - M_1\|_F \leq c_0 \lambda_{\min}$ .
- For any  $t = 1, 2, \dots, T$  :
  - $\min \left\{ \frac{\lambda_{\min}^2}{\sigma_0^2 + \sigma_1^2}, d_1 d_2 \log d_1 \right\} \geq C_1 \sum_{\tau=1}^t \frac{(\eta_{\tau} \lambda_{\max})^2}{\epsilon_{\tau}} \frac{r \log^2 d_1}{d_2}$ .
  - $\max_{\tau \in [t]} \frac{\eta_{\tau}^2}{\epsilon_{\tau}^2} \leq C_2 \sum_{\tau=1}^t \frac{\eta_{\tau}^2}{\epsilon_{\tau}}$ .

# Theorem 1

## Matrix Estimation

- With probability at least  $1 - 8td^{-200}$ :
- $$\|\hat{M}_{i,t} - M_i\|_F^2 \leq C_3 \|\hat{M}_{i,0} - M_i\|_F^2 \prod_{\tau=1}^t \left(1 - \frac{\eta_\tau \lambda_{\min}}{4d_1 d_2}\right) + C_4 \sigma_i^2 \frac{r \log^2 d_1}{d_2} \sum_{\tau=1}^t \frac{(\eta_\tau \lambda_{\max})^2}{\epsilon_\tau}$$
- $$\|\hat{M}_{i,t} - M_i\|_{\max} \leq C_3 \frac{\lambda_{\min}^2 r^3}{d_1 d_2} \prod_{\tau=1}^t \left(1 - \frac{\eta_\tau \lambda_{\min}}{4d_1 d_2}\right) + C_4 \sigma_i^2 \frac{r^3 \log^2 d_1}{d_1 d_2^2} \sum_{\tau=1}^t \frac{(\eta_\tau \lambda_{\max})^2}{\epsilon_t}$$



# Theorem 1

## Matrix Estimation

- With probability at least  $1 - 8td^{-200}$ :
- $\|\hat{M}_{i,t} - M_i\|_F^2 \leq C_3 \|\hat{M}_{i,0} - M_i\|_F^2 \prod_{\tau=1}^t \left(1 - \frac{\eta_\tau \lambda_{\min}}{4d_1 d_2}\right) + \text{Second term}$
- $\|\hat{M}_{i,t} - M_i\|_{\max} \leq C_3 \frac{\lambda_{\min}^2 r^3}{d_1 d_2} \prod_{\tau=1}^t \left(1 - \frac{\eta_\tau \lambda_{\min}}{4d_1 d_2}\right) + \text{Second term}$
- Correspond to the convergence of gradient descent.
- Depends on step size.

# Theorem 1

## Matrix Estimation

- With probability at least  $1 - 8td^{-200}$ :

- $\|\hat{M}_{i,t} - M_i\|_F^2 \leq \text{First Term} + C_4 \sigma_i^2 \frac{r \log^2 d_1}{d_2} \sum_{\tau=1}^t \frac{(\eta_{\tau} \lambda_{\max})^2}{\epsilon_{\tau}}$
- $\|\hat{M}_{i,t} - M_i\|_{\max} \leq \text{First Term} + C_4 \sigma_i^2 \frac{r^3 \log^2 d_1}{d_1 d_2^2} \sum_{\tau=1}^t \frac{(\eta_{\tau} \lambda_{\max})^2}{\epsilon_{\tau}}$

- Correspond to stochastic noise and random sampling,
- Depend on noise, exploration and stepsize.

# Corollary 1

## Matrix Estimation: A specific rate.

- Some extra assumptions:
- Fix  $\gamma \in [0,1)$ ,  $\epsilon \in (0, \frac{1}{2})$ .
- Exploration-Then-Commit Scheme:
- $T_0 = C_0 T^{1-\gamma} \log\{\lambda_{\min}(\sigma_0 \wedge \sigma_1)\}$ .
- (Explore) When  $t \leq T_0$ :  $\epsilon_t = \epsilon$ ,  $\eta_t = \eta := cd_1d_2/(T^{1-\gamma}\lambda_{\max})$ .
- (Commit) When  $T_0 \leq t \leq T$ :  $\epsilon_t = c_2 t^{-\gamma}$ ,  $\eta_t = \epsilon_t \eta$ .

# Corollary 1

**Matrix Estimation: A specific rate.**

- $T \geq C_1 r^3 d_1^{1/(1-\gamma)} \log^2 d_1, \frac{\lambda_{\min}^2}{\sigma_0^2 + \sigma_1^2} \geq C_2 \frac{r d_1^2 d_2 \log^2 d_1}{T^{1-\gamma}}.$
- We obtain the following rates:
- $\|\hat{M}_{i,T} - M_i\|_F^2 \leq C_3 \sigma_i^2 \frac{r d_1^2 d_2 \log^4 d_1}{T^{1-\gamma}},$
- $\|\hat{M}_{i,T} - M_i\|_{\max}^2 \leq C_3 \sigma_i^2 \frac{r d_1 \log^4 d_1}{T^{1-\gamma}}.$

# Corollary 1

**Matrix Estimation: A specific rate.**

- $\|\hat{M}_{i,T} - M_i\|_F^2 \leq C_3 \sigma_i^2 \frac{r d_1^2 d_2 \log^4 d_1}{T^{1-\gamma}},$
- $\|\hat{M}_{i,T} - M_i\|_{\max}^2 \leq C_3 \sigma_i^2 \frac{r d_1 \log^4 d_1}{T^{1-\gamma}}.$
- A Frobenius norm at  $\tilde{O}_p\left(\frac{r d_1^2 d_2}{T^{1-\gamma}}\right).$
- A sup norm rate at  $\tilde{O}_p(r d_1 / T^{1-\gamma}).$

# Corollary 1

## Matrix Estimation: A specific rate.

- A Frobenius norm at  $\tilde{O}_p(\frac{rd_1^2 d_2}{T^{1-\gamma}})$ .
- A sup norm rate at  $\tilde{O}_p(\frac{rd_1}{T^{1-\gamma}})$ .
- In the offline setting of Ma et al (2017):
- Assuming  $T \gg r^3 d_1 \log^3 d_1$ .
- The optimal (up to log factors) Frobenius rate:  $O_p(\frac{rd_1^2 d_2 \log d_1}{T})$ .
- The optimal (up to log factors) sup norm rate :  $O_p(\frac{rd_1 \log d_1}{T})$ .

Implicit Regularization in Nonconvex Statistical Estimation:  
Gradient Descent Converges Linearly for Phase Retrieval,  
Matrix Completion, and Blind Deconvolution

Cong Ma\*      Kaizheng Wang\*      Yuejie Chi<sup>†</sup>      Yuxin Chen<sup>‡</sup>

November 2017;    Revised July 2019

# Regret Analysis

## Key Ideas

- $R_T := \mathbb{E} \left[ \sum_{t=1}^T \max_{i \in \{0,1\}} \langle M_i, X_t \rangle - \langle M_{a_t}, X_t \rangle \right].$
- $R_T \lesssim \|M_1 - M_0\|_{\max} \sum_{t=1}^T \epsilon_t + \sum_{t=1}^T \max_{i \in \{0,1\}} \|\hat{U}_{i,t} \hat{V}_{i,t}^T - M_i\|_{\max}.$
- Wrong decision made by exploration + estimation error.

# Regret Analysis

## Theorem 2

- Based on conditions in Corollary 1.
- Define:
- $\bar{m} = \|M_0\|_{\max} + \|M_1\|_{\max}$ .
- $\bar{\sigma} = \max\{\sigma_0, \sigma_1\}$ .
- $R_T \leq C_5 \left[ \bar{m} r T^{1-\gamma} + \bar{\sigma} T^{(1+\gamma)/2} \sqrt{r d_1} \log^2 d_1 \right]$ .



# Regret Analysis

## Theorem 2 Remarks

- $R_T \lesssim \|M_1 - M_0\|_{\max} \sum_{t=1}^T \epsilon_t - \sum_{t=1}^T \max_{i \in \{0,1\}} \|\hat{U}_{i,t} \hat{V}_{i,t}^T - M_i\|_{\max}.$

- $R_T \leq C_5 \left[ \bar{m} r T^{1-\gamma} + \bar{\sigma} T^{(1+\gamma)/2} \sqrt{r d_1} \log^2 d_1 \right]$

- $\gamma = 0$ : Trivial bound  $O(T)$ .
- Pick  $\gamma$  such that  $T^{1-3\gamma} = (\bar{\sigma}/\bar{m})^2 d_1$  : A  $\tilde{O}(T^{2/3} d_1^{1/3})$  bound.

# Asymptotical Normality

- $\hat{M}_1^{IPW} = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \hat{M}_{1,t-1} + \frac{d_1 d_2}{T - T_0} \sum_{t=T_0+1}^T \frac{1\{a_t = 1\}}{\pi_t} (r_t - \langle \hat{M}_{1,t-1}, X_t \rangle) X_t.$
- $\hat{L}_1, \hat{R}_1$ : matrices formed by top-r left(right) singular vectors of  $\hat{M}_1^{IPW}$ .
- $\hat{M}_1 = \hat{L}_1 \hat{L}_1^T \hat{M}_1^{IPW} \hat{R}_1 \hat{R}_1^T.$
- Goal: Asymptotic distribution of  $\langle \hat{M}_1 - \hat{M}_0, Q \rangle.$

# Asymptotical Normality

## Extra Notation

- $M_1 = L_1 \Lambda_1 R_1^T, L_1 \in \mathbb{R}^{d_1 \times r}, R_1 \in \mathbb{R}^{d_2 \times r}.$
- $L_{1\perp} \in \mathbb{R}^{d_1 \times (d_1 - r)}, \text{col}(L_{1\perp}) = \text{col}^c(L_1).$
- $R_{1\perp} \in \mathbb{R}^{d_2 \times (d_2 - r)}, \text{col}(R_{1\perp}) = \text{col}^c(R_1).$
- When considering offline noisy matrix completion, Xia and Yuan (2021) and Ma et al (2023) give:
- $\frac{T}{d_1 d_2} \text{var}(\langle M_1, Q \rangle) \approx \sigma_1^2 \|P_{M_1}(Q)\|_F^2, \text{ and } P_{M_1}(Q) := Q - L_{1\perp} L_{1\perp}^T Q R_{1\perp} R_{1\perp}^T.$

# Asymptotical Normality

## Extra Definitions

- Variance of  $\hat{M}_1$  : Effective sample size + IPW variance inflation.
- Assume  $\delta$  be the reward gap between optimal/sub-optimal gaps.
- $\Omega_1(\delta) = \{X \in \mathcal{X}; \langle M_1 - M_0, X \rangle > \delta\} \rightarrow$  Effective sample size.
- $\Omega_0(\delta) = \{X \in \mathcal{X}; \langle M_0 - M_1, X \rangle > \delta\} \rightarrow$  IPW variance inflation.
- $\Omega_{\emptyset}(\delta) = \{\Omega_1(\delta) \cup \Omega_0(\delta)\}^c$ .
- $P_{\Omega}(M) : M_{i,j} = 0$  if  $e_i e_j^T \notin \Omega$ .

# Asymptotical Normality

## Arm Optimality Condition

- Given  $\frac{T}{d_1 d_2} \text{var}(\langle M_1, Q \rangle) \approx \sigma_1^2 \|P_{M_1}(Q)\|_F^2$ ,
- $\|P_{\Omega_1} P_{M_1}(Q)\|_F^2$ : variance induced by effective sample size.
- $\|P_{\Omega_0} P_{M_1}(Q)\|_F^2$ : variance induced by IPW variance inflation.
- Arm Optimality Assumption: there exists  $\delta > 0$ , such that
- $\|P_{\Omega_\emptyset} P_{M_i}(Q)\|_F^2 / \min\{\|P_{\Omega_0} P_{M_i}(Q)\|_F^2, \|P_{\Omega_1} P_{M_i}(Q)\|_F^2\} = o(1), i = 1, 2$ .

# Asymptotical Normality

- Under conditions of Corollary 1 and Arm optimality condition.

- $S_1^2 = T^{-\gamma} \|P_{\Omega_1} P_{M_1}(Q)\|_F^2 + C_\gamma \|P_{\Omega_0} P_{M_1}(Q)\|_F^2.$

- $S_0^2 = T^{-\gamma} \|P_{\Omega_0} P_{M_0}(Q)\|_F^2 + C_\gamma \|P_{\Omega_1} P_{M_0}(Q)\|_F^2.$

- Some regularity conditions omitted.

- $$\frac{\langle \hat{M}_0, Q \rangle - \langle M_0, Q \rangle}{\sigma_0 S_0 \sqrt{d_1 d_2 / T^{1-\gamma}}} \rightarrow N(0,1), \frac{\langle \hat{M}_1, Q \rangle - \langle M_1, Q \rangle}{\sigma_1 S_1 \sqrt{d_1 d_2 / T^{1-\gamma}}} \rightarrow N(0,1).$$

# Asymptotical Normality

## Key Elements Estimation

- $$\hat{\sigma}_0^2 = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \frac{1\{a_t = 0\}}{1 - \pi_t} (r_t - \langle \hat{M}_{0,t-1}, X_t \rangle)^2.$$
- $$\hat{\sigma}_1^2 = \frac{1}{T - T_0} \sum_{t=T_0+1}^T \frac{1\{a_t = 1\}}{\pi_t} (r_t - \langle \hat{M}_{1,t-1}, X_t \rangle)^2.$$
- $$\hat{S}_0^2 = (\|P_{\hat{\Omega}_{0,T}} P_{\hat{M}_0}(Q)\|_F^2 / T^\gamma + C_\gamma \|P_{\hat{\Omega}_{1,T}} P_{\hat{M}_0}(Q)\|_F^2) \frac{T}{T - T_0}.$$
- $$\hat{S}_0^2 = (\|P_{\hat{\Omega}_{1,T}} P_{\hat{M}_1}(Q)\|_F^2 / T^\gamma + C_\gamma \|P_{\hat{\Omega}_{0,T}} P_{\hat{M}_1}(Q)\|_F^2) \frac{T}{T - T_0}.$$

# Asymptotical Normality

## Theorem 4 and Corollary 2

- Under conditions in Theorem 3, the above estimator are consistent, and we have

- $$\frac{\langle \hat{M}_1 - \hat{M}_0, Q \rangle - \langle M_1 - M_0, Q \rangle}{\sqrt{(\hat{\sigma}_0^2 \hat{S}_0^2 + \hat{\sigma}_1^2 \hat{S}_1^2) d_1 d_2 / T^{1-\gamma}}} \rightarrow N(0,1) .$$



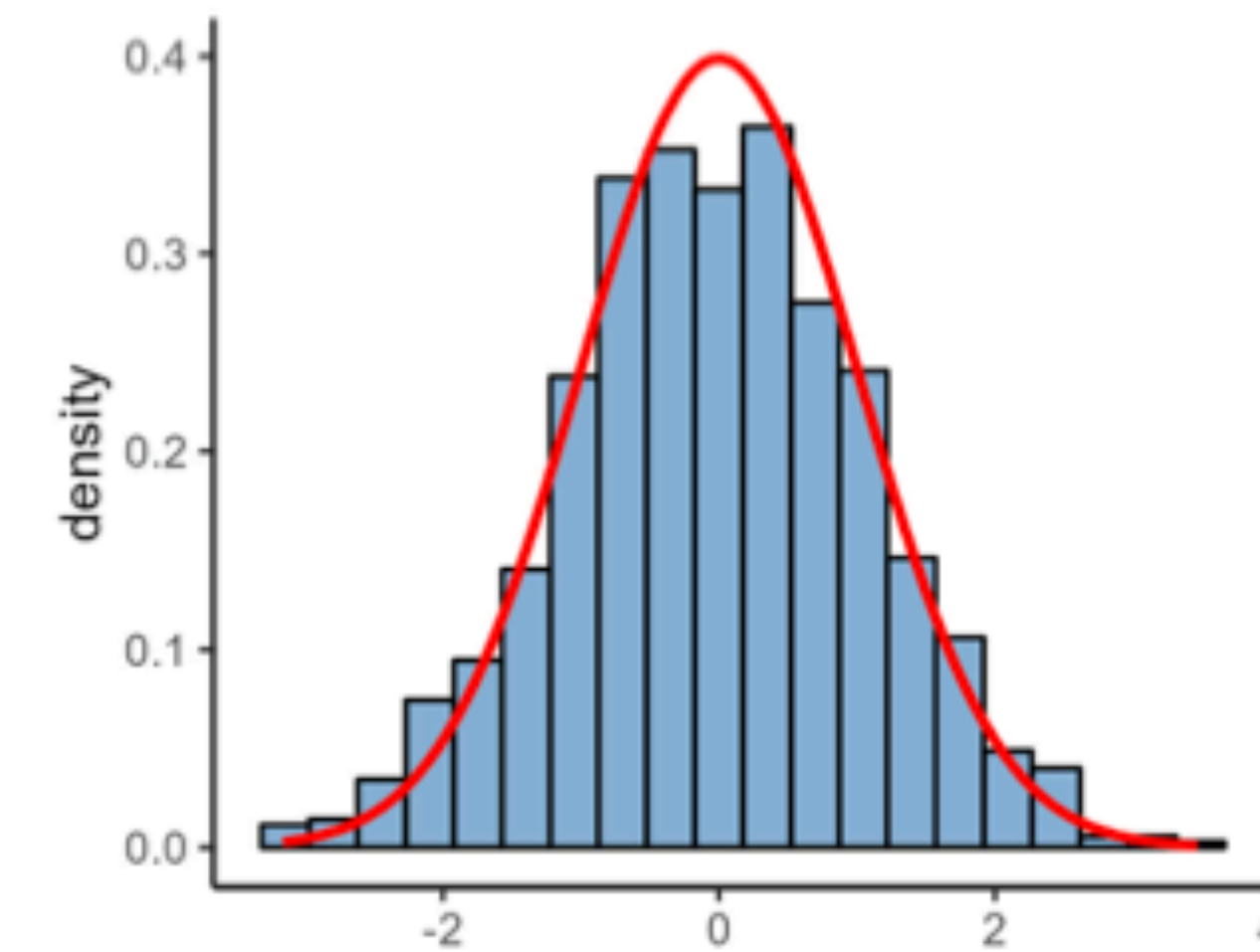
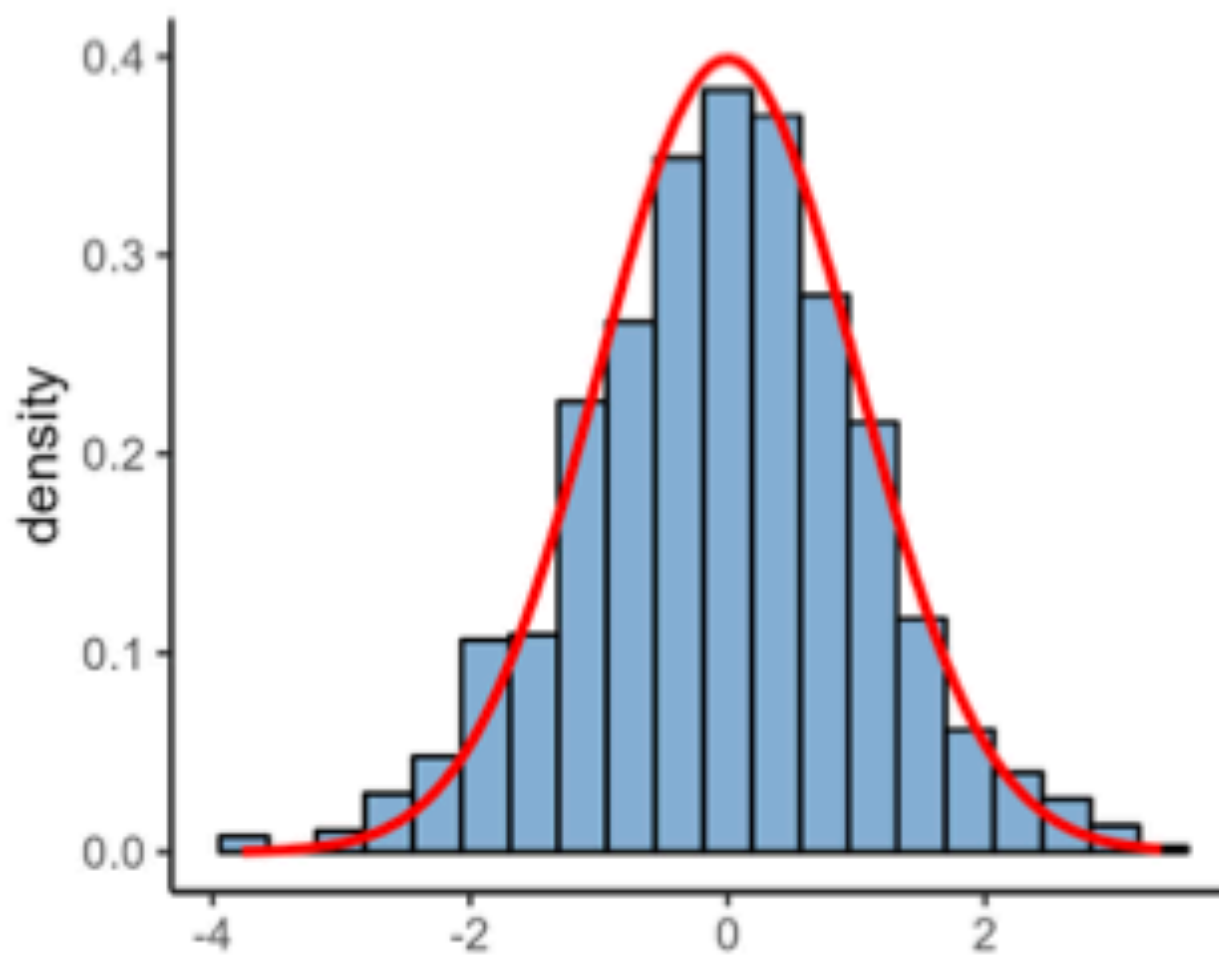
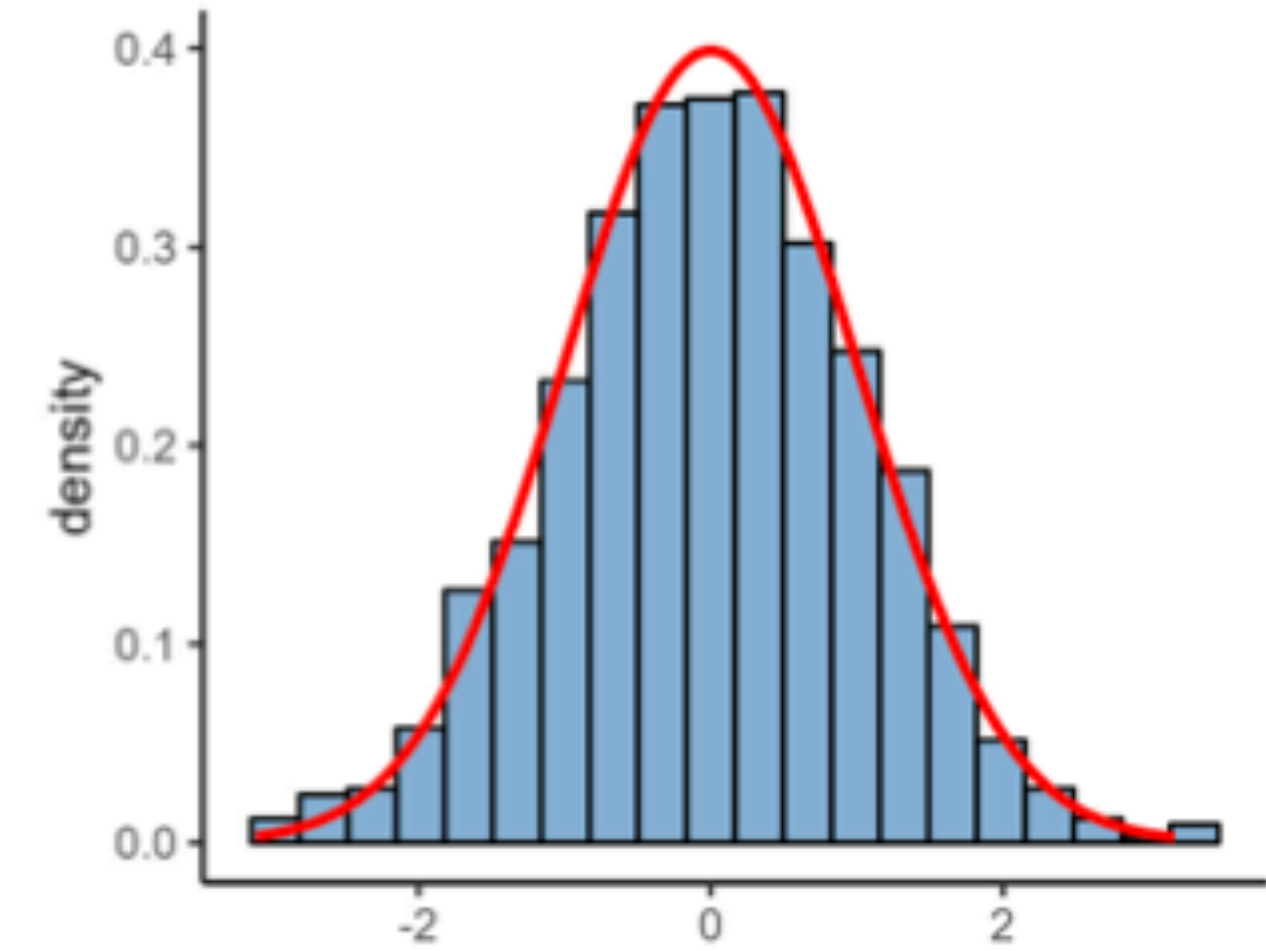
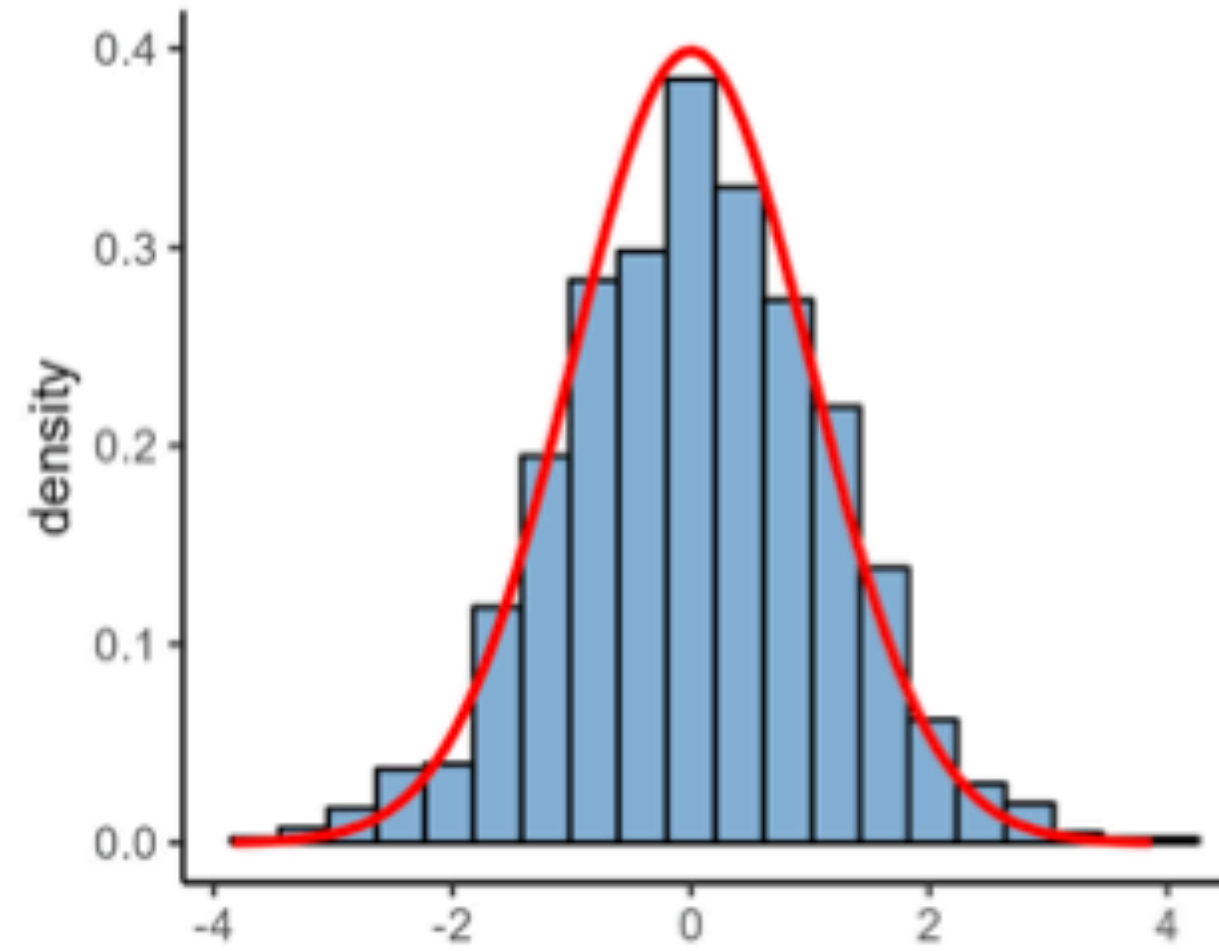
# Simulation

## Basic Settings

- $d_1 = d_2 = 300$ .
- $r = 2$ .
- $M_0, M_1$  generated from uniform distribution.
- $T = 60000, T_0 = 20000$ .
- Four settings:  $\langle M_0, e_1 e_5^T \rangle, \langle M_0, e_1 e_5^T \rangle, \langle M_0 - M_1, e_1 e_r^T \rangle, \langle M_0, e_1 e_r^T - e_2 e_2^T \rangle$ .

# Simulation

## Policy Inference



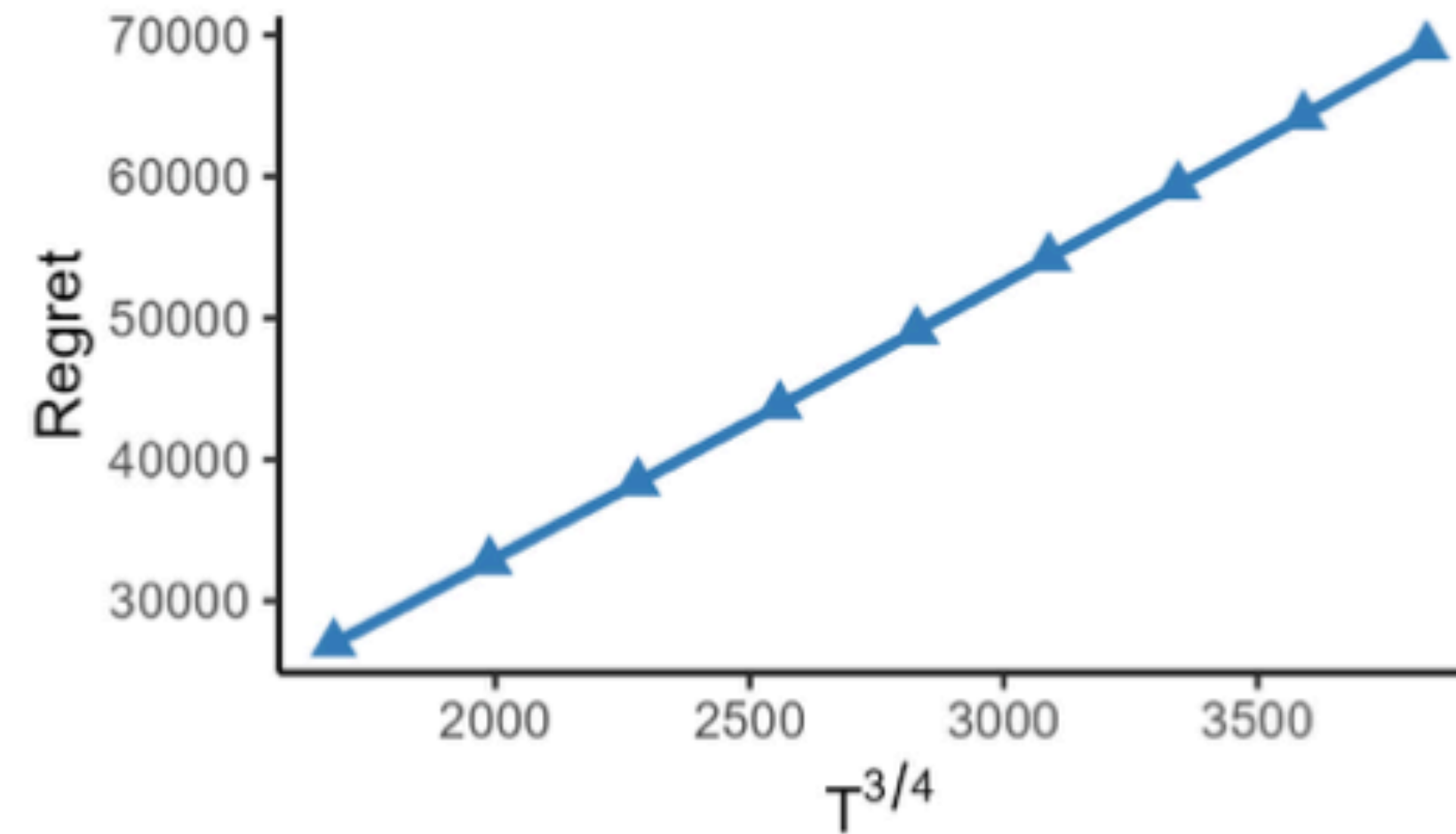
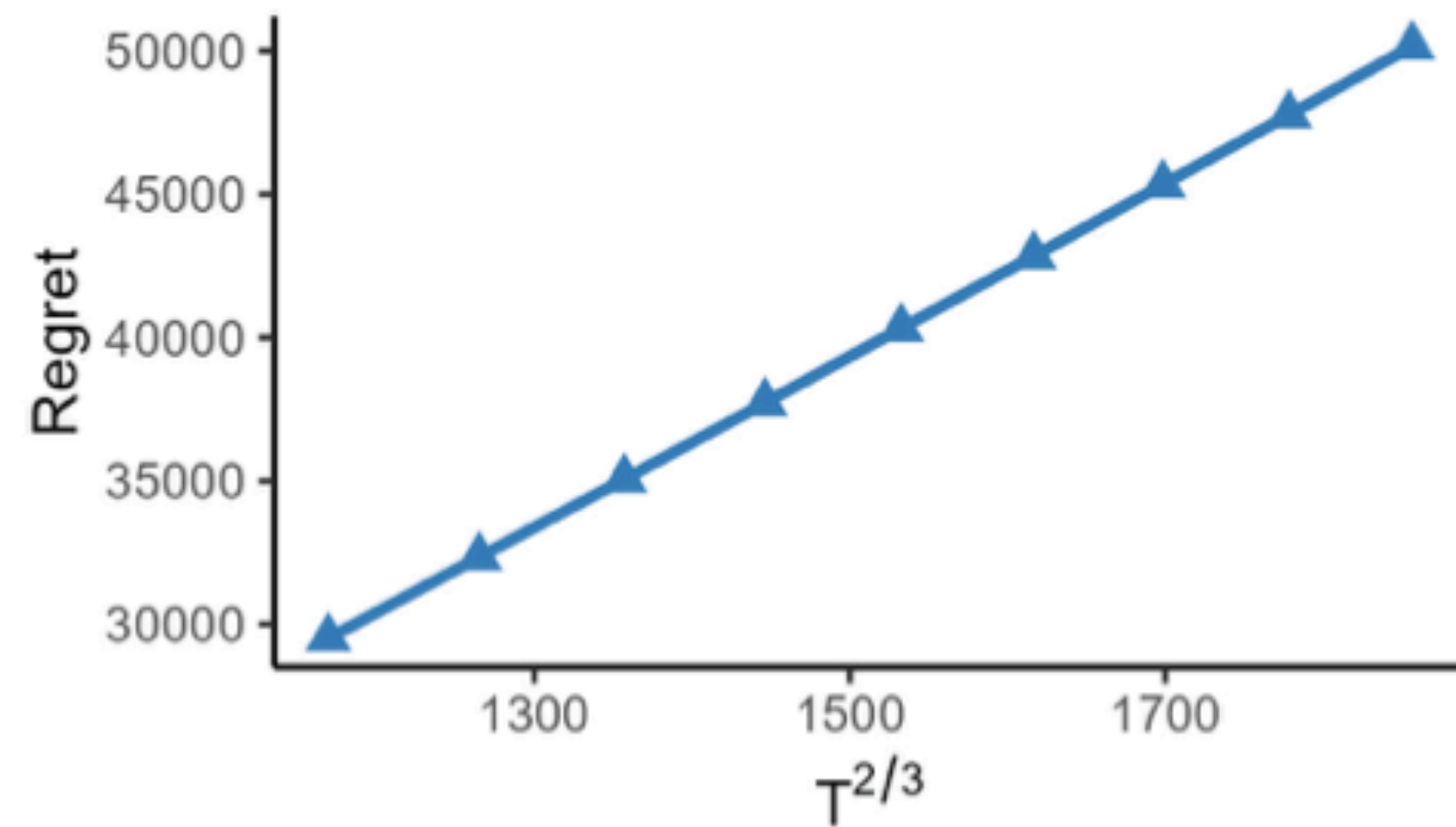
# Simulation

## Regret Analysis

- When  $\gamma = 1/3$  :
- $T$  vary from 40000 to 80000,  $T_0 = 13.5T^{1-\gamma}$  .
- When  $\gamma = 1/4$  :
- $T$  vary from 20000 to 60000,  $T_0 = 4.5T^{1-\gamma}$  .
- Theoretical Regret:  $O(T^{2/3})$  when  $\gamma = 1/3$  and  $O(T^{3/4})$  when  $\gamma = 1/4$ .
- Run 100 simulations and plot average cumulative return.

# Simulation

## Regret Analysis



# Real Data Analysis

## San Francisco Parking Problem

- SFPark pilot project: effectively manage parking towards availability targets in SF.
- Adjust price by hour, day, and block.
- Target occupancy rate: between 60% and 80%.
- Implementation period: 6 to 8 weeks.
- If meet target rate: unchanged.
- Larger than 80%: increase hourly price.
- Lower than 60%: decrease hourly price.

# Real Data Analysis

## San Francisco Parking Problem

- Dataset: includes hourly occupancy rate and price for each block at every hour.

•

	A	B	C	D	E	F	G	H
1	BLOCK_ID	STREET_NAME	BLOCK_NUM	STREET_BLOCK	AREA_TYPE	PM_DISTRICT_NAME	RATE	START_TIME_DT
436	20200	02ND ST	0	02ND ST 0	Pilot	Downtown	3.5	7/6/12 14:00
437	20204	02ND ST	4	02ND ST 400	Pilot	South Embarcadero	1.25	7/6/12 8:00
438	20204	02ND ST	4	02ND ST 400	Pilot	South Embarcadero	1.25	7/6/12 10:00
439	36004	CLEMENT ST	4	CLEMENT ST 400	Control	Inner Richmond		7/6/12 8:00
440	36006	CLEMENT ST	6	CLEMENT ST 600	Control	Inner Richmond		7/6/12 16:00
441	41321	FILBERT ST	21	FILBERT ST 2100	Pilot	Marina		7/6/12 3:00
442	56303	MCALLISTER ST	3	MCALLISTER ST 300	Pilot	Civic Center	3	7/6/12 11:00
443	47100	HARRISON ST	0	HARRISON ST 0	Pilot	South Embarcadero		4/19/13 4:00
444	47105	HARRISON ST	5	HARRISON ST 500	Pilot	South Embarcadero	1.5	4/19/13 13:00
445	56304	MCALLISTER ST	4	MCALLISTER ST 400	Pilot	Civic Center	2	4/19/13 7:00
446	56304	MCALLISTER ST	4	MCALLISTER ST 400	Pilot	Civic Center	2	4/19/13 11:00
447	33103	BRYANT ST	3	BRYANT ST 300	Pilot	South Embarcadero		4/19/13 2:00
448	41522	FILLMORE ST	22	FILLMORE ST 2200	Pilot	Fillmore		4/19/13 3:00
449	41522	FILLMORE ST	22	FILLMORE ST 2200	Pilot	Fillmore	3.25	4/19/13 9:00
450	41524	FILLMORE ST	24	FILLMORE ST 2400	Pilot	Fillmore		4/19/13 8:00
451	41529	FILLMORE ST	29	FILLMORE ST 2900	Control	Union		4/19/13 11:00
452	41530	FILLMORE ST	30	FILLMORE ST 3000	Control	Union		4/19/13 5:00
453	41530	FILLMORE ST	30	FILLMORE ST 3000	Control	Union		4/19/13 8:00
454	50001	JACKSON ST	1	JACKSON ST 100	Pilot	Downtown		4/19/13 4:00

# Real Data Analysis

## San Francisco Parking Problem

- Focus on the Downtown area, 2011 to 2012.
- Four price adjustment times: Aug 1st, Oct 11th, Dec 13th, Feb 14th. (Five periods)
- $d_1 = 34$  blocks.
- $d_2 = 22$  time points,
- (7 am to 6 pm on weekdays & weekends)
- $M$ : target deviation matrix.
- $M_0, M_1$  :Low/high parking price.
- Each hour corresponds to one request.
- $T = 105,825$ .
- $r = 5$  from rough estimation.

	7am	8am	9am	10am	11am	...	6pm
Haight Street							
The Castro							
Union Square							
Mission Street							

# Real Data Analysis

## San Francisco Parking Problem

- An estimation workflow:
- A block  $j_1 \in [d_1]$  at hour  $j_2 \in [d_2]$ ,  $X_t = e_{j_1} e_{j_2}^T$ .
- Pick action  $a_t$ , if inside target range (60% to 80%):  $r_t = 0$ . Else  $r_t = -0.1$ .
- Discard the observation whose observed action  $\neq$  online algorithm action.



# Real Data Analysis

## San Francisco Parking Problem

- A representative block: 02ND ST 200.
- 

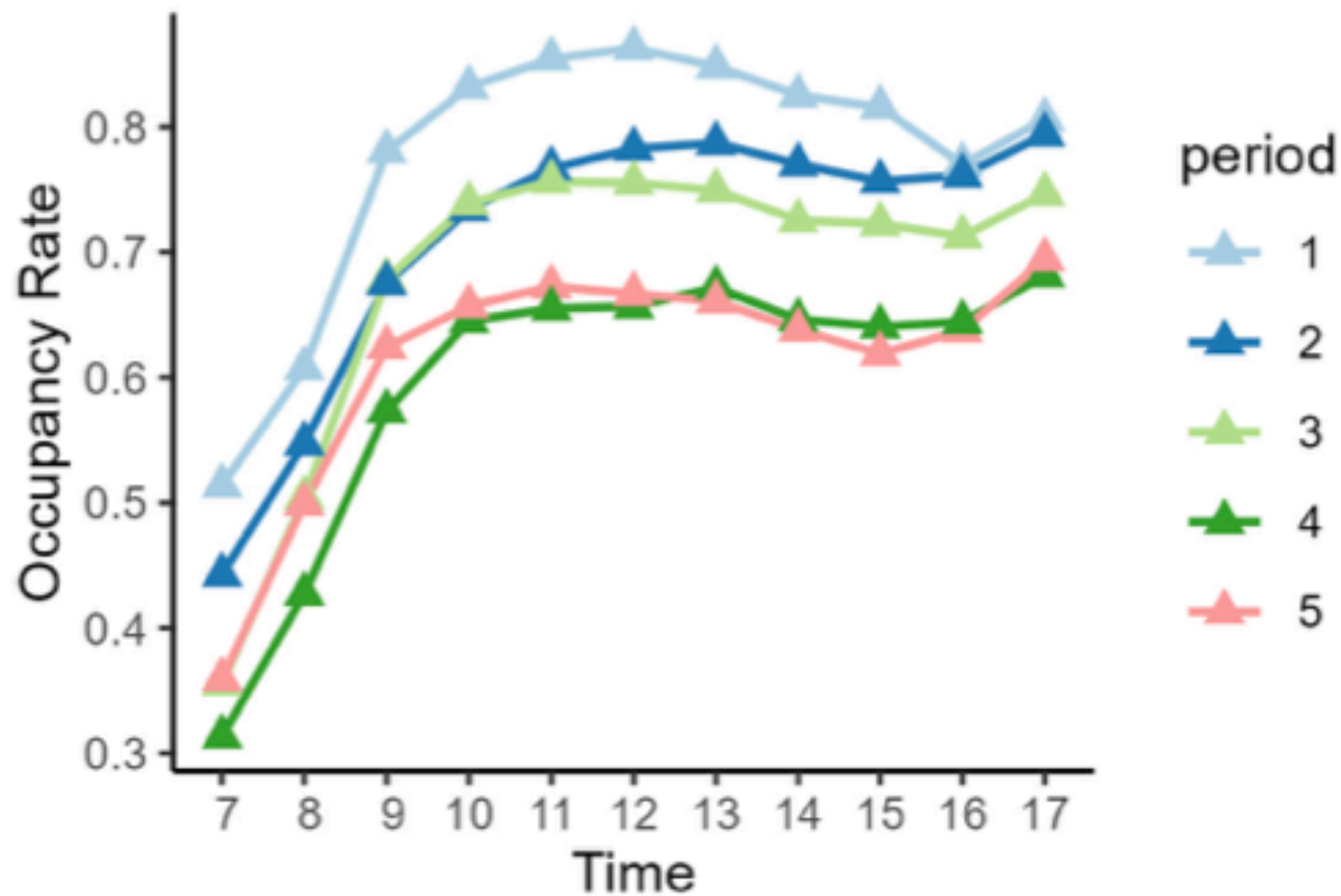
Time	7	8	9	10	11
p-value	0.282	0.016	0.009	0.002	0.304

$$H_0 : \langle M_0 - M_1, e_{j_1} e_{j_2}^T \rangle \leq 0.$$

Time	12	13	14	15	16	17
p-value	<0.001	<0.001	<0.001	0.773	<0.001	0.999

$$H_0 : \langle M_1 - M_0, e_{j_1} e_{j_2}^T \rangle \leq 0.$$

Low parking price before 11am and high parking price after 12 pm is optimal!



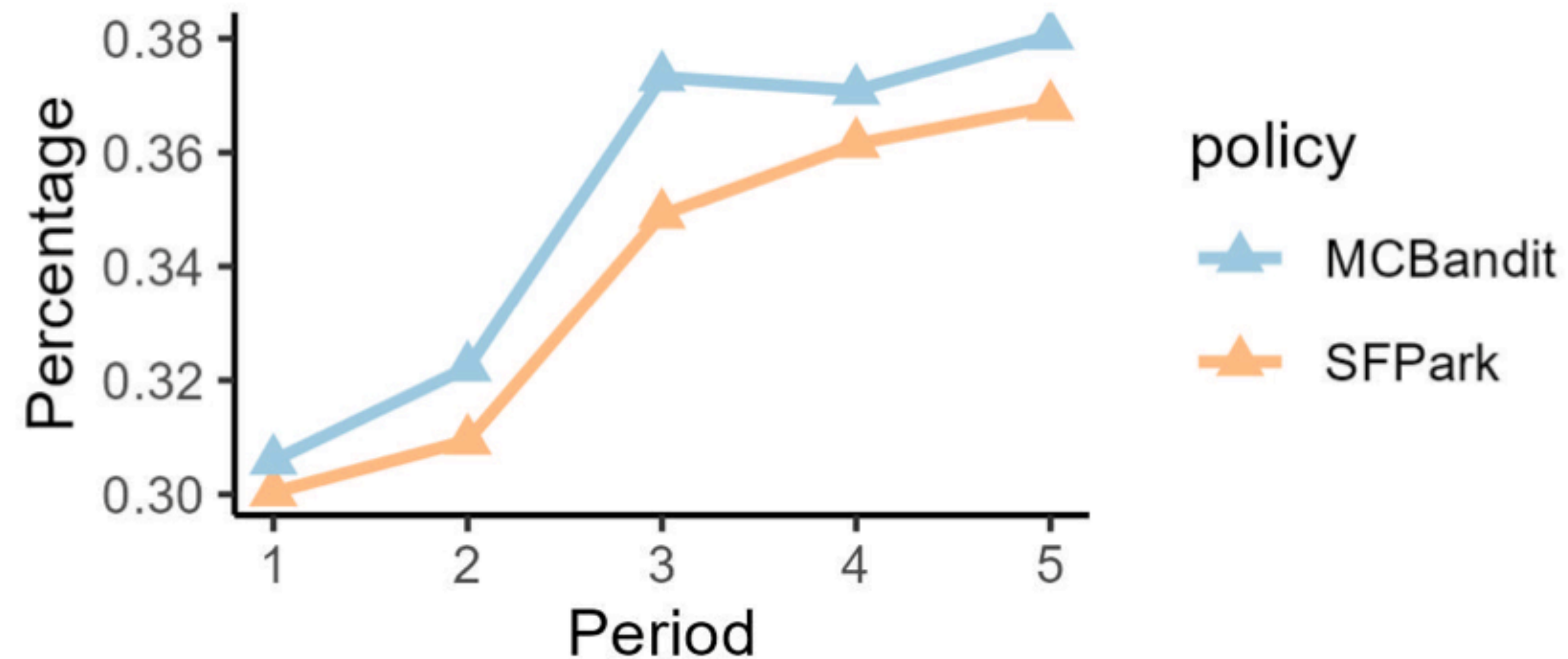
# Real Data Analysis

## San Francisco Parking Problem

- SFPark vs MCBandit.
- Comparing overall performance through a percentage of reaching the target.
- Calculating MCBandit performance:
- Keep the data whose action aligns with the observed action.
- Replace others by “nearest neighbors”.

# Real Data Analysis

## San Francisco Parking Problem



# Summary

- An online algorithm for matrix completion bandit.
- Optimal estimation bound (up to log factors).
- Regret bound  $O(T^{1-\gamma} + d_1^{1/2} T^{(1+\gamma)/2})$ .
- An inference procedure (online adaption from Xia and Yuan 2021 and Ma et al 2023) for making policy inferences.