*Submitted to the Annals of Statistics*

# ONLINE STATISTICAL INFERENCE IN DECISION-MAKING WITH MATRIX CONTEXT

BY QIYU HAN[1,a], WILL WEI SUN[1,b] AND YICHEN ZHANG[1,c]

[1] *Mitchell E. Daniels, Jr. School of Business, Purdue University,* [a]*han541@purdue.edu;* [b]*sun244@purdue.edu;* [c]*yichen@purdue.edu*

The study of online decision-making problems that leverage contextual information has drawn notable attention due to their significant applications in fields ranging from healthcare to autonomous systems. In modern applications, contextual information can be rich and is often represented as a matrix. Moreover, while existing online decision algorithms mainly focus on reward maximization, less attention has been devoted to statistical inference. To address these gaps, in this work, we consider an online decision-making problem with a matrix context where the true model parameters have a low-rank structure. We propose a *fully online* procedure to conduct statistical inference with adaptively collected data. The low-rank structure of the model parameter and the adaptive nature of the data collection process make this difficult: standard low-rank estimators are biased and cannot be obtained in a sequential manner while existing inference approaches in sequential decision-making algorithms fail to account for the low-rankness and are also biased. To overcome these challenges, we introduce a new online debiasing procedure to simultaneously handle both sources of bias. Our inference framework encompasses both parameter inference and optimal policy value inference. In theory, we establish the asymptotic normality of the proposed online debiased estimators and prove the validity of the constructed confidence intervals for both inference tasks. Our inference results are built upon a newly developed low-rank stochastic gradient descent estimator and its convergence result, which are also of independent interest.

**1. Introduction.** From personalized medicine to recommendation systems, exploiting personalized information in decision-making has gained popularity during the last decades (Kosorok and Laber, 2019; Fang, Wang and Wang, 2023; Qi, Pang and Liu, 2023). In the widely studied framework of online decision-making with contextual information, decisions are sequentially made for users based on the current context and historical interactions (Li et al., 2010; Agrawal and Goyal, 2013; Li, Lu and Zhou, 2017; Lattimore and Szepesvári, 2020). In traditional settings, the context is typically formulated in a vector. However, contextual information in modern online decision-making problems is often in a matrix form. In the skin treatment example shown in Figure 1, the decision-making policy determines whether an immediate intervention should be applied based on the patient's current image of skin condition (a matrix context) and the health outcomes of historical interventions (Akrout et al., 2019). The inspiration for this example can be traced to the recently growing application of mobile Health, which targets to deliver immediate interventions, such as motivational messages, to individuals through mobile devices according to their current health condition (Istepanian, Laxminarayan and Pattichis, 2007; Deliu, Williams and Chakraborty, 2022). In such examples, the context is an image that can be formulated as a matrix. The goal of the

*MSC2020 subject classifications:* Primary 62L20, 90B50; secondary 62E20.

*Keywords and phrases:* online inference, online decision-making, low-rank matrix, reinforcement learning, stochastic gradient descent.

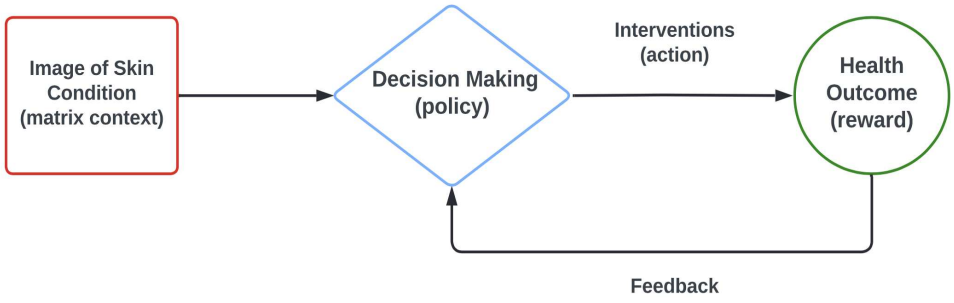



Fig 1: An illustration of our online decision-making framework with matrix context.

decision-making policy is to decide the best action at each time based on the current matrix context and all historical interactions.

In this paper, we consider an online decision-making problem with matrix contexts. In particular, at time $t$, given a matrix context $X_t \in \mathbb{R}^{d_1 \times d_2}$, the policy takes an action $a_t \in \{0, 1\}$ and observes a noisy reward $y_t \in \mathbb{R}$ as

$$y_t = a_t \langle M_1, X_t \rangle + (1 - a_t) \langle M_0, X_t \rangle + \xi_t, \tag{1}$$

where $\xi_t \in \mathbb{R}$ is the random noise and $\langle M_i, X_t \rangle = tr(X_t^\top M_i)$, for $i \in \{0, 1\}$, denotes the matrix inner product. The true matrix parameter $M_i$ is assumed to be of low rank with a rank $r \ll \min\{d_1, d_2\}$. In our motivation example, a group of pixels in the image that form a region can impose a collaborative effect on describing the health outcome, allowing the matrix parameter to have a low-rank structure (Chen et al., 2019; Xia, 2019; Xia and Yuan, 2021). In addition, such a low-rank structure is crucial in online decision-making due to its high dimensionality compared to its limited sample size. In (1), when $a_t = 1$ (with intervention), the reward is given by $\langle M_1, X_t \rangle + \xi_t$ (health outcome with intervention); when $a_t = 0$ (without intervention), the reward is given by $\langle M_0, X_t \rangle + \xi_t$ (health outcome without intervention). Without loss of generality, our work mainly focuses on a binary action, i.e., $a_t \in \{0, 1\}$ at each time $t$, and it can be easily extended to multiple actions in a discrete action space.

While existing sequential decision-making algorithms mainly focused on choosing the best action to maximize the cumulative reward (Li et al., 2010; Agrawal and Goyal, 2013; Li, Lu and Zhou, 2017; Lattimore and Szepesvári, 2020), less attention has been paid to statistical inference in sequential decision-making frameworks. In real-world applications, we are often not just interested in obtaining the point estimate of the reward function but also a measure of the statistical uncertainty associated with the estimate. This is especially relevant in fields such as personalized medicine, mobile health, and automated driving, where it is often risky to run a policy without a statistically sound estimate of its quality. For example, online randomized experiments like A/B testing have been widely conducted by technological/pharmaceutical companies to compare a new product with an old one. Recent studies (Li et al., 2021; Shi et al., 2021a, 2023) have used various bandit or reinforcement learning methods to form sequential testing procedures. In these online evaluation tasks, it is important to quantify the uncertainty of the point estimate for constructing valid hypothesis testing.

Statistical inference significantly enhances scientific knowledge by applying insights from prior experiments to improve future research designs, extending beyond the immediate objectives of in-experiment learning aimed at optimizing decision-making performance. This knowledge is crucial for capturing the extensive, long-term consequences of actions and associated rewards. For example, if an inference result learns that certain variables have a significant impact on the outcomes, this insight can be used to improve the design of future

experiments (Shi et al., 2022; Zhang, Janson and Murphy, 2021, 2022; Shi et al., 2024). Different from in-experiment learning focusing on maximizing reward within the trial, statistical inference can lead to more strategic and informed decision-making over time (Simchi-Levi and Wang, 2023). Therefore, our work aims to provide a comprehensive online inferential framework applicable throughout a wide range of sequential decision-making algorithms.

Motivated by the importance of statistical inference, we first provide a procedure to conduct entry-wise inference on the true matrix parameter $M_i$ under the sequential decision-making framework. We introduce a matrix $T \in \mathbb{R}^{d_1 \times d_2}$ such that $\langle M_i, T\rangle$ characterizes the entries of interest for hypothesis testing. For example, setting $T = e_{j_1} e_{j_2}^\top$, where $\{e_{j_1}\}_{j_1 \in [d_1]}$ and $\{e_{j_2}\}_{j_2 \in [d_2]}$ denote the canonical basis vector in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively, our work allows a valid confidence interval of $\langle M_i, T\rangle = M_i(j_1, j_2)$ for hypothesis testing on whether the $(j_1, j_2)$-th entry of the matrix $M_i$ is zero, i.e.,

$$H_0 : M_i(j_1, j_2) = 0 \quad \text{v.s.} \quad H_1 : M_i(j_1, j_2) \neq 0, \tag{2}$$

where $M_i(j_1, j_2)$ denotes the $(j_1, j_2)$ entry of $M_i$. In this case, we can test the effectiveness of a certain entry in the matrix context for describing the reward. It is worth pointing out that the form of $T$ is flexible. For example, setting $T = e_{j_1} e_{j_2}^\top - e_{j_3} e_{j_4}^\top$ can test whether $M_i(j_1, j_2)$ and $M_i(j_3, j_4)$ are significantly different. Moreover, our work also enables us to check whether different actions result in different effectiveness of a certain context entry by testing

$$H_0 : M_1(j_1, j_2) - M_0(j_1, j_2) = 0 \quad \text{v.s.} \quad H_1 : M_1(j_1, j_2) - M_0(j_1, j_2) \neq 0. \tag{3}$$

As Poldrack, Mumford and Nichols (2011) introduced in their neuroimaging book, statistical inference on the pixel level is able to test whether an individual pixel in an image has a significant effect on measuring the outcome. In our motivational example in Figure 1, hypothesis test (2) provides the answer of whether a certain pixel is significant in determining the reward, while hypothesis test (3) helps us understand if the intervention causes a significant difference in the patient's health outcome.

In addition to the parameter inference, we further extend our online inference framework to the optimal policy value. This value represents the best-expected reward a decision-maker can achieve given complete knowledge of the environment. The need to infer this optimal value becomes crucial in real-world applications whenever the experimenters need to assess the best possible reward they can achieve given the currently available interventions. Such assessment determines the adequacy of current actions in achieving desirable outcomes or necessitates refinement of the action set. In particular, the optimal policy value attainable under the current environment is defined as

$$V^* = \mathbb{E}\left[\left\langle M_{a^*(X)}, X\right\rangle\right], \quad \text{with} \;\; a^*(X) = I\{\langle M_1 - M_0, X\rangle > 0\}, \tag{4}$$

where $a^*(X)$ indicates the optimal policy for a given context $X$ under our reward function described in (1). To provide additional clarification, experimenters can assess whether the current best treatment outcome surpasses a certain threshold $(V_0)$ by conducting the following one-sided statistical test:

$$H_0 : V^* \leq V_0 \quad \text{v.s.} \quad H_1 : V^* > V_0. \tag{5}$$

After exploring the essential aspects of both parameter inference and optimal policy value inference, we now present our proposed methodology, a procedural framework specifically designed to address these key areas of statistical estimation and inference in online decision-making. In particular, we iteratively update a low-rank estimation of $M_i$ under a sequential decision-making framework with low computational cost. Meanwhile, we simultaneously

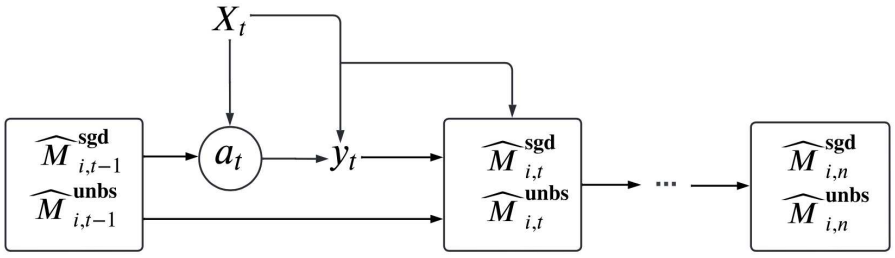



Fig 2: The flow chart of the proposed sequential procedure for a total of $n$ iterations.

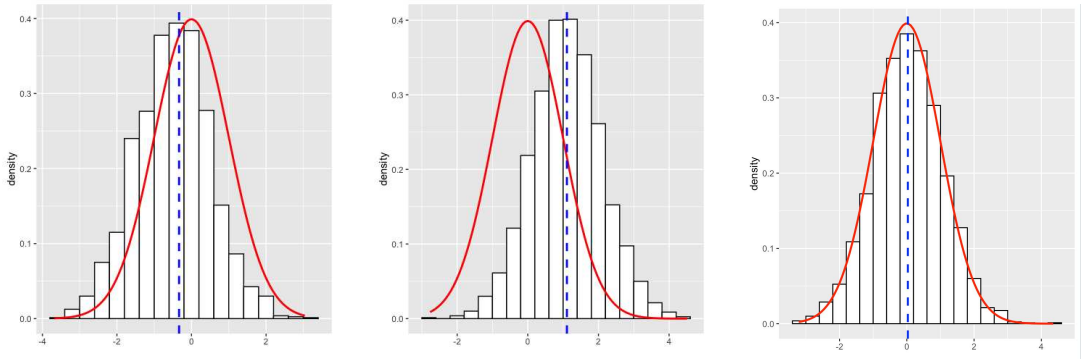
maintain an unbiased estimator in an online fashion for inference purposes. We briefly illustrate this online procedure in Figure 2 where the low-rank estimation of $M_i$ is denoted as $\widehat{M}^{\mathbf{sgd}}_{i,t}$, and the unbiased estimator for the inference purpose is denoted as $\widehat{M}^{\mathbf{unbs}}_{i,t}$. We summarize the role and properties of both estimators below.

- $\widehat{M}^{\mathbf{sgd}}_{i,t}$: Low-rank but biased, sequentially updated low-rank estimation for $M_i$.
- $\widehat{M}^{\mathbf{unbs}}_{i,t}$: Unbiased but not low-rank, designed for conducting inference of $M_i$.

In our problem, it is important to maintain both estimators to handle the two tasks of sequential decision-making and online inference. The methodological contributions of our proposed procedure can be viewed from three aspects. First, in existing low-rank literature, a low-rank estimator is typically obtained by solving nuclear-norm penalized optimization using offline samples (Candes and Plan, 2011; Koltchinskii and Xia, 2015; Chen et al., 2019; Xia, 2019). However, the offline methods become impractical when handling large-scale matrices due to the substantial storage costs. For instance, storing a single $500 \times 500$ single-precision matrix requires about one megabyte, underscoring the significant storage demands in an offline setting where thousands of such matrices are necessary. In contrast, our proposed online estimation method exhibits distinct advantages in terms of data storage efficiency by eliminating the need for local storage of the complete dataset. Our online estimation procedure uses a single observation at a time and then discards it, which makes this technique particularly well-suited for high-dimensional datasets. In our method, we sequentially update the low-rank factorization of $\widehat{M}^{\mathbf{sgd}}_{i,t}$ via stochastic gradient descent (SGD) to preserve its low-rankness. While it is suitable for sequential decision-making, $\widehat{M}^{\mathbf{sgd}}_{i,t}$ is not directly applicable for statistical inference due to its bias. This motivates our new design of an unbiased estimator $\widehat{M}^{\mathbf{unbs}}_{i,t}$ by sequentially debiasing $\widehat{M}^{\mathbf{sgd}}_{i,t}$ for online inference.

Second, the debiasing procedure to obtain $\widehat{M}^{\mathbf{unbs}}_{i,t}$ also requires delicate design since it needs to compensate for two sources of bias: (1) the bias in $\widehat{M}^{\mathbf{sgd}}_{i,t}$ caused by preserving the low-rankness, and (2) the bias in adaptive sample collection due to the fact that the samples are not collected randomly, but rather through the distribution of $a_t$ which is determined by the historical information. To illustrate these two types of bias, Figure 3a demonstrates the bias of the estimator caused by adaptive sample collection, and Figure 3b demonstrates the bias of the estimator caused by the low-rankness. To fill in the gap, we introduce a new debiasing approach to handle both sources of bias simultaneously in a sequential manner. Figure 3c shows that our proposed estimator is unbiased and enables a valid statistical inference.

Third, we further introduce an online estimator tailored for optimal policy value inference. While most of the existing literature focuses on offline value inference, our proposed estimator for the optimal policy value equips the experimenters with the ability to monitor the




(a) Bias of the estimator caused by adaptively collected data.

(b) Bias of the estimator caused by the low-rankness.

(c) Our proposed debiased estimator

Fig 3: The empirical distributions of two biased estimators and our debiased method. The center of each empirical distribution is shown in the blue dashed line, and the standard normal curve is shown in red.

confidence interval of the optimal policy value in a timely manner. Unlike the approach for parameter inference, which requires a sufficient sample size for both action 1 and action 0 to ensure adequate information is collected for $M_1$ and $M_0$, the optimal policy value estimator only leverages samples obtained through the estimated optimal action at each time. As a result, our approach to inferring the optimal policy value enables the exploration probability to gradually decrease over time. Additionally, our framework is adaptable to handle scenarios in which the probabilities of selecting each action, as determined by the decision-making policy, are unknown and estimated empirically.

In addition to the aforementioned methodological contributions, we further summarize our theoretical contributions and discuss the technical challenges in our analysis.

- We provide a non-asymptotic convergence result for the sequentially updated low-rank estimator $\widehat{M}_{i,t}^{\mathbf{sgd}}$ in Theorem 2.2. That is, with high probability,

$$\|\widehat{M}_{i,t}^{\mathbf{sgd}} - M_i\|_{\mathrm{F}} \leq C\sigma_i \sqrt{\frac{dr\log^2 d}{t^{\varsigma}}},$$

  for some positive constant $C$, where $d = \max\{d_1, d_2\}$, and $\varsigma \in (0.5, 1)$. The existing SGD literature for the low-rank estimation is limited except Jin, Kakade and Netrapalli (2016) considers a noiseless matrix completion problem with i.i.d. samples. Our work, on the other hand, deals with noisy reward and the adaptive sampling in the sequential decision-making setting. In the noiseless scenario, stochastic objective functions share the same minimizer, with each gradient descent iteration steadily progressing toward this common minimizer. However, the introduction of noise leads to the steps of SGD targeting varying minimizers, causing the SGD updates to oscillate or move away from the optimal solution's local region. To prevent this from happening, it is crucial to add stabilization measures to ensure the optimization trajectory consistently advances toward the right direction.
- We establish the asymptotic normality of $\widehat{m}_T^{(i)}$ for estimating $m_T^{(i)} = \langle M_i, T\rangle$ in Theorem 3.1. Due to the fact that our data are collected adaptively and sequentially, the analysis based on offline *i.i.d.* samples is no longer applicable in our case. Traditional debiasing approach in the offline low-rank literature (Xia and Yuan, 2021) involves splitting the

dataset into two independent sets, using one to correct biases in the low-rank estimator derived from the other one. However, in online decision-making, where data is passed only once, a sequential debiasing method is necessary. Gathering all data for debiasing at the end is computationally infeasible and renders existing methods ineffective. Our sequential method eliminates the need to store historical data, allowing efficient debiasing at each step in the online decision-making process. Due to these significant differences, new proof techniques are necessary to address the dependency on data. In addition, due to both low-rankness and data adaptivity, our proof involves controlling the additional variance introduced by our debiasing procedure. As an important step, the convergence result of $\widehat{M}_{i,t}^{\mathbf{sgd}}$ shown in Theorem 2.2 ensures this additional variance is well controlled.

- For the purpose of statistical inference of the parameter, we propose a fully online estimator for the variance of $\widehat{m}_T^{(i)}$ without storing historical data. We prove the consistency of this estimator, which provides the guarantees that the asymptotic normality in Theorem 3.3 holds with the estimated standard deviation. This ensures the validity of our constructed confidence interval for the true matrix parameter.
- Finally, we establish the asymptotic normality of our optimal policy value estimator in Theorem 4.1, showing that the asymptotic bias of the estimator approaches zero with data accumulation. We additionally propose a variance estimator for constructing confidence intervals, and Theorem 4.2 demonstrates the reliability of this estimator, affirming the empirical validity of the generated confidence intervals. Besides addressing the theoretical challenges posed by non-*i.i.d.* data collection and the low-rank structure, establishing the asymptotic normality of the optimal policy value estimator also involves ensuring convergence of the estimated optimal action towards the true optimal action. This is crucial for controlling the bias resulting from the accumulation of differences between the estimated and true optimal actions, which is shown to be sufficiently small compared to the variance of the optimal policy value estimator.

1.1. *Related Literature.* This section discusses three lines of related work, including online inference based on SGD, statistical inference in bandit and Reinforcement Learning (RL) settings, and statistical inference for low-rank matrices. The literature review presents the fundamental differences compared to our work in terms of motivation and problem settings, which end up with different algorithms and technical tools for theoretical analysis.

**Online Inference Based on SGD.** Our work is related to a recent growing literature on statistical inference based on SGD. Fang, Xu and Yang (2018) proposed an online bootstrap procedure for the estimation of confidence intervals of the SGD estimator. Chen et al. (2020) studied the statistical inference of the true model parameters by proposing two consistent estimators of the asymptotic covariance of the averaged SGD estimator, extended by Zhu, Chen and Wu (2023) to a fully online scenario. Shi et al. (2021b) developed an online estimation procedure for high-dimensional statistical inference. Chen et al. (2021) studied the online inference when the gradient information is unavailable and Tang et al. (2023) extends the analysis to SGD with momentum. All of these works consider i.i.d. samples and are not applicable to adaptively collected data. Recently, Chen, Lu and Song (2021a); Chen et al. (2022) conducted the statistical inference of the model parameters via SGD under online decision-making settings. Ramprasad et al. (2023); Liu et al. (2023) studied the online inference in linear stochastic optimization with Markov noise. However, none of these works handles the low-rankness in a matrix estimation.

**Statistical Inference in Bandit and RL Settings.** Chen, Lu and Song (2021b) studied the asymptotic behavior of the parameters under the traditional linear contextual bandit framework. Bibaut et al. (2021) studied the asymptotic behavior of the treatment effect with

contextual adaptive data collection. Zhan et al. (2021) and Hadad et al. (2021) developed adaptive weighting methods to construct estimators that are suitable for policy value inference with adaptive collected data. Deshpande, Javanmard and Mehrabi (2023) and Khamaru et al. (2021) considered the adaptive linear regression. Zhang, Janson and Murphy (2021, 2022) provided statistical inference for the M-estimators in the contextual bandit and non-Markovian environment. Ye, Cai and Song (2023) employed a doubly robust estimator for the optimal policy value inference within an online decision-making framework. In addition to these references, there are also related inference works in RL. For example, Shi et al. (2022) constructed the confidence interval for the policy value in the Markov decision process, and Shi et al. (2024); Bian et al. (2024) further extended the statistical inference to the confounded Markov decision processes and doubly inhomogeneous environments, respectively. The above works are tailored for vector contexts and not for matrix contexts.

**Statistical Inference for Low-Rank Matrix.** With the sample splitting procedure for obtaining an unbiased estimator, Carpentier et al. (2015) constructed confidence sets for the matrix of interest with regard to its Frobenius norm. Xia (2019) conducted the inference on the matrix's singular subspace, reflecting the information about matrix geometry. To conduct inference on matrix entries, Carpentier and Kim (2018) proposed a new estimator that was established using the iterative thresholding method. Chen et al. (2019) proposed a debiased estimator for a matrix completion problem. Xia and Yuan (2021) studied the inference of a matrix linear form, which established the entry-level confidence intervals. However, none of the above works is applicable when the data are adaptively collected. As shown in Figure 3, we need to handle two sources of bias in our setting, which demands a new debiasing procedure.

1.2. *Notations and Organization.* For a matrix $M \in \mathbb{R}^{d_1 \times d_2}$, we use $\|M\|_{\mathrm{F}}$ to denote its Frobenius norm, $\|M\|$ to denote its matrix operator norm, and $\|M\|_{\ell_1}$ to denote its vectorized $\ell_1$ norm. We use $M(i,j)$ to denote the entry of $M$ at row $i$ and column $j$. Assume a matrix has rank $r$, then we denote the $\lambda_1, \lambda_r$ as its largest and smallest singular values, respectively, and we denote $\kappa(M) = \lambda_1/\lambda_r$ as the condition number of $M$. Given a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we denote $\langle M, A\rangle$ as the matrix inner product, i.e., $\langle M, A\rangle = \mathrm{tr}(M^\top A)$. For a matrix $U \in \mathbb{R}^{d\times r}$, then we denote its orthogonal complement as $U_\perp \in \mathbb{R}^{d\times(d-r)}$. We use the notation $C_1, C_2, \ldots$ to represent the absolute constants, and we use $a \lesssim b$ to represent $a \le Cb$ for some absolute constant $C$. We denote $\xrightarrow{p}$ and $\xrightarrow{d}$ as convergence in probability and in distribution, respectively. Finally, we use $I\{\cdot\}$ to denote the indicator function.

The rest of the paper is organized as follows. In Section 2, we introduce our problem setting and decision-making procedure under the online decision-making framework. In Section 3, we propose the online debiasing procedure to construct an unbiased estimator for inference purposes. We also present the asymptotic normality of the proposed estimator and prove the validity of the proposed statistical inference procedure. In Section 4, we outline a procedure for inferring the value of the optimal policy. In Section 5, we present numerical experiments to demonstrate the merit of our proposed method. Finally, the supplementary material includes additional numerical studies, further discussions on assumptions, and comprehensive proofs of main theorems and technical lemmas.

**2. Online Decision Making and Low-Rank Estimation.** In this section, we first present the online decision-making procedure designed to address the exploration-exploitation dilemma. Subsequently, we propose a sequential low-rank estimation for $M_i$, denoted as $\widehat{M}_{i,t}^{\mathbf{sgd}}$ for $i = 0, 1$ and $t = 1, 2, \ldots$. The convergence properties of the proposed SGD estimator are discussed in the later part of this section.

2.1. *Sequential Decision Making.* In sequential decision-making, the objective is to select a series of actions over time aiming to maximize the cumulative reward. As described by our reward model, denoted by (1), the reward, represented by $y_t$ at time $t$, is observed after the execution of an action $a_t$. Let $\mathcal{F}_t$ denote the filtration generated by all the historical randomness up to time $t$, i.e., $\mathcal{F}_t = \sigma(X_1, a_1, y_1, ..., X_t, a_t, y_t)$. Then the policy function, denoted as $\pi_t$, can be formally expressed as

$$\mathbb{P}(a_t = 1|\mathcal{F}_{t-1}, X_t) = \pi_t(X_t, \widehat{M}^{\mathbf{sgd}}_{1,t-1}, \widehat{M}^{\mathbf{sgd}}_{0,t-1}),$$

and correspondingly, $\mathbb{P}(a_t = 0|\mathcal{F}_{t-1}, X_t) = 1 - \pi_t(X_t, \widehat{M}^{\mathbf{sgd}}_{1,t-1}, \widehat{M}^{\mathbf{sgd}}_{0,t-1})$. Here, the domain and range of policy function can be specified as $\pi_t : \mathbb{R}^{d_1 \times d_2} \times \mathbb{R}^{d_1 \times d_2} \times \mathbb{R}^{d_1 \times d_2} \to [0, 1]$. To streamline notation, we employ $\pi_t$ to represent the probability of selecting action $a_t = 1$ at time $t$, while $1 - \pi_t$ denotes the probability associated with selecting $a_t = 0$ accordingly.

The estimation and inference procedure introduced in this work is applicable to a wide range of randomized bandit policies, and here we list three examples.

- **$\varepsilon$-Greedy.** One widely used policy demonstrating the exploration-exploitation tradeoff is the $\varepsilon$-greedy approach (Lattimore and Szepesvári, 2020) which allocates $\varepsilon_t/2$ as the exploration probability while $1 - \varepsilon_t/2$ for exploitation at each iteration. With any pre-specified $\varepsilon_t \in (0, 1)$, $\pi_t$ can be explicitly expressed using $\varepsilon_t$. Specifically, probability of taking action $a_t = 1$ at time $t$ is described as

$$\mathbb{P}(a_t = 1|\mathcal{F}_{t-1}, X_t) = (1 - \varepsilon_t) I\left\{\langle \widehat{M}^{\mathbf{sgd}}_{1,t-1} - \widehat{M}^{\mathbf{sgd}}_{0,t-1}, X_t\rangle > 0\right\} + \frac{\varepsilon_t}{2}.$$

- **Softmax Policy.** Our proposed method can also be employed effectively with softmax policies that utilize exponential weighting schemes to balance exploration and exploitation. Consider the following probability model for choosing action $a_t = 1$,

$$\mathbb{P}(a_t = 1|\mathcal{F}_{t-1}, X_t) = \frac{\exp(\langle \widehat{M}^{\mathbf{sgd}}_{1,t-1}, X_t\rangle)}{\exp(\langle \widehat{M}^{\mathbf{sgd}}_{0,t-1}, X_t\rangle) + \exp(\langle \widehat{M}^{\mathbf{sgd}}_{1,t-1}, X_t\rangle)}.$$

  The action with a higher estimated reward is assigned with a higher probability through a softmax transformation. Popular applications include EXP3, EXP4 (Auer et al., 2002), and softmax policy gradient (Mei et al., 2020; Boutilier et al., 2020; Agarwal et al., 2021).

- **Thompson Sampling.** Thompson Sampling (Lattimore and Szepesvári, 2020) balances the exploration-exploitation trade-off by sampling from the posterior distribution over the expected reward for each action. At time $t$, the algorithm samples the matrix parameter $\bar{M}_{i,t}$ from the posterior distribution $\mathcal{P}^{(i)}(\cdot|\mathcal{F}_{t-1})$, and chooses the action to be the one that gives the maximum reward, i.e., $a_t = \arg\max_i \langle \bar{M}_{i,t}, X_t\rangle$. As the posterior distribution may not have an explicit form, approximate sampling could be employed and we discuss an adapted approach in the supplementary material.

Although our focus in the main paper remains on the aforementioned randomized policies with known action probabilities to enhance clarity, we also detail a methodology and accompanying theoretical analysis for scenarios where action probabilities are unknown. This discussion is provided in the supplementary material. These popular bandit algorithms typically select actions at time $t$ based on current estimations of model parameters. Therefore, an accurate estimation of $M_i$ enables more precise reward predictions, thereby enhancing the decision-making performance. In the following section, we introduce the methodology for deriving a sequential and sample-efficient estimator for $M_i$.

2.2. *Online Low-Rank Estimation via SGD.* In this section, we introduce the procedure to obtain the online low-rank estimator $\widehat{M}_{i,t}^{\mathbf{sgd}}$. The estimation method needs to meet two requirements: (1) the estimator should be updated sequentially under the online decision-making framework, and (2) the estimator should leverage the inherent low-rank structure to ensure sample efficiency. To accomplish these tasks, we apply SGD to iteratively update the estimation of the low-rank factorization of $M_i$. Specifically, for $i = 0, 1$, we solve the following stochastic optimization problem via SGD,

$$\min_{\mathcal{U}_i \in \mathbb{R}^{d_1 \times r}, \mathcal{V}_i \in \mathbb{R}^{d_2 \times r}} F(\mathcal{U}_i, \mathcal{V}_i) \;=\; \mathbb{E}\Big[ f(\mathcal{U}_i, \mathcal{V}_i; \{X, y\}) \Big], \tag{6}$$

where the expectation is taken with respect to the randomness of $\{X, y\}$, and the individual loss function is defined as

$$f(\mathcal{U}_i, \mathcal{V}_i; \{X, y\}) = \frac{1}{2} \left( y - \left\langle \mathcal{U}_i \mathcal{V}_i^\top, X \right\rangle \right)^2. \tag{7}$$

If we denote $\mathcal{U}_{i,t}$ and $\mathcal{V}_{i,t}$ as the estimated $\mathcal{U}_i$ and $\mathcal{V}_i$ at time $t$, respectively, a naive SGD approach for implementing the update at time $t$ with learning rate $\eta_t$ is given by

$$\begin{pmatrix} \mathcal{U}_{i,t} \\ \mathcal{V}_{i,t} \end{pmatrix} = \begin{pmatrix} \mathcal{U}_{i,t-1} \\ \mathcal{V}_{i,t-1} \end{pmatrix} - \eta_t I\{a_t = i\} \nabla f(\mathcal{U}_{i,t-1}, \mathcal{V}_{i,t-1}; \{X_t, y_t\}), \tag{8}$$

where $\nabla f$ is the gradient of the individual loss function in (7), i.e.,

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

However, this naive implementation is not applicable to our analysis for two reasons. First, the stochastic gradient given in the above form is no longer an unbiased estimator of the population gradient $\nabla F(\mathcal{U}_{i,t-1}, \mathcal{V}_{i,t-1})$ because this stochastic gradient depends on the adaptive distribution of $a_t$ while the population gradient does not. Second, our analysis requires that $\mathcal{U}_{i,t}$ and $\mathcal{V}_{i,t}$ stay in a neighborhood such that $F(\mathcal{U}_{i,t}, \mathcal{V}_{i,t})$ enjoys the smoothness and strong convexity, but this naive approach may destroy this geometric property of $F$ as discussed later in Section 2.3. To address the aforementioned two concerns, we propose our stochastic gradient as

$$\begin{aligned} &g(\mathcal{U}_{i,t-1}, \mathcal{V}_{i,t-1}; \{X_t, y_t, a_t, \pi_t\}) \\ =&\frac{I\{a_t = i\}}{i\pi_t + (1-i)(1-\pi_t)} \begin{pmatrix} (\langle \mathcal{U}_{i,t-1} \mathcal{V}_{i,t-1}^\top, X_t \rangle - y_t) X_t \mathcal{V}_{i,t-1} R_{\mathcal{V}} D_{\mathcal{V}}^{-\frac{1}{2}} Q_{\mathcal{V}} Q_{\mathcal{U}}^\top D_{\mathcal{U}}^{\frac{1}{2}} R_{\mathcal{U}}^\top \\ (\langle \mathcal{U}_{i,t-1} \mathcal{V}_{i,t-1}^\top, X_t \rangle - y_t) X_t \mathcal{U}_{i,t-1} R_{\mathcal{U}} D_{\mathcal{U}}^{-\frac{1}{2}} Q_{\mathcal{U}} Q_{\mathcal{V}}^\top D_{\mathcal{V}}^{\frac{1}{2}} R_{\mathcal{V}}^\top \end{pmatrix}. \end{aligned} \tag{9}$$

We describe the procedure of obtaining the above auxiliary matrices at each iteration in Algorithm 1. The inverse weight $1/[i\pi_t + (1-i)(1-\pi_t)]$ is applied to compensate for the bias in the naive stochastic gradient in (8) caused by the adaptive distribution of $a_t$, where we recall that $\pi_t$ is the shorthand notation for $\mathbb{P}(a_t = 1 | \mathcal{F}_{t-1}, X_t)$. Besides the inverse weighting, our form of $g$ also serves as a computationally efficient method for re-normalizing $\mathcal{U}_{i,t-1}$ and $\mathcal{V}_{i,t-1}$ to ensure that each iterate stays in a neighborhood. We provide more explanations and benefits of choosing $g$ as our stochastic gradient in Section 2.3. Given the designed stochastic gradient $g$, our updating rule is

$$\begin{pmatrix} \mathcal{U}_{i,t} \\ \mathcal{V}_{i,t} \end{pmatrix} = \begin{pmatrix} \mathcal{U}_{i,t-1} \\ \mathcal{V}_{i,t-1} \end{pmatrix} - \eta_t g(\mathcal{U}_{i,t-1}, \mathcal{V}_{i,t-1}; \{X_t, y_t, a_t, \pi_t\}), \tag{10}$$

where we require the learning rate $\eta_t$ to decay as $t$ grows to diminish the effect of the noise in the convergence analysis. We defer the discussion of the learning rate to Section 2.4. To further clarify this updating rule, we take $a_t = 1$ at time $t$ for example,

**Algorithm 1** One-Step SGD Update at time $t$

1: **Input**: $\mathcal{U}_{i,t-1}$, $\mathcal{V}_{i,t-1}$ for $i = 0, 1$, $X_t$, $y_t$, $a_t$, $\pi_t$, $\eta_t$

2: $R_{\mathcal{U}} D_{\mathcal{U}} R_{\mathcal{U}}^\top \leftarrow \text{SVD}\left(\mathcal{U}_{a_t,t-1}^\top \mathcal{U}_{a_t,t-1}\right)$, $R_{\mathcal{V}} D_{\mathcal{V}} R_{\mathcal{V}}^\top \leftarrow \text{SVD}\left(\mathcal{V}_{a_t,t-1}^\top \mathcal{V}_{a_t,t-1}\right)$.

3: $Q_{\mathcal{U}} D Q_{\mathcal{V}} \leftarrow \text{SVD}\left(D_{\mathcal{U}}^{\frac{1}{2}} R_{\mathcal{U}}^\top R_{\mathcal{V}} D_{\mathcal{V}}^{\frac{1}{2}}\right)$.

4: For $i = 0, 1$, update $\mathcal{U}_{i,t}$, $\mathcal{V}_{i,t}$ using (10).

5: **Output**: $\mathcal{U}_{i,t}$, $\mathcal{V}_{i,t}$, $R_{\mathcal{U}}$, $D_{\mathcal{U}}$, $R_{\mathcal{V}}$, $D_{\mathcal{V}}$

then $g(\mathcal{U}_{0,t-1}, \mathcal{V}_{0,t-1}; \{X_t, y_t, a_t, \pi_t\}) = (0,0)^\top$, which implies $\mathcal{U}_{0,t}$, $\mathcal{V}_{0,t}$ (for the action $a_t = 0$) are not updated. Meanwhile, the singular value decomposition (SVD) is applied to $\mathcal{U}_{1,t-1}^\top \mathcal{U}_{1,t-1}$ and $\mathcal{V}_{1,t-1}^\top \mathcal{V}_{1,t-1}$ after $\mathcal{U}_{1,t-1}$ and $\mathcal{V}_{1,t-1}$ are updated according to (10). The one-step update at time $t$ is summarized in Algorithm 1. Finally, we set $\widehat{M}_{i,t}^{\mathbf{sgd}} = \mathcal{U}_{i,t}\mathcal{V}_{i,t}^\top$, which will be used for the decision policy in the next iteration.

2.3. *Explanation of the Form of Stochastic Gradient.* We first discuss the necessity of applying the inverse weighting to compensate for the bias caused by the adaptive distribution of $a_t$. Then we discuss the necessity of renormalizing $\mathcal{U}_{i,t-1}$ and $\mathcal{V}_{i,t-1}$ at each time $t$. Finally, we demonstrate that Algorithm 1 only requires computing the SVD for an $r \times r$ matrix instead of a $d_1 \times d_2$ matrix at each iteration for re-normalization, which makes our algorithm computationally efficient.

As the SGD update is implemented under the online decision-making setting, the samples are collected through the action $a_t$ according to our decision-making policy at each time. This implies that the sample used for each update is not collected randomly but based on the "past experience" inherited in the distribution of $a_t$. Since the action $a_t$ determines either $(\mathcal{U}_{1,t}, \mathcal{V}_{1,t})$, or $(\mathcal{U}_{0,t}, \mathcal{V}_{0,t})$ to be updated at time $t$, we need to eliminate this bias so that the estimation for both $i = 0$ and 1 can be treated equally. Inspired by Chen, Lu and Song (2021a), we apply the inverse weight that serves as a distribution correction that compensates for the aforementioned bias using the fact $\mathbb{E}\left[I\{a_t = i\} | X_t, \mathcal{F}_{t-1}\right] = i\pi_t + (1-i)(1-\pi_t)$.

To ensure the convergence of our algorithm, we need $\mathcal{U}_{i,t}$ and $\mathcal{V}_{i,t}$ to stay in a local region. The naive implementation of SGD such as (8) might end up with an estimator $\mathcal{U}_{i,t}$ very large and $\mathcal{V}_{i,t}$ very small or vice versa even though $\mathcal{U}_{i,t}\mathcal{V}_{i,t}^\top$ is a reasonable estimate of $M_i$ (Jin, Kakade and Netrapalli, 2016). To see it, assuming we have matrices $A \in \mathbb{R}^{d_1 \times r}$ and $B \in \mathbb{R}^{d_2 \times r}$, then $AB^\top = \tilde{A}\tilde{B}^\top$ even if $\tilde{A}$ is very small while $\tilde{B}$ very large, e.g. $\tilde{A} = \delta A$ and $\tilde{B} = \delta^{-1} B$ for some very small scalar $\delta$. To avoid this situation, we can apply re-normalization at the beginning of each iteration by setting $\tilde{\mathcal{U}}_{a_t,t-1} = W_{\mathcal{U}} D^{\frac{1}{2}}$ and $\tilde{\mathcal{V}}_{a_t,t-1} = W_{\mathcal{V}} D^{\frac{1}{2}}$, where $W_{\mathcal{U}} D W_{\mathcal{V}}^\top$ is the top-$r$ SVD of $\mathcal{U}_{a_t,t-1}\mathcal{V}_{a_t,t-1}^\top$, meaning that $W_{\mathcal{U}}$ and $W_{\mathcal{V}}$ are the top-$r$ singular vectors. On the other hand, we leave $(\tilde{\mathcal{U}}_{1-a_t,t-1}, \tilde{\mathcal{V}}_{1-a_t,t-1})$ unchanged from the last iteration, i.e., $(\tilde{\mathcal{U}}_{1-a_t,t-1}, \tilde{\mathcal{V}}_{1-a_t,t-1}) = (\mathcal{U}_{1-a_t,t-1}, \mathcal{V}_{1-a_t,t-1})$. Then a straightforward way to deal with this concern is to plug the renormalized version $\tilde{\mathcal{U}}_{a_t,t-1}$ and $\tilde{\mathcal{V}}_{a_t,t-1}$ into (8) with the inverse weighting

$$\begin{pmatrix} \mathcal{U}_{i,t} \\ \mathcal{V}_{i,t} \end{pmatrix} = \begin{pmatrix} \tilde{\mathcal{U}}_{i,t-1} \\ \tilde{\mathcal{V}}_{i,t-1} \end{pmatrix} - \eta_t \frac{I\{a_t = i\}}{i\pi_t + (1-i)(1-\pi_t)} \nabla f(\tilde{\mathcal{U}}_{i,t-1}, \tilde{\mathcal{V}}_{i,t-1}; \{X_t, y_t\}). \tag{11}$$

In this case, the strong convexity and smoothness of $F$ can be guaranteed within the neighborhood of $(\tilde{\mathcal{U}}_{i,t-1}, \tilde{\mathcal{V}}_{i,t-1})$. Unfortunately, this naive approach requires computing the SVD of a $d_1 \times d_2$ matrix at each iteration, which incurs a huge computational cost. Nonetheless, the

low-rankness of $\mathcal{U}_{i,t}$ and $\mathcal{V}_{i,t}$ allows us to compute a cheaper SVD on $r \times r$ matrices $\mathcal{U}_{i,t}^\top \mathcal{U}_{i,t}$ and $\mathcal{V}_{i,t}^\top \mathcal{V}_{i,t}$ instead. The resulting alternative approach, described in Algorithm 1 using (9) as the stochastic gradient, handles the re-normalization issue in a computationally efficient way. It only remains to show the equivalency between (10) and (11), which demonstrates that the re-normalization can be done by applying the SVD of $r \times r$ matrices.

LEMMA 2.1 (Jin, Kakade and Netrapalli 2016). *The updating rules given by* (10) *and* (11) *are equivalent in the sense that, at any time* $t$*, the updates* $\mathcal{U}_{i,t}$*,* $\mathcal{V}_{i,t}$ *from* (10)*, and* $\mathcal{U}'_{i,t}$ *and* $\mathcal{V}'_{i,t}$ *from* (11)*, satisfy the relation* $\mathcal{U}'_{i,t}\mathcal{V}'^\top_{i,t} = \mathcal{U}_{i,t}\mathcal{V}_{i,t}^\top$.

Lemma 2.1 follows directly from Lemma 3.2 in Jin, Kakade and Netrapalli (2016), establishing computational equivalence between two SVD procedures. While the renormalization technique is adapted for computational efficiency, our statistical convergence analysis for stochastic gradient descent differs due to two reasons. Firstly, our framework encompasses noisy observations, where each stochastic gradient descent iteration does not progress toward a common minimizer. Secondly, our approach requires the integration of decision-making policies throughout data collection. These differences call for new tools to analyze the convergence of our low-rank estimation.

2.4. *Convergence Analysis of Low-Rank Estimation.* Before presenting the convergence results, we introduce the following assumptions for our true model.

ASSUMPTION 1. *We consider the reward model* (1)*. For* $i \in \{0, 1\}$*,*

(i) *The noise* $\xi_t$ *given* $a_t = i$ *are* i.i.d. *sub-Gaussian random variables with parameter* $\sigma_i$*,*

$$\mathbb{E}[\xi_t | a_t = i] = 0, \ \ \mathbb{E}[\xi_t^2 | a_t = i] = \sigma_i^2, \ \ \mathbb{E}[e^{s\xi_t} | a_t = i] \le e^{s^2\sigma_i^2}, \quad \forall s \in \mathbb{R}.$$

(ii) *The context matrix* $X_t$ *has* i.i.d *standard Gaussian entries, i.e.,* $X_t(j_1, j_2) \sim \mathcal{N}(0, 1)$*. Moreover,* $X_t$ *is independent from* $\mathcal{F}_{t-1}$ *and* $\xi_t$*, and* $\{X_t\}$ *are* i.i.d. *across all* $t$*.*

(iii) *The true matrix parameter* $M_i$ *is low-rank with rank* $r \ll \min\{d_1, d_2\}$*, and its condition number is* $\kappa(M_i) \le \kappa$ *for a positive constant* $\kappa$*.*

Assumption 1 indicates that the observed $y_t$ after taking action is corrupted by a sub-Gaussian noise with parameter $\sigma_i$, which is a common assumption in online decision-making literature (Lattimore and Szepesvári, 2020). Additionally, we assume the context matrix $X_t$ has *i.i.d.* standard Gaussian entries, which is a typical and convenient assumption in the low-rank matrix regression literature (Xia, 2019), and this contextual information received at each time is *i.i.d.* and independent from the noise. We note that the Gaussian condition is not exclusive and can be extended to include other distributions. For instance, in the supplementary material, we discuss an alternative design of the contextual matrix that can broaden the scope of our inference framework, moving beyond online low-rank regression to include the case of online low-rank matrix completion. Finally, we assume that the matrix is well conditioned with a known rank $r$, which is common in existing low-rank literature (Xia and Yuan, 2021; Zhu et al., 2022; Chen et al., 2019, 2024). A theoretical analysis for the case of unknown $r$ remains unclear even in the traditional matrix regression problems and deserves a careful investigation in future works.

We then discuss the initialization of $\mathcal{U}_i$ and $\mathcal{V}_i$ for $i = 0, 1$. Given a low-rank initialization $\widehat{M}_i^{\mathbf{init}}$ (i.e., $\widehat{M}_{i,0}^{\mathbf{sgd}}$), we can obtain $\mathcal{U}_{i,0}$ and $\mathcal{V}_{i,0}$ by applying the SVD on $\widehat{M}_i^{\mathbf{init}}$. We denote

$W_{\mathcal{U}}^{\mathbf{init}}$ and $W_{\mathcal{V}}^{\mathbf{init}}$ as the top-$r$ left and right singular vectors of $\widehat{M}_i^{\mathbf{init}}$, along with a diagonal matrix containing top-$r$ singular values denoted as $D^{\mathbf{init}}$. Then we set

$$\mathcal{U}_{i,0} = W_{\mathcal{U}}^{\mathbf{init}}(D^{\mathbf{init}})^{\frac{1}{2}}, \quad \text{and} \quad \mathcal{V}_{i,0} = W_{\mathcal{V}}^{\mathbf{init}}(D^{\mathbf{init}})^{\frac{1}{2}}. \tag{12}$$

For theoretical analysis, we require the following assumption on initialization.

ASSUMPTION 2. *With $\sigma_i$ specified in Assumption 1, the initialization $\widehat{M}_i^{\mathbf{init}}$ satisfies $\|\widehat{M}_i^{\mathbf{init}} - M_i\|_{\mathrm{F}} \le C\sigma_i$ for $i = 0, 1$, and some constant $C > 0$.*

The procedure of obtaining such initialization can be seen as the random exploration phase in the bandit problem. Since the samples are independent in the random exploration phase, such initialization condition is mild and can be satisfied by existing low-rank estimation literature (Xia, 2019).

ASSUMPTION 3. *The probabilities $\pi_t$ and $1 - \pi_t$, defined in Section 2.1, satisfy*

$$\min\{\pi_t, 1 - \pi_t\} \ge t^{-\beta} p_0,$$

*for some $0 \le \beta < 1$ and $p_0 \in (0, 1)$.*

This assumption ensures sufficient exploration by preventing the exploration probability from decaying too rapidly. When $\beta = 0$, it requires a constant lower bound $p_0$ for exploration, which is a common assumption in SGD-based inference (Chen, Lu and Song, 2021a; Chen et al., 2022). However, for estimation, Assumption 3 provides flexibility by allowing the lower bound of the exploration probability to decay over time for any $\beta > 0$ for the estimation resuls in this section and the policy value inference in Section 4.

With all these assumptions, we are ready to present the convergence result of our online low-rank estimation obtained through Algorithm 1. Recall that we define $d = \max\{d_1, d_2\}$ and set $\widehat{M}_{i,t}^{\mathbf{sgd}} = \mathcal{U}_{i,t}\mathcal{V}_{i,t}^{\top}$ at each iteration. To simplify the notations, we assume $\|M_0\| = \|M_1\| = 1$, and define $\lambda_r = \min\{\lambda_r(M_1), \lambda_r(M_0)\}$ with the condition number $\kappa \le 1/\lambda_r$.

THEOREM 2.2. *Define the learning rate $\eta_t = c \cdot (\max\{t, t^\star\})^{-\alpha}$, and $t^\star = \left(\gamma^2 dr \log^2 d\right)^{\frac{1}{\alpha-\beta}}$ for some constant $c > 0$ and $\alpha \in (\beta, 1)$. Assume the signal-to-noise ratio $\frac{\lambda_r}{\sigma_i} \ge 10C$ for some constant $C > 0$ and Assumptions 1–3 hold. For any large enough $\gamma > 0$, with probability at least $1 - \frac{4n}{d^\gamma}$, we have for $1 \le t \le n$,*

$$\left\|\widehat{M}_{i,t}^{\mathbf{sgd}} - M_i\right\|_{\mathrm{F}} \le C_1 \gamma \sigma_i \sqrt{\frac{dr \log^2 d}{t^{\alpha-\beta}}},$$

*for some positive constant $C_1$.*

REMARK 1. *Theorem 2.2 can be generalized to accommodate a relaxed initial condition $\|\widehat{M}_i^{\mathbf{init}} - M_i\|_{\mathrm{F}} \le C\lambda_r$. This generalization is formally stated in Theorem D.1 of the supplementary material. Specifically, if the initialization falls outside original region defined in Assumption 2 but within the relaxed one, a burn-in phase of estimation ensures that the same convergence rate can be achieved for sufficiently large $t$.*

When $\beta = 0$, the estimation error rate in Theorem 2.2 reduces to $\tilde{O}(\sqrt{dr/t^\alpha})$, ignoring the logarithm factors, which closely aligns with the statistically optimal rate in the offline setting (Xia, 2019) as one specifies $\alpha$ to be close to 1. For $\beta > 0$, the decision-making policy

allows for a decaying exploration probability, which may increase the estimation error but could benefit the decision-making objectives. Specifically, under an $\varepsilon$-greedy policy with $\varepsilon_t = p_0 t^{-\beta}$, the cumulative regret over a time horizon of $n$ is bounded by $\tilde{O}(n^{1-\frac{\alpha-\beta}{2}} + n^{1-\beta})$, ignoring logarithmic terms and dimensionality, where the two terms correspond to the regret due to exploitation and exploration, respectively. The parameter $\beta$ represents a tradeoff between online decision-making and the estimation error. Setting $\beta = \frac{1}{3}\alpha$ with $\alpha$ approaches 1, the cumulative regret is of the order $n^{2/3}$. A similar tradeoff in online decision making and parameter estimation has also been observed in Simchi-Levi and Wang (2023).

Having developed our online estimation method along with its associated error rate, we now proceed to present the framework for statistical inference. Section 3 details the methodology and theoretical foundation for parameter inference, while Section 4 focuses on inferring the optimal policy value.

**3. Parameter Inference.** In this section, we propose an online framework for conducting entry-wise statistical inference on the parameter $M_i$, which leverages the low-rank estimation from the earlier section. Particularly, we propose a sequential debiasing procedure that can obtain an unbiased estimator by removing the two types of bias inherited in $\widehat{M}^{\mathbf{sgd}}_{i,t}$ simultaneously as shown in Figure 3. We first introduce our proposed online debiasing procedure. We then present the asymptotic normality of our proposed unbiased estimator, which serves as the theoretical foundation for conducting the inference. Finally, we propose the estimation of the variance of this unbiased estimator and show the consistency of the estimator. It is worth pointing out that our estimation can be obtained in a fully online fashion without storing historical data.

3.1. *Online Debiasing Procedure.* As discussed in the existing low-rank matrix inference literature (Xia, 2019; Chen et al., 2019; Xia and Yuan, 2021), debiasing is a commonly used method that handles the bias caused by preserving the low-rankness. Unlike existing debiasing approaches, our debiasing procedure needs to deal with two sources of bias. First, even though the estimation method via SGD in Section 2.2 ensures that $\mathcal{U}_{i,t}$ and $\mathcal{V}_{i,t}$ are unbiased estimators for the corresponding low-rank factorization of $M_i$, there is no guarantee that $\mathcal{U}_{i,t}\mathcal{V}_{i,t}^\top$ is an unbiased estimator for $M_i$. Second, because the data collection is adaptive through the action $a_t$, we also need to handle the bias introduced by the adaptive samples in the bandit setting. To fill in the gap, we introduce a new debiasing procedure to eliminate both types of bias due to low-rankness and data adaptivity. The unbiased estimator obtained from our proposed online debiasing procedure is described as follows: taking $i = 1$ for example, we define

$$\widetilde{M}_{1,t} = \widehat{M}^{\mathbf{sgd}}_{1,t-1} + \frac{I\{a_t = 1\}}{\pi_t}(y_t - \langle \widehat{M}^{\mathbf{sgd}}_{1,t-1}, X_t\rangle)X_t,$$

at time $t$, and then update an online unbiased estimator

$$\widehat{M}^{\mathbf{unbs}}_{1,t} = (\widetilde{M}_{1,t} + (t-1)\widehat{M}^{\mathbf{unbs}}_{1,t-1})/t,$$

as the running average of $\widetilde{M}_{1,t}$. We apply the inverse weighting in $\widetilde{M}_{1,t}$ to compensate for the bias caused by the adaptive distribution of $a_t$. Additionally, $(y_t - \langle \widehat{M}^{\mathbf{sgd}}_{1,t-1}, X_t\rangle)X_t$ in the second term of $\widetilde{M}_{1,t}$ can be seen as the gradient of $f(M) = \frac{1}{2}(y_t - \langle M, X_t\rangle)^2$ at $\widehat{M}^{\mathbf{sgd}}_{1,t-1}$. This gradient does not impose low-rank constraint and thus pushes $\widehat{M}^{\mathbf{sgd}}_{1,t-1}$ towards the direction of an unbiased estimation of $M_1$. Moreover, it is important to note that we use $\widehat{M}^{\mathbf{sgd}}_{1,t-1}$ instead of $\widehat{M}^{\mathbf{sgd}}_{1,t}$ to obtain $\widetilde{M}_{1,t}$. Otherwise, $\widetilde{M}_{i,t}$ would no longer be an unbiased estimator of $M_i$

**Algorithm 2** One-Step Online Debiasing Update

1: **Input**: $\widehat{M}^{\mathbf{unbs}}_{i,t-1}$, $\widehat{M}^{\mathbf{sgd}}_{i,t-1}$, for $i = 0, 1$, $X_t$, $y_t$, $\pi_t$, $a_t$

2: For $i = 0, 1$, $\widetilde{M}_{i,t} \leftarrow \widehat{M}^{\mathbf{sgd}}_{i,t-1} + \frac{I\{a_t=i\}}{i\pi_t+(1-i)(1-\pi_t)}(y_t - \langle \widehat{M}^{\mathbf{sgd}}_{i,t-1}, X_t\rangle)X_t$.

3: $\widehat{M}^{\mathbf{unbs}}_{i,t} \leftarrow (\widetilde{M}_{i,t} + (t-1)\widehat{M}^{\mathbf{unbs}}_{i,t-1})/t$.

4: **Output**: $\widehat{M}^{\mathbf{unbs}}_{1,t}$, $\widehat{M}^{\mathbf{unbs}}_{0,t}$

because updating $\widehat{M}^{\mathbf{sgd}}_{1,t}$ uses the observation $X_t$, causing the dependence between $\widehat{M}^{\mathbf{sgd}}_{1,t}$ and $X_t$. Finally, we obtain our unbiased estimator for the inference purpose as

$$\widehat{M}^{\mathbf{unbs}}_{1,n} = \frac{1}{n}\sum_{t=1}^{n} \widehat{M}^{\mathbf{sgd}}_{1,t-1} + \frac{1}{n}\sum_{t=1}^{n} \frac{I\{a_t = 1\}}{\pi_t}(y_t - \langle \widehat{M}^{\mathbf{sgd}}_{1,t-1}, X_t\rangle)X_t, \tag{13}$$

which is essentially the average over $\widetilde{M}_{1,t}$. To see the unbiasness of $\widehat{M}^{\mathbf{unbs}}_{1,n}$ more formally, we define $\Delta_{t-1} = M_1 - \widehat{M}^{\mathbf{sgd}}_{1,t-1}$, and rewrite equation (13) by adding and subtracting $M_1$. With the definition of $y_t$ from (1), we then have

$$\widehat{M}^{\mathbf{unbs}}_{1,n} = M_1 + \underbrace{\frac{1}{n}\sum_{t=1}^{n} I\{a_t = 1\}\xi_t X_t/\pi_t}_{\widehat{Z}_1} + \underbrace{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{I\{a_t = 1\}\langle \Delta_{t-1}, X_t\rangle X_t}{\pi_t} - \Delta_{t-1}\right)}_{\widehat{Z}_2}.$$

Then both $\widehat{Z}_1$ and $\widehat{Z}_2$ are sum of martingale difference sequence by noting that for $\widehat{Z}_1$

$$\mathbb{E}\left[\frac{I\{a_t = 1\}}{\pi_t}\xi_t X_t \middle| \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{I\{a_t = 1\}}{\pi_t}\xi_t X_t \middle| \mathcal{F}_{t-1}, X_t\right] \middle| \mathcal{F}_{t-1}\right] = 0,$$

and similarly for $\widehat{Z}_2$, Assumption 1 implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{I\{a_t = 1\}\langle \Delta_{t-1}, X_t\rangle X_t}{\pi_t} - \Delta_{t-1} \middle| \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\frac{\langle \Delta_{t-1}, X_t\rangle X_t}{\pi_t}\mathbb{E}\left[I\{a_t = 1\} \middle| \mathcal{F}_{t-1}, X_t\right] - \Delta_{t-1} \middle| \mathcal{F}_{t-1}\right] = 0.
\end{aligned}$$

A similar debiasing procedure also applies to the case when $i = 0$ by replacing the $\pi_t$ by $(1 - \pi_t)$ due to the fact that $\mathbb{E}[I\{a_t = 0\}|X_t, \mathcal{F}_{t-1}] = 1 - \pi_t$. We summarize the online debiasing procedure at each time $t$ in Algorithm 2.

As we mentioned earlier, the debiasing procedure eliminates both sources of bias simultaneously disregarding maintaining the low-rankness. In this case, $\widehat{M}^{\mathbf{unbs}}_{i,n}$ obtained after $n$-iterations is not low-rank. Since the true parameter $M_i$ has a low-rank structure, we can apply a low-rank projection on the $\widehat{M}^{\mathbf{unbs}}_{i,n}$ by its left and right top-$r$ singular vectors to yield an improved estimate for the inference purpose, which is denoted as $\widehat{M}^{\mathbf{proj}}_{i,n}$. Recall that we target to conduct the statistical inference on $m^{(i)}_T = \langle M_i, T\rangle$ that we discussed in Section 1, the corresponding estimator for the inference purpose is defined as

$$\widehat{m}^{(i)}_T = \left\langle \widehat{M}^{\mathbf{proj}}_{i,n}, T \right\rangle. \tag{14}$$

While $\widehat{M}^{\mathbf{unbs}}_{i,n}$ serves as an unbiased estimator for $M_i$, it should be noted that $\widehat{M}^{\mathbf{proj}}_{i,n}$ does not necessarily possess this property. In theory, we can show that this additional bias in

$\widehat{m}_T^{(i)}$ is quantifiable and negligible under mild assumptions that we introduce in Section 3.2. Moreover, to obtain $\widehat{M}_{i,n}^{\mathbf{proj}}$, we need to compute the SVD for a $d_1 \times d_2$ matrix $\widehat{M}_{i,n}^{\mathbf{unbs}}$, and this computation is only required once after $n$-iterations. Because of its heavy computation cost, $\widehat{M}_{i,t}^{\mathbf{proj}}$ is not suitable for replacing the online estimator $\widehat{M}_{i,t}^{\mathbf{sgd}}$ for the decision-making purpose as $\widehat{M}_{i,t}^{\mathbf{sgd}}$ only requires computing the SVD of an $r \times r$ matrix at each iteration.

3.2. *Asymptotic normality of* $\widehat{m}_T^{(i)}$. We start the discussion on asymptotic normality by introducing several assumptions for the theoretical analysis. We denote $U_i$ and $V_i$ as the left and right singular vectors of the true matrix parameter $M_i$.

ASSUMPTION 4. *There exists a constant* $\alpha_T > 0$ *such that*

$$\alpha_T \|T\|_{\mathrm{F}} \sqrt{\frac{r}{d_1}} \le \|U_i^\top T\|_{\mathrm{F}}, \quad \alpha_T \|T\|_{\mathrm{F}} \sqrt{\frac{r}{d_2}} \le \|T V_i\|_{\mathrm{F}}.$$

To perform statistical inference for $m_T^{(i)} = \langle M_i, T \rangle$, Assumption 4 ensures that $T$ does not lie entirely in the null space of $M_i$ by imposing a lower bound on $\|U_i^\top T\|_{\mathrm{F}}$ and $\|T V_i\|_{\mathrm{F}}$.

ASSUMPTION 5. *There exists a constant* $\mu > 0$ *such that, for* $i \in \{0, 1\}$,

$$\max\left\{ \sqrt{\frac{d_1}{r}} \max_{j \in [d_1]} \|e_j^\top U_i\|, \sqrt{\frac{d_2}{r}} \max_{j \in [d_2]} \|e_j^\top V_i\| \right\} \le \mu.$$

Assumption 5 imposes an incoherence condition on the spectral space of the true matrix parameters $M_0, M_1$, indicating that their singular vectors should not be overly sparse. While not required to establish asymptotic normality, it simplifies the expression of the asymptotic distribution. Further discussion is provided in Section E.13 of the supplementary material.

ASSUMPTION 6. *As* $n, d_1, d_2 \to \infty$, *assume*

$$\max\left\{ \sqrt{\frac{dr \log^2 d}{n^\alpha}}, \ \frac{\sigma_i}{\lambda_r} \sqrt{\frac{d^2 r}{n}} \right\} \to 0,$$

*where* $\sigma_i$ *is defined in Assumption 1, and* $\alpha \in (0, 1)$ *is specified in Theorem 2.2. In addition, there exist constants* $\gamma$, $\gamma_d, \underline{\lambda} > 0$ *such that* $n = o(d^\gamma)$, $\lambda_r \ge \underline{\lambda}$, *and* $d_1/d_2 + d_2/d_1 \le \gamma_d$.

Assumption 6 requires conditions on the sample size and signal-to-noise ratio for reliable entry-level parameter inference. Under the additional assumption that the matrix $T$, which specifies the linear form under inference, is low-rank, the second condition may be relaxed to $(\sigma_i/\lambda_r)\sqrt{dr/n} = o(1)$. Section E.13 of the supplementary material outlines key supporting arguments for this relaxation, while a rigorous analysis is deferred to future work.

THEOREM 3.1. *Under Assumptions 1–6 with* $\beta = 0$, *and if we denote* $\pi_t(X) := \mathbb{P}(a_t = 1 | \mathcal{F}_{t-1}, X_t = X)$ *with* $\pi_t(X) \xrightarrow{p} \pi_\infty(X)$ *for any* $X$. *As* $n, d_1, d_2 \to \infty$, *we have*

$$\frac{\widehat{m}_T^{(i)} - m_T^{(i)}}{\sigma_i S_i / \sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1), \quad i = 0, 1,$$

*where*

$$S_i^2 = \int \frac{\left\langle U_{i,\perp} U_{i,\perp}^\top X V_i V_i^\top + U_i U_i^\top X V_{i,\perp} V_{i,\perp}^\top, T \right\rangle^2}{i\pi_\infty(X) + (1 - i)(1 - \pi_\infty(X))} dP_X,$$

Theorem 3.1 assumes $\beta = 0$ in Assumption 3, requiring the policy to maintain a constant lower bound $p_0$ for exploration. To ensure asymptotic normality of the parameter for each action, it mandates that each action is pulled sufficiently often to gather enough information for reliable parameter inference. As we will discuss in Section 4, the restriction on $\beta = 0$ can be relaxed for the inference of optimal policy value.

Theorem 3.1 provides a key insight: incorporating a debiasing step improves the estimation rate to $n^{-1/2}$. This improvement stems from the additional averaging performed during the debiasing procedure, which mitigates fluctuations across multiple iterates. As a result, the variance of the averaged sequence is reduced, leading to faster convergence. This acceleration behavior is analogous to the vector case studied in Polyak and Juditsky (1992).

The above result allows us to derive the asymptotic normality of the difference between two estimators. The following corollary demonstrates the asymptotic behavior of the difference between $\widehat{m}_T^{(1)} - \widehat{m}_T^{(0)}$, and thus provides the theoretical guarantee for the hypothesis testing mentioned in (3).

COROLLARY 3.2. *Under Assumptions of Theorem 3.1, as $n, d_1, d_2 \to \infty$, we have*

$$\frac{\left(\widehat{m}_T^{(1)} - \widehat{m}_T^{(0)}\right) - \left(m_T^{(1)} - m_T^{(0)}\right)}{\sqrt{(\sigma_1^2 S_1^2 + \sigma_0^2 S_0^2)/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

The intuition of proving Corollary 3.2 is that the main terms in $\widehat{m}_T^{(i)} - m_T^{(i)}$, $i = 0, 1$, are uncorrelated while the remainder terms are negligible. Therefore, the asymptotic variance of $(\widehat{m}_T^{(1)} - \widehat{m}_T^{(0)}) - (m_T^{(1)} - m_T^{(0)})$ is given by the sum of two individual variances.

3.3. *Parameter Inference.* With the asymptotic normality shown in Theorem 3.1, we are in a position to answer the inferential question about $m_T^{(i)}$ by constructing an online data-dependent confidence interval. In this section, we show that the asymptotic normality of $\widehat{m}_T^{(i)}$ remains valid after we replace $S_i^2$ and $\sigma_i^2$ by their estimators. To achieve this goal, we only need to prove the consistency of the proposed variance estimator.

Throughout this section, we use $\widehat{\mathrm{U}}_{i,t}$ and $\widehat{\mathrm{V}}_{i,t}$ to denote the left and right top-$r$ singular vectors of $\widehat{M}_{i,t}^{\mathbf{sgd}}$, and $\widehat{\mathrm{U}}_{i,t\perp}$, $\widehat{\mathrm{V}}_{i,t\perp}$ as their orthogonal complements. To obtain a consistent estimator of $S_i^2$ in Theorem 3.1, we need first to demonstrate that the $\widehat{\mathrm{U}}_{i,t}\widehat{\mathrm{U}}_{i,t}^\top$ and $\widehat{\mathrm{V}}_{i,t}\widehat{\mathrm{V}}_{i,t}^\top$ are consistent estimators for $U_i U_i^\top$ and $V_i V_i^\top$, where $U_i$ and $V_i$ denote the left and right top-$r$ singular vectors of $M_i$ respectively. Indeed, by the matrix perturbation theorem (Davis and Kahan, 1970; Wedin, 1972), for some positive constant $C$ we have

$$\max\left\{\|\widehat{\mathrm{U}}_{i,t}\widehat{\mathrm{U}}_{i,t}^\top - U_i U_i^\top\|_{\mathrm{F}}, \|\widehat{\mathrm{V}}_{i,t}\widehat{\mathrm{V}}_{i,t}^\top - V_i V_i^\top\|_{\mathrm{F}}\right\} \le C \cdot \frac{\|\widehat{M}_{i,t}^{\mathbf{sgd}} - M_i\|_{\mathrm{F}}}{\lambda_r}.$$

The convergence rate of $\widehat{M}_{i,t}^{\mathbf{sgd}}$ shown in Theorem 2.2 enables us to prove the consistency of the variance estimator, which leads to the following asymptotic normality of $\widehat{m}_T^{(i)}$ with the estimated $S_i^2$ and $\sigma_i^2$.

THEOREM 3.3. *Under Assumptions of Theorem 3.1, as $n, d_1, d_2 \to \infty$, we have*

$$\frac{\widehat{m}_T^{(i)} - m_T^{(i)}}{\hat{\sigma}_i \hat{S}_i / \sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1), \quad i = 0, 1,$$

*where*

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{t=1}^{n} \frac{I\{a_t = i\}}{i\pi_t + (1-i)(1-\pi_t)}(y_t - \langle \widehat{M}_{i,t-1}^{\mathbf{sgd}}, X_t\rangle)^2, \tag{15}$$

$$\hat{S}_i^2 = \frac{1}{n}\sum_{t=1}^{n} \frac{I\{a_t = i\}\left\langle \widehat{\mathsf{U}}_{i,t-1\perp}\widehat{\mathsf{U}}_{i,t-1\perp}^\top X_t \widehat{\mathsf{V}}_{i,t-1}\widehat{\mathsf{V}}_{i,t-1}^\top + \widehat{\mathsf{U}}_{i,t-1}\widehat{\mathsf{U}}_{i,t-1}^\top X_t \widehat{\mathsf{V}}_{i,t-1\perp}\widehat{\mathsf{V}}_{i,t-1\perp}^\top, T\right\rangle^2}{i\pi_t^2 + (1-i)(1-\pi_t)^2}. \tag{16}$$

It is worth pointing out that acquiring estimators $\hat{S}_i^2$ and $\hat{\sigma}_i^2$ only requires storing the partial sums instead of all historical data. At time $t$, estimators $\hat{S}_i^2$ and $\hat{\sigma}_i^2$ get updated by computing the running average of (15) and (16) for both $i = 0$ and 1, and note that only $\widehat{\mathsf{U}}_{a_t,t-1}\widehat{\mathsf{U}}_{a_t,t-1}^\top$ and $\widehat{\mathsf{V}}_{a_t,t-1}\widehat{\mathsf{V}}_{a_t,t-1}^\top$ need to be calculated at each iteration. We present the method of obtaining $\widehat{\mathsf{U}}_{a_t,t-1}\widehat{\mathsf{U}}_{a_t,t-1}^\top$ and $\widehat{\mathsf{V}}_{a_t,t-1}\widehat{\mathsf{V}}_{a_t,t-1}^\top$ in the fourth to the last line inside the *for* loop of Algorithm 3. Meanwhile, we can obtain the corresponding orthogonal complements used in (16) via

$$\widehat{\mathsf{U}}_{a_t,t-1\perp}\widehat{\mathsf{U}}_{a_t,t-1\perp}^\top = I - \widehat{\mathsf{U}}_{a_t,t-1}\widehat{\mathsf{U}}_{a_t,t-1}^\top, \quad \text{and} \quad \widehat{\mathsf{V}}_{a_t,t-1\perp}\widehat{\mathsf{V}}_{a_t,t-1\perp}^\top = I - \widehat{\mathsf{V}}_{a_t,t-1}\widehat{\mathsf{V}}_{a_t,t-1}^\top,$$

where $I$ denotes the identity matrix.

Given the result of Theorem 3.3, we can thus construct the data-dependent confidence interval for the true parameter $m_T^{(i)}$. In particular, at any confidence level $\alpha \in (0,1)$ we can construct the confidence interval

$$\left[\widehat{m}_T^{(i)} - z_{\alpha/2}\hat{\sigma}_i\hat{S}_i/\sqrt{n},\ \widehat{m}_T^{(i)} + z_{\alpha/2}\hat{\sigma}_i\hat{S}_i/\sqrt{n}\right], \tag{17}$$

where $z_{\alpha/2}$ denotes the standard score of normal distribution for the upper $\alpha/2$-quantile. The whole procedure of conducting the inference for $m_T^{(i)}$ is summarized in Algorithm 3. It is also worth pointing out that due to Corollary 3.2, we extend the result of Theorem 3.3 to

$$\frac{(\widehat{m}_T^{(1)} - \widehat{m}_T^{(0)}) - (m_T^{(1)} - m_T^{(0)})}{\sqrt{(\hat{\sigma}_0^2\hat{S}_0^2 + \hat{\sigma}_1^2\hat{S}_1^2)/n}} \xrightarrow{d} \mathcal{N}(0,1),$$

which allows us to test the difference in effectiveness between the actions.

**4. Inference for Optimal Policy Value.** In this section, we investigate the statistical inference of optimal policy value as defined in (4). In contrast with Section 3, which requires the exploration probability to be lower bounded by constant, we relax this condition by permitting the exploration probability to gradually diminish over time for optimal policy value inference. Echoing the debiasing technique outlined in Equation (13) from Section 3.1, we adopt a similar strategy to develop an estimator for inferring the optimal policy value. The construction of this estimator also incorporates a correction term designed for bias reduction. Due to space limitations, this section focuses on scenarios where exploration probabilities are known. We defer the optimal policy value inference procedure when these probabilities are unknown yet estimated to Section A of the supplementary material.

4.1. *Estimator for Optimal Policy Value.* We now present our estimator for the optimal policy value. This estimator after $n$ iterations is defined as follows:

$$\widehat{V}_n = \frac{1}{n}\sum_{t=1}^{n}\left\langle \widehat{M}_{\hat{a}(X_t),t-1}^{\mathbf{sgd}}, X_t\right\rangle + \frac{1}{n}\sum_{t=1}^{n}\frac{I\{a_t = \hat{a}(X_t)\}}{1-e_t}\left(y_t - \left\langle \widehat{M}_{\hat{a}(X_t),t-1}^{\mathbf{sgd}}, X_t\right\rangle\right), \tag{18}$$

**Algorithm 3** Online Inference of $m_T^{(i)}$

**Input**: $\widehat{M}_1^{\mathbf{init}}, \widehat{M}_0^{\mathbf{init}}, \mathcal{U}_{i,0}, \mathcal{V}_{i,0}\ r$.
**Initialization**: $\widehat{M}_{i,0}^{\mathbf{unbs}} \leftarrow \widehat{M}_i^{\mathbf{init}}$, $\widehat{M}_{i,0}^{\mathbf{sgd}} \leftarrow \widehat{M}_i^{\mathbf{init}}$, for $i = 0, 1$.
**for** $t \leftarrow 1$ *to* $n$ **do**
  Observe a contextual matrix $X_t$.
  Compute $\pi_t$ according to the policy.
  Decide the action $a_t$ by $Ber(\pi_t)$.
  Receive reward $y_t$ according to (1).
  For $i = 0, 1$, $\widehat{M}_{i,t}^{\mathbf{unbs}} \leftarrow$ Algorithm 2 ($\widehat{M}_{i,t-1}^{\mathbf{unbs}}, \widehat{M}_{i,t-1}^{\mathbf{sgd}}, X_t, y_t, a_t, \pi_t$).
  $\mathcal{U}_{i,t}, \mathcal{V}_{i,t}, R_{\mathcal{U}}, D_{\mathcal{U}}, R_{\mathcal{V}}, D_{\mathcal{V}} \leftarrow$ Algorithm 1 ($\mathcal{U}_{i,t-1}, \mathcal{V}_{i,t-1}, X_t, y_t, a_t, \pi_t$).
  $\widehat{\mathsf{U}}_{a_t,t-1}\widehat{\mathsf{U}}_{a_t,t-1}^\top \leftarrow R_{\mathcal{U}} D_{\mathcal{U}}^{-1} R_{\mathcal{U}}^\top$, $\widehat{\mathsf{V}}_{a_t,t-1}\widehat{\mathsf{V}}_{a_t,t-1}^\top \leftarrow R_{\mathcal{V}} D_{\mathcal{V}}^{-1} R_{\mathcal{V}}^\top$.
  $\widehat{\mathsf{U}}_{a_t,t-1\perp}\widehat{\mathsf{U}}_{a_t,t-1\perp}^\top \leftarrow I - \widehat{\mathsf{U}}_{a_t,t-1}\widehat{\mathsf{U}}_{a_t,t-1}^\top$, $\widehat{\mathsf{V}}_{a_t,t-1\perp}\widehat{\mathsf{V}}_{a_t,t-1\perp}^\top \leftarrow I - \widehat{\mathsf{V}}_{a_t,t-1}\widehat{\mathsf{V}}_{a_t,t-1}^\top$.
  Update $\hat{\sigma}_i^2$ and $\hat{S}_i^2$ by computing the running average of (15) and (16).
  $\widehat{M}_{i,t}^{\mathbf{sgd}} \leftarrow \mathcal{U}_{i,t}\mathcal{V}_{i,t}^\top$.
Compute the top-$r$ singular vectors of $\widehat{M}_{i,n}^{\mathbf{unbs}}$ to obtain $\widehat{M}_{i,n}^{\mathbf{proj}}$, and then we calculate $\widehat{m}_T^{(i)}$ by (14).
Obtain the confidence interval as (17).

where

$$\hat{a}(X_t) = I\{\langle \widehat{M}_{1,t-1}^{\mathbf{sgd}} - \widehat{M}_{0,t-1}^{\mathbf{sgd}}, X_t \rangle > 0\}, \tag{19}$$

and $e_t := 1 - \mathbb{P}(a_t = \hat{a}(X_t)|\mathcal{F}_{t-1}, X_t)$. In the formation of this optimal policy value estimator, $\hat{a}(X_t)$ represents the estimated optimal action at time $t$, and $e_t$ represents the probability for exploration. To elaborate, if $\hat{a}(X_t) = 1$, the exploration probability becomes $e_t = \mathbb{P}(a_t = 0|\mathcal{F}_{t-1}, X_t) = 1 - \pi_t$. Similar to the debiasing process used in parameter inference described in (13), we also employ inverse probability weighting to correct distributional bias in this scenario. However, there is a key distinction: in parameter inference, the weighting factor is derived from the probability of taking each possible action, while here it suffices to use only the exploitation probability for the inverse weighting. This distinction arises because bias correction in parameter inference leverages samples gathered from each action individually. In the case of the optimal policy value estimator, however, we exclusively use samples collected from the estimated optimal action, regardless of whether it is action 1 or 0, to formulate this bias reduction. This forms the key reason that we allow a relaxed exploration probability in this section.

In Equation (18), we can view the first term as a direct estimator for the optimal policy value. However, relying on this direct estimate exclusively can lead to potential failure when $\widehat{M}_{i,t}^{\mathbf{sgd}}$ does not offer an accurate estimate of $M_i$. In the context of our study, where $\widehat{M}_{i,t}^{\mathbf{sgd}}$ is inherently biased, the latter term of (18) serves as a corrective mechanism, functioning in a manner analogous to how we formulated $\widehat{M}_{i,t}^{\mathbf{unbs}}$ in Section 3. For optimal policy value inference, samples contributed to the estimation should be selectively obtained from the exploitation part, which explains the reason that our estimator presented in (18) only takes the samples generated by the estimated optimal action.

4.2. *Asymptotic Normality.* We start the discussion on the asymptotic normality of the optimal policy value estimator (18) by introducing the following assumptions.

ASSUMPTION 7. *For $\alpha$ in the learning rate specified in Theorem 2.2 and $\beta$ specified in Assumption 3 such that $\alpha - \beta > \frac{1}{2}$, as $n, d_1, d_2 \to \infty$,*

$$\max\left\{\sqrt{\frac{dr\log^2 d}{n^{\alpha-\beta}}},\ \frac{\sigma_i\|M_1 - M_0\|_{\mathrm{F}}^{-1} dr\log^2 d}{n^{\alpha-\beta-\frac{1}{2}}}\right\} \to 0.$$

*In addition, there exist constants $\gamma$, $\gamma_d > 0$ such that $n = o(d^\gamma)$ and $d_1/d_2 + d_2/d_1 \le \gamma_d$.*

Assumption 7 consists of two components: the first part ensures that $\widehat{M}_i^{\mathbf{sgd}}$ serves as a consistent estimator of $M_i$, and the second condition ensures that the gap between $M_1$ and $M_0$ is sufficiently large compared to the noise, making the optimal action distinguishable. With these considerations, we are now prepared to discuss the asymptotic normality of $\sqrt{n}(\widehat{V}_n - V^*)$.

THEOREM 4.1. *Under the conditions of Theorem 2.2 and Assumption 7, if we denote $e_t^*(X) = \mathbb{P}(a_t \neq a^*(X_t)|\mathcal{F}_{t-1}, X_t = X)$ with $e_t^*(X) \xrightarrow{p} e_\infty^*(X)$ for any $X$. Then as $n, d_1, d_2 \to \infty$, we have*

$$\frac{\widehat{V}_n - V^*}{S_V/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

*where*

$$S_V^2 = \int \frac{a^*(X)\sigma_1^2 + (1-a^*(X))\sigma_0^2}{1 - e_\infty^*(X)} dP_X + \mathrm{Var}_X\left[\langle M_{a^*(X)}, X\rangle\right].$$

Theorem 4.1 establishes the asymptotic normality of our proposed optimal policy value estimator. This asymptotic variance consists of two distinct components. The first term in $S_V^2$ serves as the weighted average variance of the noise, conditional on the optimal action for a given context. On the other hand, the second term in $S_V^2$ captures the variance associated with the context. If the estimated optimal action $\hat{a}(X_t)$ converges to the true optimal action $a^*(X_t)$, then the weight assigned to the first component of $S_V^2$ is determined by the limiting probability associated with exploitation. Note that the asymptotic probability of exploration $e_\infty^*(X)$ is allowed to be zero in this scenario, which marks the fundamental difference from the parameter inference in Theorem 3.1.

4.3. *Optimal Policy Value Inference.* With the asymptotic normality introduced in Theorem 4.1, we next construct a valid confidence interval for the optimal policy value. We first propose the empirical estimator for $S_V^2$ in a fully online fashion without requiring any storage for $d_1 \times d_2$ context matrix $X_t$. Define the online estimator as

$$\begin{aligned}\widehat{S}_V^2 =& \frac{1}{n}\sum_{t=1}^n \frac{\hat{\sigma}_{1,t}^2 I\left\{\left\langle \widehat{M}_{1,t-1}^{\mathbf{sgd}} - \widehat{M}_{0,t-1}^{\mathbf{sgd}}, X_t\right\rangle > 0\right\} + \hat{\sigma}_{0,t}^2 I\left\{\left\langle \widehat{M}_{1,t-1}^{\mathbf{sgd}} - \widehat{M}_{0,t-1}^{\mathbf{sgd}}, X_t\right\rangle \le 0\right\}}{1 - e_t} \\ &+ \frac{1}{n}\sum_{t=1}^n \left\langle \widehat{M}_{\hat{a}(X_t),t-1}^{\mathbf{sgd}}, X_t\right\rangle^2 - \left(\frac{1}{n}\sum_{t=1}^n \left\langle \widehat{M}_{\hat{a}(X_t),t-1}^{\mathbf{sgd}}, X_t\right\rangle\right)^2, \end{aligned} \tag{20}$$

where for $i = 0, 1$,

$$\hat{\sigma}_{i,t}^2 = \frac{1}{t}\sum_{s=1}^t \frac{I\{a_s = i\}}{i\pi_s + (1-i)(1-\pi_s)}\left(y_s - \left\langle \widehat{M}_{i,s-1}^{\mathbf{sgd}}, X_s\right\rangle\right)^2. \tag{21}$$

**Algorithm 4** Online Inference of Optimal Policy Value $V^*$

**Input**: $\widehat{M}_1^{\mathbf{init}}$, $\widehat{M}_0^{\mathbf{init}}$, $\mathcal{U}_{i,0}$, $\mathcal{V}_{i,0}$, $r$.
**Initialization**: $\widehat{M}_{i,0}^{\mathbf{sgd}} \leftarrow \widehat{M}_i^{\mathbf{init}}$, for $i = 0, 1$.
**for** $t \leftarrow 1$ *to* $n$ **do**
  Observe a contextual matrix $X_t$.
  Obtain $\pi_t = \mathbb{P}(a_t = 1|\mathcal{F}_{t-1}, X_t)$ according to the decision-making policy.
  Update $\hat{a}(X_t)$ by equation (19), and calculate $e_t \leftarrow 1 - \mathbb{P}(a_t = \hat{a}(X_t)|\mathcal{F}_{t-1}, X_t)$.
  Decide the action $a_t$ by $Ber(\pi_t)$.
  $\mathcal{U}_{i,t}, \mathcal{V}_{i,t} \leftarrow$ Algorithm 1 $(\mathcal{U}_{i,t-1}, \mathcal{V}_{i,t-1}, X_t, y_t, a_t, \pi_t)$
  $\widehat{M}_{i,t}^{\mathbf{sgd}} \leftarrow \mathcal{U}_{i,t}\mathcal{V}_{i,t}^\top$.
  Get the estimator value $\widehat{V}_t$ by equation (18).
  Update the variance estimator $\widehat{S}_V^2$ by equation (20).
Obtain the two-sided confidence interval with critical value $z$: $(\widehat{V}_n - z\widehat{S}_V/\sqrt{n},\ \ \widehat{V}_n + z\widehat{S}_V/\sqrt{n})$.

It is important to note that the running summation in (20) and (21) can be sequentially updated. Theorem 4.2 below shows that $\widehat{S}_V^2$ is a consistent estimator for $S_V^2$, and thus the asymptotic normality is also guaranteed with the estimated variance.

THEOREM 4.2. *Under the conditions of Theorem 4.1, we have $\widehat{S}_V^2$ is a consistent estimator of $S_V^2$, i.e., $\widehat{S}_V^2 \xrightarrow{p} S_V^2$. Furthermore, as $n, d_1, d_2 \to \infty$, we have*

$$\frac{\widehat{V}_n - V^*}{\widehat{S}_V/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

In light of Theorem 4.2, constructing a confidence interval for the optimal policy value $V^*$ becomes feasible. This opens the door to hypothesis testing to evaluate the performance of the currently available actions in achieving a desired level of outcome, even under the optimal policy. This addresses inferential questions posed in Equation (5). Unlike the parameter inference discussed in Section 3, which necessitates computing the SVD for a $d_1 \times d_2$ matrix at the end of the online sequence for low-rank projection, the value inference approach introduced in this section sidesteps the computational overhead associated with SVD calculations. Finally, we summarize the optimal policy value inference procedure in Algorithm 4.

**5. Simulation Studies.** In this section, we present extensive numerical studies to evaluate the performance of our online inference procedure. In the presented synthetic simulations, we consider a Gaussian noise $\xi_t|a_t = i \sim N(0, \sigma_i^2)$ with the noise level $\sigma_i = 0.1$ for both $i = 0, 1$. We generate the true low-rank matrices $M_1$ and $M_0$ with rank $r = 3$, and dimensions $d = d_1 = d_2 = 50$. The singular vectors, $U_i, V_i \in \mathbb{R}^{d\times r}$, are generated from the singular space of random Gaussian matrices. We set top-$r$ singular values of $M_i$ to be 1, i.e., $\lambda_1(M_i) = \lambda_2(M_i) = \lambda_3(M_i) = 1$. For the simulation study of the parameter inference, we adopt $\varepsilon$-greedy policy with $\varepsilon = 0.1$. The additional simulation results for optimal value inference with $\varepsilon \to 0$ are illustrated in Section B of the supplementary material. We set the learning rate $\eta_t = 0.1(\max\{t, t^\star\})^{-0.99}$ with $t^\star = 300$. Finally, the initialization $\widehat{M}_i^{\mathbf{init}}$ is obtained from a nuclear-norm penalized estimation (Negahban and Wainwright, 2011) with pre-collected offline data.

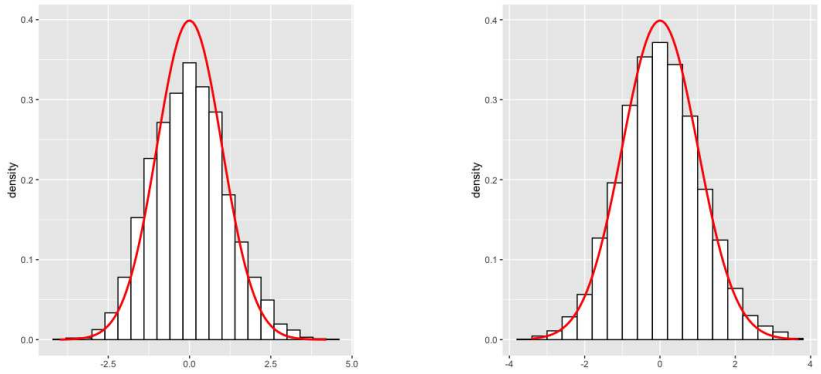
We first validate the asymptotic normality of $\widehat{m}_T^{(i)}$ with $T = e_1 e_1^\top$ by plotting the histogram of $\sqrt{n}(\widehat{m}_T^{(i)} - m_T^{(i)})/\hat{\sigma}_i\hat{S}_i$ from 5000 independent trails with $n = 1000$ and 3000. We present the histogram of $\sqrt{n}(\widehat{m}_T^{(i)} - m_T^{(i)})/\hat{\sigma}_i\hat{S}_i$ for $i = 1$ in Figure 4. The result for $i = 0$ is similar

TABLE 1
*Coverage Probability, Average Confidence Interval Length and corresponding standard deviation for the scenario $T = T_1$ and $T = T_2$ based on 5000 independent trails.*

| | | | Coverage Probability | Average CI Length |
|---|---|---|---|---|
| $T_1$ | $n = 1000$ | $i = 0$ | 0.909 | 0.018 |
| | | $i = 1$ | 0.913 | 0.010 |
| | $n = 2000$ | $i = 0$ | 0.923 | 0.013 |
| | | $i = 1$ | 0.925 | 0.008 |
| | $n = 3000$ | $i = 0$ | 0.929 | 0.011 |
| | | $i = 1$ | 0.936 | 0.006 |
| $T_2$ | $n = 1000$ | $i = 0$ | 0.906 | 0.065 |
| | | $i = 1$ | 0.908 | 0.042 |
| | $n = 2000$ | $i = 0$ | 0.924 | 0.048 |
| | | $i = 1$ | 0.923 | 0.031 |
| | $n = 3000$ | $i = 0$ | 0.931 | 0.039 |
| | | $i = 1$ | 0.930 | 0.026 |

TABLE 2
*Coverage Probability, Average Confidence Interval Length for $r = 3, 5, 7$ for $T = T_1$ and $n = 3000$ based on 5000 independent trails.*

| | | Coverage Probability | Average CI Length |
|---|---|---|---|
| $r = 3$ | $i = 0$ | 0.929 | 0.011 |
| | $i = 1$ | 0.936 | 0.006 |
| $r = 5$ | $i = 0$ | 0.917 | 0.015 |
| | $i = 1$ | 0.921 | 0.014 |
| $r = 7$ | $i = 0$ | 0.913 | 0.021 |
| | $i = 1$ | 0.906 | 0.021 |

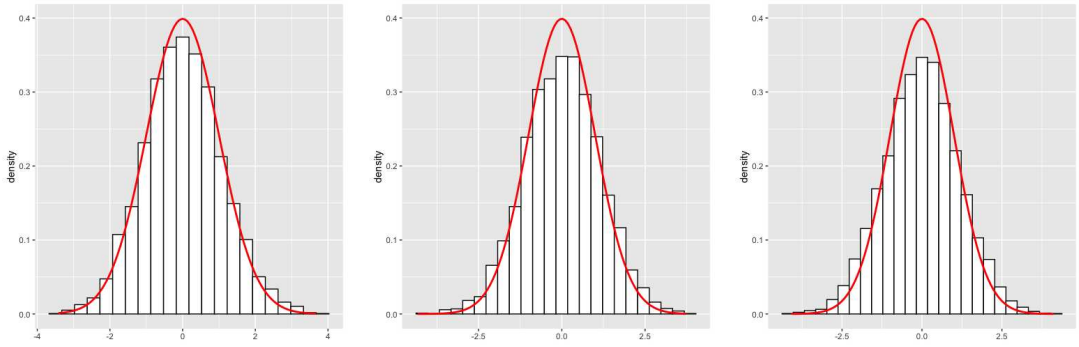



(a) $n = 1000$, $r = 3$ (b) $n = 3000$, $r = 3$

Fig 4: Empirical distribution of $\sqrt{n}(\widehat{m}_T^{(1)} - m_T^{(1)})/\hat{\sigma}_1\hat{S}_1$ based on 5000 independent trails for $T = e_1e_1^\top$. The red curve refers to the density of standard normal.

and hence is omitted. As shown in Figure 4, as $n$ increases, the empirical distribution of $\sqrt{n}(\widehat{m}_T^{(i)} - m_T^{(i)})/\hat{\sigma}_i\hat{S}_i$ gets closer to the standard normal distribution.

In Table 1, we present the coverage probability and average confidence interval length in two scenarios with $T = T_1 = e_1e_1^\top$ and $T = T_2 = e_1e_1^\top + 2e_2e_2^\top - 3e_3e_3^\top$. The coverage probability is calculated as the ratio of the 5000 independent trails that fall into $\left(\widehat{m}_T^{(i)} - 1.96\hat{\sigma}_i\hat{S}_i, \widehat{m}_T^{(i)} + 1.96\hat{\sigma}_i\hat{S}_i\right)$, which is the 95% confidence interval constructed by the




(a) $n = 3000$, $r = 3$ (b) $n = 3000$, $r = 5$ (c) $n = 3000$, $r = 7$

Fig 5: Empirical distribution of $\sqrt{n}(\widehat{m}_T^{(1)} - m_T^{(1)})/\hat{\sigma}_1\hat{S}_1$ based on 5000 independent trails for ranks $r = 3,\ 5,\ 7$ and $T = e_1 e_1^\top$.

standard deviation estimation. The interval length is calculated as $2 \times 1.96\hat{\sigma}_i\hat{S}_i$. We present the result as $n = 1000$, 2000, and 3000. As shown in Table 1, for both $T_1$ and $T_2$, as $n$ grows, the coverage probability is closer to 0.95, and the confidence interval length decreases. In addition, when we increase the $\|T\|_{\mathrm{F}}$, i.e., from $\|T_1\|_{\mathrm{F}}$ to $\|T_2\|_{\mathrm{F}}$, the true $S_i$ gets larger which causes the average length of confidence interval increases.

In Table 2, we compare the converge probability and the average confidence interval lengths across different true ranks $r$. As the rank $r$ increases, the coverage probability shrinks, and the confidence interval length increases. We also compare the histograms for $r = 3, 5, 7$ in Figure 5, and the normal approximation gets slightly worse as the true rank increases.

**Acknowledgment.** The authors thank the editor Professor Lan Wang, the associate editor and three anonymous reviewers for their valuable comments and suggestions which led to a much improved paper. Will Wei Sun acknowledges support from the National Science Foundation (SES 2217440). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not reflect the views of the National Science Foundation.

## SUPPLEMENTARY MATERIAL

**Supplement to "Online Statistical Inference in Decision Making with Matrix Context"**. The supplementary material includes extensions of optimal policy value inference, additional numerical studies, and discussions on the assumptions. Finally, it provides proofs of all theoretical results and supporting technical lemmas.

## REFERENCES

AGARWAL, A., KAKADE, S., LEE, J. and MAHAJAN, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research* **22** 4431–4506.

AGRAWAL, S. and GOYAL, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*.

AKROUT, M., FARAHMAND, A.-M., JARMAIN, T. and ABID, L. (2019). Improving skin condition classification with a visual symptom checker trained using reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* **32** 48–77.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973.

BIAN, Z., SHI, C., QI, Z. and WANG, L. (2024). Off-policy evaluation in doubly inhomogeneous environments. *Journal of the American Statistical Association* 1–27.

BIBAUT, A., DIMAKOPOULOU, M., KALLUS, N., CHAMBAZ, A. and VAN DER LAAN, M. (2021). Post-contextual-bandit inference. *Advances in Neural Information Processing Systems* **34** 28548–28559.

BOUTILIER, C., HSU, C.-W., KVETON, B., MLADENOV, M., SZEPESVARI, C. and ZAHEER, M. (2020). Differentiable meta-learning of bandit policies. *Advances in Neural Information Processing Systems* **33** 2122–2134.

CANDES, E. J. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory* **57** 2342–2359.

CARPENTIER, A. and KIM, A. K. (2018). An iterative hard thresholding estimator for low rank matrix recovery with explicit limiting distribution. *Statistica Sinica* **28** 1371–1393.

CARPENTIER, A., EISERT, J., GROSS, D. and NICKL, R. (2015). Uncertainty quantification for matrix compressed sensing and quantum tomography problems. In *High Dimensional Probability VIII* 385–430. Springer.

CHEN, H., LU, W. and SONG, R. (2021a). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association* **116** 708–719.

CHEN, H., LU, W. and SONG, R. (2021b). Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association* **116** 240–255.

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences* **116** 22931–22937.

CHEN, X., LEE, J. D., TONG, X. T. and ZHANG, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics* **48** 251–273.

CHEN, X., LAI, Z., LI, H. and ZHANG, Y. (2021). Online statistical inference for stochastic optimization via Kiefer-Wolfowitz methods. *arXiv preprint arXiv:2102.03389*.

CHEN, X., LAI, Z., LI, H. and ZHANG, Y. (2022). Online statistical inference for contextual bandits via stochastic gradient descent. *arXiv preprint arXiv:2212.14883*.

CHEN, E. Y., XIA, D., CAI, C. and FAN, J. (2024). Semi-parametric tensor factor analysis by iteratively projected singular value decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* **7** 1–46.

DELIU, N., WILLIAMS, J. J. and CHAKRABORTY, B. (2022). Reinforcement learning in modern biostatistics: constructing optimal adaptive interventions. *arXiv preprint arXiv:2203.02605*.

DESHPANDE, Y., JAVANMARD, A. and MEHRABI, M. (2023). Online debiasing for adaptively collected high-dimensional data with applications to time series analysis. *Journal of the American Statistical Association* **118** 1126–1139.

DESHPANDE, Y., MACKEY, L., SYRGKANIS, V. and TADDY, M. (2018). Accurate inference for adaptive linear models. In *International Conference on Machine Learning*. PMLR.

FANG, E. X., WANG, Z. and WANG, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association* **118** 1733–1746.

FANG, Y., XU, J. and YANG, L. (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research* **19** 3053–3073.

HADAD, V., HIRSHBERG, D. A., ZHAN, R., WAGER, S. and ATHEY, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences* **118** e2014602118.

HALL, P. and HEYDE, C. C. (1980). *Martingale limit theory and its application*. Academic press.

ISTEPANIAN, R., LAXMINARAYAN, S. and PATTICHIS, C. S. (2007). *M-health: Emerging mobile health systems*. Springer Science & Business Media.

JAIN, P. and PAL, S. (2022). Online low rank matrix completion. *arXiv preprint arXiv:2209.03997*.

JIN, C., KAKADE, S. M. and NETRAPALLI, P. (2016). Provable efficient online matrix completion via non-convex stochastic gradient descent. *Advances in Neural Information Processing Systems*.

KHAMARU, K., DESHPANDE, Y., MACKEY, L. and WAINWRIGHT, M. J. (2021). Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*.

KOLTCHINSKII, V. and XIA, D. (2015). Optimal estimation of low rank density matrices. *The Journal of Machine Learning Research* **16** 1757–1792.

KOREN, Y. (2009). The bellkor solution to the netflix grand prize. *Netflix Prize Documentation* **81** 1–10.

KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annual Review of Statistics and its Application* **6** 263–286.

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: isoperimetry and processes* **23**. Springer Science & Business Media.

LI, L., LU, Y. and ZHOU, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*.

LI, L., CHU, W., LANGFORD, J. and SCHAPIRE, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web* 661–670.

LI, Y., XIE, H., LIN, Y. and LUI, J. C. (2021). Unifying offline causal inference and online bandit learning for data driven decision. In *Proceedings of the Web Conference 2021* 2291–2303.

LIU, W., TU, J., ZHANG, Y. and CHEN, X. (2023). Online estimation and inference for robust policy evaluation in reinforcement learning. *arXiv preprint arXiv:2310.02581*.

LU, Y., MEISAMI, A. and TEWARI, A. (2021). Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics* 460–468. PMLR.

LU, X. and VAN ROY, B. (2017). Ensemble sampling. *Advances in neural information processing systems* **30**.

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics* **44** 713.

MEI, J., XIAO, C., SZEPESVARI, C. and SCHUURMANS, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning* 6820–6829. PMLR.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 1069–1097.

POLDRACK, R. A., MUMFORD, J. A. and NICHOLS, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* **30** 838–855.

QI, Z., PANG, J.-S. and LIU, Y. (2023). On robustness of individualized decision rules. *Journal of the American Statistical Association* **118** 2143–2157.

RAMPRASAD, P., LI, Y., YANG, Z., WANG, Z., SUN, W. W. and CHENG, G. (2023). Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association* **118** 2901–2914.

RUSSO, D. J., VAN ROY, B., KAZEROUNI, A., OSBAND, I. and WEN, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning* **11** 1–96.

SHI, L., WANG, J. and WU, T. (2023). Statistical Inference on Multi-armed Bandits with Delayed Feedback.

SHI, C., LUO, S., ZHU, H. and SONG, R. (2021a). An online sequential test for qualitative treatment effects. *Journal of Machine Learning Research* **22** 1–51.

SHI, C., SONG, R., LU, W. and LI, R. (2021b). Statistical inference for high-dimensional models via recursive online-score estimation. *Journal of the American Statistical Association* **116** 1307–1318.

SHI, C., ZHANG, S., LU, W. and SONG, R. (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84** 765–793.

SHI, C., WANG, X., LUO, S., ZHU, H., YE, J. and SONG, R. (2023). Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association* **118** 2059–2071.

SHI, C., ZHU, J., YE, S., LUO, S., ZHU, H. and SONG, R. (2024). Off-policy confidence interval estimation with confounded Markov decision process. *Journal of the American Statistical Association* **119** 273–284.

SIMCHI-LEVI, D. and WANG, C. (2023). Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics*.

TANG, K., LIU, W., ZHANG, Y. and CHEN, X. (2023). Acceleration of stochastic gradient descent with momentum by averaging: finite-sample rates and asymptotic normality. *arXiv preprint arXiv:2305.17665*.

WEDIN, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* **12** 99–111.

XIA, D. (2019). Confidence region of singular subspaces for low-rank matrix regression. *IEEE Transactions on Information Theory* **65** 7437–7459.

XIA, D. (2021). Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics* **15** 3798–3851.

XIA, D. and YUAN, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83** 58–77.

YE, S., CAI, H. and SONG, R. (2023). Doubly robust interval estimation for optimal policy evaluation in online learning. *Journal of the American Statistical Association* 1–20.

ZHAN, R., HADAD, V., HIRSHBERG, D. A. and ATHEY, S. (2021). Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference*.

ZHANG, K., JANSON, L. and MURPHY, S. (2020). Inference for batched bandits. *Advances in neural information processing systems*.

ZHANG, K., JANSON, L. and MURPHY, S. (2021). Statistical inference with M-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*.

ZHANG, K. W., JANSON, L. and MURPHY, S. A. (2022). Statistical inference after adaptive sampling in non-markovian environments. *arXiv preprint arXiv:2202.07098*.

ZHOU, J., HAO, B., WEN, Z., ZHANG, J. and SUN, W. W. (2024). Stochastic low-rank tensor bandits for multi-dimensional online decision making. *Journal of the American Statistical Association* 1-24.

ZHU, W., CHEN, X. and WU, W. B. (2023). Online covariance matrix estimation in stochastic gradient descent. *Journal of the American Statistical Association* **118** 393–404.

ZHU, Z., LI, X., WANG, M. and ZHANG, A. (2022). Learning Markov models via low-rank optimization. *Operations Research* **70** 2384–2398.