Appendices for Reward Imputation with Sketching for Contextual Batched Bandits

Xiao Zhang^{1,2}, **Ninglu Shao**^{1,2,*}, **Zihua Si**^{1,2,*}, **Jun Xu**^{1,2,†},
Wenhan Wang^{3}, **Hanjing Su**^{3}, **Ji-Rong Wen**^{1,2}
^{1} Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
^{2} Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
^{3} Tencent Inc., Shenzhen, China
`{zhangx89, ninglu_shao, zihua_si, junxu, jrwen}@ruc.edu.cn`
`{justinsu, ezewang}@tencent.com`

Abstract

In these appendices, we give the detailed problem formulation and proof in the problem analysis (Appendix A), present the proof of the proposed theoretical results as well as their dependence structure (Figure 1 in Appendix B), provide more details and results in the experiments (Appendix C), and discuss the potential social impact and limitations of our work (Appendix D). Our main theoretical results include:

- The instantaneous regret bound in full-information CBB Setting (see Appendix A.2).
- The instantaneous regret bound of our reward imputation approach (see Appendix B.1).
- The approximation properties of sketching which are necessary to achieve approximation error bound of the sketched ridge regression (See Appendix B.2).
- The approximation error bound of reward imputation using sketching (see Appendix B.3).
- The regret bound of SPUIR (see Appendix B.4).
- The regret bounds of the extensions of SPUIR in Section "Extensions of Our Approach", including SPUIR-Exp, SPUIR-Poly, and SPUIR-Kernel (see Appendix B.5).

A Detailed Problem Formulation and Proof in Problem Analysis

In this part, we give the detailed problem formulation and the detailed proof of Theorem 1 in problem analysis in Section 2.

A.1 Detailed Problem Formulation of CBB

In this paper, we focus on the setting of contextual batched bandits (CBB), which can be formulated as a 6-tuple $\langle \mathcal{S}, \mathcal{A}, p, R, N, B \rangle$:

Context space $\mathcal{S} \subseteq \mathbb{R}^d$ means a vector space containing the context information received at each step, e.g., context summarizes the information of both the user and items in recommendation scenarios.

---

[*] Ninglu Shao and Zihua Si have made equal contributions to this paper.

[†] Corresponding author: Jun Xu.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Algorithm 1 Batch UCB Policy Updating in the $(n+1)$-th episode in Full-Information CBB Setting

```
INPUT: Policy $p_n$, data buffer $\mathcal{D}_{n+1}$, action space $\mathcal{A} = \{A_j\}_{j \in [M]}$, $\boldsymbol{\theta}^0_{A_j} = \mathbf{0}$, $j \in [M]$, batch size $B$
OUTPUT: Updated policy $p_{n+1}$
Let $\widetilde{\boldsymbol{L}}_n \in \mathbb{R}^{(n+1)B \times d}$ be the matrix that stores all the context vectors till the $n$-th episode as the row vectors
For $\forall A \in \mathcal{A}$, let $\widetilde{\boldsymbol{T}}^n_A \in \mathbb{R}^{(n+1)B}$ be the reward vector that stores all the rewards of action $A \in \mathcal{A}$ till the $n$-th episode
// Policy Updating
$\boldsymbol{\Upsilon}_{n+1} \leftarrow \widetilde{\boldsymbol{L}}_n^\intercal \widetilde{\boldsymbol{L}}_n$
for all action $A \in \mathcal{A}$ do
    $\boldsymbol{\theta}^{n+1}_A \leftarrow (\boldsymbol{I}_d + \boldsymbol{\Upsilon}_{n+1})^{-1} \widetilde{\boldsymbol{L}}_n^\intercal \widetilde{\boldsymbol{T}}^n_A$
end for
For a new context $\boldsymbol{s}$, $p_{n+1}(\boldsymbol{s})$ is to choose the action following: $A \leftarrow \arg\max_{A \in \mathcal{A}} \langle \boldsymbol{\theta}^{n+1}_A, \boldsymbol{s} \rangle$
Return $\{\boldsymbol{\theta}^{n+1}_A\}_{A \in \mathcal{A}}$
```

Action space $\mathcal{A} = \{A_j\}_{j \in [M]}$ contains M candidate actions for execution. As an example, in recommender systems, each action corresponds to a candidate item, and selecting an action means that the corresponding item is recommended.

Policy p determines which action to take at each step, which is a function of the context $\boldsymbol{s} \in \mathcal{S}$ and outputs an action for execution (or a selection distribution over action space $\mathcal{A}$).

Reward R in CBB is a *partial-information feedback* where rewards are unobserved for the non-executed actions. Consider a stochastic bandit setting, where the expectation of the true reward is assumed to be a function of the context $\boldsymbol{s} \in \mathcal{S}$. In particular, different from the shared expectation function of true rewards in existing batch bandits (Han et al., 2020), we assume that the expectation functions of true rewards are different for each action, where each expectation function corresponds to an unknown parameter vector $\boldsymbol{\theta}^*_A \in \mathbb{R}^d$, $A \in \mathcal{A}$. This setting for rewards matches many real-world applications, e.g., each action corresponds to a different category of candidate coupons in coupon recommendation.

Number of episodes N. The decision process in CBB is partitioned into N episodes. Within one episode, the agent updates the policy using the collected data, and then interacts with the environment for multiple steps using the updated and fixed policy.

Batch size B is the number of steps in each episode. That is, in each episode, the agent interacts with the environment B times using a fixed policy, and stores the contexts, executed actions, and observed rewards into a data buffer $\mathcal{D}$ at the end of each episode.

A.2 Detailed Description and Proof of Theorem 1 in Problem Analysis

We present some theoretical findings about the regret difference between the partial-information feedback and the full-information feedback. Assuming that the agent in CBB setting can observe the rewards of all the candidate actions from the environment at each step, we apply the batched UCB policy (Han et al., 2020) to this setting (see Algorithm 1). We demonstrate an instantaneous regret bound in Theorem A.1, where Theorem A.1 is a detailed version of Theorem 1 in Section 2.

Theorem A.1 (Instantaneous Regret Bound in Full-Information CBB Setting, Detailed Version of Theorem 1)**.** *Let $\widetilde{\boldsymbol{L}}_{n-1} \in \mathbb{R}^{nB \times d}$ be the matrix that stores all the context vectors till the $(n-1)$-th episode as the row vectors, and $\widetilde{\boldsymbol{T}}^{n-1}_A \in \mathbb{R}^{nB}$ be the reward vector that stores all the rewards of action $A \in \mathcal{A}$ till the $(n-1)$-th episode. Given the action space $\mathcal{A} = \{A_j\}_{j \in [M]}$, in the n-th episode, assume that the rewards are independent and bounded by C_R. Then, with probability at least $1-\delta$, for any $b \in [B]$ and $\forall A \in \mathcal{A}$, we have the following instantaneous regret bound in the n-th episode*

$$|\langle \boldsymbol{\theta}^n_A, \boldsymbol{s}_{n,b} \rangle - \langle \boldsymbol{\theta}^*_A, \boldsymbol{s}_{n,b} \rangle| \leq \left[\|\boldsymbol{\theta}^*_A\|_2 + \sqrt{2C_R^2 \log(2MB/\delta)} \right] \sqrt{\boldsymbol{s}^\intercal_{n,b} \left(\boldsymbol{I}_d + \boldsymbol{\Upsilon}_n\right)^{-1} \boldsymbol{s}_{n,b}}, \tag{1}$$

where $\boldsymbol{\Upsilon}_n = \widetilde{\boldsymbol{L}}^\intercal_{n-1} \widetilde{\boldsymbol{L}}_{n-1}$ and the parameter of reward model $\boldsymbol{\theta}^n_A$ in the batched UCB policy is obtained by

$$\boldsymbol{\theta}^n_A := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\| \widetilde{\boldsymbol{L}}_{n-1} \boldsymbol{\theta} - \widetilde{\boldsymbol{T}}^{n-1}_A \right\|_2^2 + \|\boldsymbol{\theta}\|_2^2 = \left(\boldsymbol{I}_d + \boldsymbol{\Upsilon}_n\right)^{-1} \widetilde{\boldsymbol{L}}^\intercal_{n-1} \widetilde{\boldsymbol{T}}^{n-1}_A.$$

Further, the instantaneous regret bound (1) *in FI-CBB setting is tighter than that in CBB setting (i.e., using the partial-information feedback). In particular, the variance term* $\sqrt{\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}}$ *is smaller than that in CBB setting.*

Proof of Theorem A.1. By the formulation of $\boldsymbol{\theta}_A^n$ and the triangle inequality, we first obtain that

$$\begin{aligned}
&\left|\langle\boldsymbol{\theta}_A^n,\boldsymbol{s}_{n,b}\rangle-\langle\boldsymbol{\theta}_A^*,\boldsymbol{s}_{n,b}\rangle\right| \\
=&\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\widetilde{\boldsymbol{T}}_A^{n-1}-\boldsymbol{s}_{n,b}^{\mathsf{T}}\boldsymbol{\theta}_A^*\right| \\
=&\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\left[\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\widetilde{\boldsymbol{T}}_A^{n-1}-\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)\boldsymbol{\theta}_A^*\right]\right| \\
=&\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\left[\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\widetilde{\boldsymbol{T}}_A^{n-1}-\left(\boldsymbol{I}_d+\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\widetilde{\boldsymbol{L}}_{n-1}\right)\boldsymbol{\theta}_A^*\right]\right| \\
=&\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\left(\widetilde{\boldsymbol{T}}_A^{n-1}-\widetilde{\boldsymbol{L}}_{n-1}\boldsymbol{\theta}_A^*\right)-\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{\theta}_A^*\right| \\
\leq&\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\left(\widetilde{\boldsymbol{T}}_A^{n-1}-\widetilde{\boldsymbol{L}}_{n-1}\boldsymbol{\theta}_A^*\right)\right|+\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{\theta}_A^*\right|
\end{aligned}\tag{2}$$

Next, we bound the two terms in the last row of (2).

Bounding $\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\left(\widetilde{\boldsymbol{T}}_A^{n-1}-\widetilde{\boldsymbol{L}}_{n-1}\boldsymbol{\theta}_A^*\right)\right|$**:**

Since $\mathrm{E}\left[\widetilde{\boldsymbol{T}}_A^{n-1}\right]=\widetilde{\boldsymbol{L}}_{n-1}\boldsymbol{\theta}_A^*$ and the received rewards are independent, by the Azuma-Hoeffding bound, we have

$$\begin{aligned}
&\Pr\left\{\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\left(\widetilde{\boldsymbol{T}}_A^{n-1}-\widetilde{\boldsymbol{L}}_{n-1}\boldsymbol{\theta}_A^*\right)\right|\geq\nu\sqrt{\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}}\right\} \\
&\leq 2\exp\left\{-\frac{\nu^2\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}}{2C_R^2\|\widetilde{\boldsymbol{L}}_{n-1}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}\|_2^2}\right\},
\end{aligned}\tag{3}$$

where $\nu>0$ is some constant. Since

$$\begin{aligned}
\|\widetilde{\boldsymbol{L}}_{n-1}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}\|_2^2&=\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\widetilde{\boldsymbol{L}}_{n-1}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b} \\
&\leq\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\left(\boldsymbol{I}_d+\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\widetilde{\boldsymbol{L}}_{n-1}\right)\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b} \\
&\leq\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b} \\
&=\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b},
\end{aligned}$$

combing with (3) implies the following results

$$\begin{aligned}
&\Pr\left\{\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\left(\widetilde{\boldsymbol{T}}_A^{n-1}-\widetilde{\boldsymbol{L}}_{n-1}\boldsymbol{\theta}_A^*\right)\right|\geq\nu\sqrt{\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}}\right\} \\
&\leq 2\exp\left\{-\frac{\nu^2}{2C_R^2}\right\}.
\end{aligned}\tag{4}$$

Combing (4) with the union bound, yields that, with probability at least $1-\delta$, for any $b\in[B]$ and $\forall A\in\mathcal{A}$,

$$\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\widetilde{\boldsymbol{L}}_{n-1}^{\mathsf{T}}\left(\widetilde{\boldsymbol{T}}_A^{n-1}-\widetilde{\boldsymbol{L}}_{n-1}\boldsymbol{\theta}_A^*\right)\right|\leq\nu\sqrt{\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}},\tag{5}$$

where the failure probability is

$$\delta=2MB\exp\left\{-\frac{\nu^2}{2C_R^2}\right\},$$

yielding that $\nu=\sqrt{2C_R^2\log(2MB/\delta)}$.

Bounding $\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{\theta}_A^*\right|$**:**

Since $\boldsymbol{\Upsilon}_n$ is positive semi-definite, combining with the Hölder inequality, we obtain

$$\begin{aligned}\left|\boldsymbol{s}_{n,b}^{\mathsf{T}}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{\theta}_A^*\right| &\le \|\boldsymbol{\theta}_A^*\|_2\left\|\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}\right\|_2\\ &= \|\boldsymbol{\theta}_A^*\|_2\sqrt{\boldsymbol{s}_{n,b}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}}\\ &\le \|\boldsymbol{\theta}_A^*\|_2\sqrt{\boldsymbol{s}_{n,b}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}}\\ &= \|\boldsymbol{\theta}_A^*\|_2\sqrt{\boldsymbol{s}_{n,b}\left(\boldsymbol{I}_d+\boldsymbol{\Upsilon}_n\right)^{-1}\boldsymbol{s}_{n,b}}.\end{aligned} \tag{6}$$

Combing (5) and (6) concludes the proof.

Similarly to the proof of (19), we can obtain that the variance term in full-information setting is smaller than that in partial-information setting. ☐

B Detailed Proofs in Theoretical Analysis

In this section, we provide the instantaneous regret bound in each episode, prove the approximation error of sketching, and analyze the regret for policy updating in CBB setting. Figure 1 describes the dependence structure of our theoretical results.

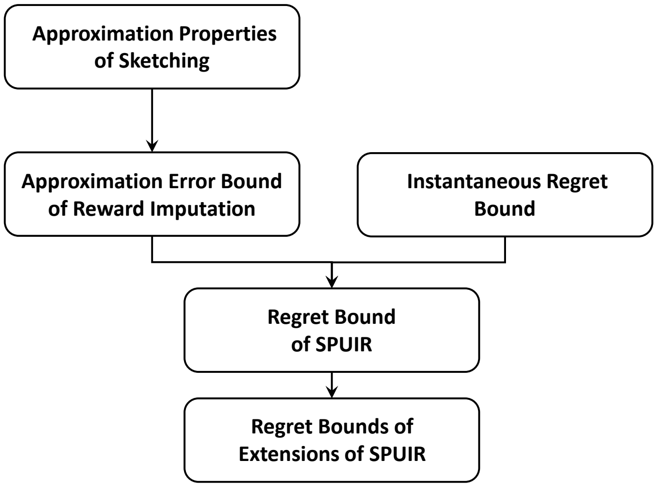

Approximation Properties of Sketching
Approximation Error Bound of Reward Imputation
Instantaneous Regret Bound
Regret Bound of SPUIR
Regret Bounds of Extensions of SPUIR

Figure 1: The dependence structure of our theoretical results, where the proof of instantaneous regret bound (Theorem 2 in the manuscript) is provided in Appendix B.1, the analysis of approximation properties of sketching (Theorem B.1) is given in Appendix B.2, the approximation error bound of reward imputation (Theorem 3 in the manuscript) is proven in Appendix B.3, the regret of SPUIR (Theorem 4 in the manuscript) is analyzed in Appendix B.4, and the regret bounds of the extensions of SPUIR (Corollary B.1) are provided in Appendix B.5

B.1 Proof of Theorem 2

Before we provide the detailed proof of Theorem 2 in the manuscript, we first demonstrate a lemma about the convergence and monotonicity of the sum of functions, which is the main tool for analyzing the additional bias of reward imputation.

Lemma B.1 (Convergence and Monotonicity). *Let $f(n)=\sum_{j=1}^{n}a^{n-j}\cdot g(j)$, where $a\in(0,1)$ and n is a positive integer. Then,*

1) when $g(j)$ is convergent, the limit $\lim_{n\to\infty}f(n)$ exists. Moreover,

$$\lim_{n\to\infty}f(n)=\frac{1}{1-a}\lim_{n\to\infty}g(n). \tag{7}$$

2) $f(n)$ is a monotonic decreasing function if and only if $g(j)$ satisfies, for any positive integer $j \geq 2$,

$$g(j) \leq \begin{cases} (j-1)a^{j-1}g(1) & a = 1/2, \\ \dfrac{(1-a)\left[a^{j-1} - (1-a)^{j-1}\right]}{2a-1} g(1) & a \neq 1/2. \end{cases} \tag{8}$$

Proof of Lemma B.1. Letting $b(j) = a^{-j} \cdot g(j), \forall j \in [n]$, and $S(n) = \sum_{j=1}^{n} b(j)$, $f(n)$ can be rewritten as $f(n) = a^n S(n)$.

1) Rewriting $f(n) = S(n)/a^{-n}$, from the Stolz's theorem, we have

$$\begin{aligned} \lim_{n\to\infty} f(n) &= \lim_{n\to\infty} \frac{S(n) - S(n-1)}{a^{-n} - a^{-(n-1)}} \\ &= \lim_{n\to\infty} \frac{b(n)}{a^{-n} - a^{-(n-1)}} \\ &= \lim_{n\to\infty} \frac{a^{-n} \cdot g(n)}{a^{-n} - a^{-(n-1)}} \\ &= \frac{1}{1-a} \lim_{n\to\infty} g(n). \end{aligned}$$

2) The condition that $f(\cdot)$ is a monotonic decreasing function is equivalent to the following condition: for any positive integer n,

$$\begin{aligned} f(n+1) \leq f(n) &\Leftrightarrow a^{n+1}S(n+1) \leq a^n S(n) \\ &\Leftrightarrow a[S(n) + b(n+1)] \leq S(n) \\ &\Leftrightarrow b(n+1) \leq (1/a - 1)\, S(n). \end{aligned} \tag{9}$$

From the equivalent condition (9), we obtain the following recursion formula:

$$\begin{aligned} b(n+1) &\leq (1/a-1)\, S(n) \\ b(n) &\leq (1/a-1) \sum_{j=1}^{n-1} b(j) \\ &\vdots \\ b(3) &\leq (1/a-1)\, [b(1) + b(2)] \\ b(2) &\leq (1/a-1)\, b(1), \end{aligned}$$

yielding that, for any positive integer $j \geq 2$,

$$b(j) \leq \left[(1/a-1) + (1/a-1)^2 + \cdots + (1/a-1)^{j-1}\right] b(1). \tag{10}$$

From (10) , for $a \neq 1/2$,

$$b(j) \leq \frac{(1/a-1)\left[1-(1/a-1)^{j-1}\right]}{1-(1/a-1)} b(1) = \frac{(1-a)\left[1-(1/a-1)^{j-1}\right]}{2a-1} b(1), \tag{11}$$

and substituting the definition of $b(j)$ into (11) yields the equivalent condition

$$g(j) \leq \frac{(1-a)\left[1-(1/a-1)^{j-1}\right]}{2a-1} a^{j-1}\, g(1).$$

For $a = 1/2$, we have the condition $b(j) \leq (j-1)b(1)$, which is equivalent to

$$a^{-j} \cdot g(j) \leq (j-1)a^{-1} \cdot g(1) \;\Leftrightarrow\; g(j) \leq (j-1)a^{j-1}g(1).$$

□

Next, we provide the detailed proof of Theorem 2 in the manuscript.

Proof of Theorem 2. From the formulation of $\bar{\boldsymbol{\theta}}_A^n$ and the triangle inequality, we can obtain that, for each action $A \in \mathcal{A}$,

$$\begin{aligned}
&\left|\left\langle \bar{\boldsymbol{\theta}}_A^n, \boldsymbol{s}_{n,b}\right\rangle - \left\langle \boldsymbol{\theta}_A^*, \boldsymbol{s}_{n,b}\right\rangle\right| \\
={}& \left|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\left(\boldsymbol{b}_A^n + \gamma \widehat{\boldsymbol{b}}_A^n\right) - \boldsymbol{s}_{n,b}^{\intercal}\boldsymbol{\theta}_A^*\right| \\
={}& \left|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\left[\left(\boldsymbol{L}_A^{n-1}\right)^{\intercal}\boldsymbol{T}_A^{n-1} + \gamma\left(\widehat{\boldsymbol{L}}_A^{n-1}\right)^{\intercal}\widehat{\boldsymbol{T}}_A^{n-1} - \boldsymbol{\Psi}_A^n\boldsymbol{\theta}_A^*\right]\right| \\
={}& \left|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\left[\left(\boldsymbol{L}_A^{n-1}\right)^{\intercal}\boldsymbol{T}_A^{n-1} + \gamma\left(\widehat{\boldsymbol{L}}_A^{n-1}\right)^{\intercal}\widehat{\boldsymbol{T}}_A^{n-1} - \left(\lambda\boldsymbol{I}_d + \boldsymbol{\Phi}_A^n + \gamma\widehat{\boldsymbol{\Phi}}_A^n\right)\boldsymbol{\theta}_A^*\right]\right| \\
={}& \Bigg|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\Bigg\{\left(\boldsymbol{L}_A^{n-1}\right)^{\intercal}\boldsymbol{T}_A^{n-1} + \gamma\left(\widehat{\boldsymbol{L}}_A^{n-1}\right)^{\intercal}\widehat{\boldsymbol{T}}_A^{n-1} - \\
&\left[\lambda\boldsymbol{I}_d + \left(\boldsymbol{L}_A^{n-1}\right)^{\intercal}\boldsymbol{L}_A^{n-1} + \gamma\left(\widehat{\boldsymbol{L}}_A^{n-1}\right)^{\intercal}\widehat{\boldsymbol{L}}_A^{n-1}\right]\boldsymbol{\theta}_A^*\Bigg\}\Bigg| \\
={}& \Bigg|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\left(\boldsymbol{L}_A^{n-1}\right)^{\intercal}\left(\boldsymbol{T}_A^{n-1} - \boldsymbol{L}_A^{n-1}\boldsymbol{\theta}_A^*\right) - \lambda\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\boldsymbol{\theta}_A^* + \\
&\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\gamma\left(\widehat{\boldsymbol{L}}_A^{n-1}\right)^{\intercal}\left(\widehat{\boldsymbol{T}}_A^{n-1} - \widehat{\boldsymbol{L}}_A^{n-1}\boldsymbol{\theta}_A^*\right)\Bigg| \\
\leq{}& \underbrace{\left|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\left(\boldsymbol{L}_A^{n-1}\right)^{\intercal}\left(\boldsymbol{T}_A^{n-1} - \boldsymbol{L}_A^{n-1}\boldsymbol{\theta}_A^*\right)\right|}_{X_A^{(1)}} + \lambda\underbrace{\left|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\boldsymbol{\theta}_A^*\right|}_{X_A^{(2)}} + \\
&\underbrace{\left|\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\gamma\left(\widehat{\boldsymbol{L}}_A^{n-1}\right)^{\intercal}\left(\widehat{\boldsymbol{T}}_A^{n-1} - \widehat{\boldsymbol{L}}_A^{n-1}\boldsymbol{\theta}_A^*\right)\right|}_{X_A^{(3)}}.
\end{aligned}$$

Next, we bound $X_A^{(1)}$, $X_A^{(2)}$, and $X_A^{(3)}$. For convenience, we drop all the superscripts and subscripts about n and b. Similarly to the proof of Theorem 1, we bound $X_A^{(1)} + X_A^{(2)}$ as follows: with probability at least $1-\delta$,

$$X_A^{(1)} + X_A^{(2)} \leq (\lambda\|\boldsymbol{\theta}_A^*\|_2 + \nu)\sqrt{\boldsymbol{s}^{\intercal}\boldsymbol{\Psi}_A^{-1}\boldsymbol{s}}, \tag{12}$$

where $\nu = \sqrt{2C_R^2\log(2MB/\delta)}$. For $X_A^{(3)}$, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
X_A^{(3)} &\leq \gamma\left\|\widehat{\boldsymbol{L}}_A\boldsymbol{\Psi}_A^{-1}\boldsymbol{s}\right\|_2\left\|\widehat{\boldsymbol{T}}_A - \widehat{\boldsymbol{L}}_A\boldsymbol{\theta}_A^*\right\|_2 \\
&= \sqrt{\gamma}\sqrt{\boldsymbol{s}^{\intercal}\boldsymbol{\Psi}_A^{-1}\left(\gamma\widehat{\boldsymbol{L}}_A^{\intercal}\widehat{\boldsymbol{L}}_A\right)\boldsymbol{\Psi}_A^{-1}\boldsymbol{s}}\,\left\|\widehat{\boldsymbol{T}}_A - \widehat{\boldsymbol{L}}_A\boldsymbol{\theta}_A^*\right\|_2 \\
&\leq \sqrt{\gamma}\sqrt{\boldsymbol{s}^{\intercal}\boldsymbol{\Psi}_A^{-1}\boldsymbol{s}}\,\left\|\widehat{\boldsymbol{T}}_A - \widehat{\boldsymbol{L}}_A\boldsymbol{\theta}_A^*\right\|_2.
\end{aligned} \tag{13}$$

Now we need to bound the term $\left\|\widehat{\boldsymbol{T}}_A - \widehat{\boldsymbol{L}}_A\boldsymbol{\theta}_A^*\right\|_2$. Since using the discount parameter $\eta \in (0,1)$ is equivalent to multiplying both the imputed contexts and the imputed rewards by the parameter $\sqrt{\eta}$ in each episode, we have, in the n-th episode,

$$\left\|\widehat{\boldsymbol{T}}_A^{n-1} - \widehat{\boldsymbol{L}}_A^{n-1}\boldsymbol{\theta}_A^*\right\|_2 = \left\|\boldsymbol{\Delta}_{n-1}^{\eta}\right\|_2, \tag{14}$$

where $\boldsymbol{\Delta}_{n-1}^{\eta} = \{\eta^{(n-i-1)/2}\,\mathrm{IR}_{i,b}\}_{i\in[n-1],b\in[B]}$ denotes an exponential-decay vector of the instantaneous regrets, and $\mathrm{IR}_{i,b}$ denotes the instantaneous regret at step b in the i-th episode, i.e, $\mathrm{IR}_{i,b} = \left|\left\langle\bar{\boldsymbol{\theta}}_A^i, \boldsymbol{s}_{i,b}\right\rangle - \left\langle\boldsymbol{\theta}_A^*, \boldsymbol{s}_{i,b}\right\rangle\right|$. From (14), letting

$$\mathrm{CIR}_i = \sum_{b\in[B]}\mathrm{IR}_{i,b} \tag{15}$$

be the cumulative instantaneous regret in the i-th episode, we can obtain the upper bound of (14) as follows:

$$\begin{aligned}\left\|\widehat{\boldsymbol{T}}_A^{n-1}-\widehat{\boldsymbol{L}}_A^{n-1}\boldsymbol{\theta}_A^*\right\|_2 &= \left\|\boldsymbol{\Delta}_{n-1}^{\eta}\right\|_2 \\ &\leq \left\|\boldsymbol{\Delta}_{n-1}^{\eta}\right\|_1 \\ &= \sum_{i\in[n-1],b\in[B]}\left|\eta^{(n-i-1)/2}\,\mathrm{IR}_{i,b}\right| \\ &= \sum_{i\in[n-1]}\eta^{(n-i-1)/2}\,\mathrm{CIR}_i \\ &= \eta^{-\frac{1}{2}}f_{\mathrm{Imp}}(n),\end{aligned} \tag{16}$$

where

$$f_{\mathrm{Imp}}(n) := \sum_{i\in[n-1]}\left(\sqrt{\eta}\right)^{n-i}\ \mathrm{CIR}_i. \tag{17}$$

From monotone bounded theorem, we have that the limit of CIR_i exists. From (7) in Lemma B.1, we get that $f_{\mathrm{Imp}}(n)$ is convergent and then has an upper bound. We denotes the upper bound of $f_{\mathrm{Imp}}(n)$ by $C_{\mathrm{Imp}}>0$, and then from (16) we have

$$\left\|\widehat{\boldsymbol{T}}_A^{n-1}-\widehat{\boldsymbol{L}}_A^{n-1}\boldsymbol{\theta}_A^*\right\|_2 \leq \eta^{-\frac{1}{2}}C_{\mathrm{Imp}}. \tag{18}$$

Substituting (18) into (13) yields the upper bound of $X_A^{(3)}$.

Then, we prove that

$$\sqrt{\boldsymbol{s}^{\mathsf{T}}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\boldsymbol{s}} \leq \sqrt{\boldsymbol{s}^{\mathsf{T}}\left(\lambda\boldsymbol{I}_d+\boldsymbol{\Phi}_A^n\right)^{-1}\boldsymbol{s}}. \tag{19}$$

holds, which is equivalent to

$$\boldsymbol{s}^{\mathsf{T}}\left(\lambda\boldsymbol{I}_d+\boldsymbol{\Phi}+\gamma\widehat{\boldsymbol{\Phi}}\right)^{-1}\boldsymbol{s} \leq \boldsymbol{s}^{\mathsf{T}}\left(\lambda\boldsymbol{I}_d+\boldsymbol{\Phi}\right)^{-1}\boldsymbol{s}. \tag{20}$$

Letting $\boldsymbol{\Theta}=\lambda\boldsymbol{I}_d+\boldsymbol{\Phi}$, by Sherman-Morrison-Woodbury formula, we have

$$\begin{aligned}\left(\boldsymbol{\Theta}+\gamma\widehat{\boldsymbol{\Phi}}\right)^{-1}=\left(\boldsymbol{\Theta}+\gamma\widehat{\boldsymbol{S}}^{\mathsf{T}}\widehat{\boldsymbol{S}}\right)^{-1} &= \boldsymbol{\Theta}^{-1}-\gamma\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}\left(\boldsymbol{I}_d+\gamma\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}\right)^{-1}\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1} \\ &= \boldsymbol{\Theta}^{-1}-\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}\left(\frac{\boldsymbol{I}_d}{\gamma}+\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}\right)^{-1}\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1},\end{aligned} \tag{21}$$

yielding that (20) is equivalent to

$$\boldsymbol{s}^{\mathsf{T}}\boldsymbol{\Gamma}\boldsymbol{s}\geq 0, \tag{22}$$

where

$$\boldsymbol{\Gamma}=\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}\left(\boldsymbol{I}_d/\gamma+\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}\right)^{-1}\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}.$$

Let $\boldsymbol{S}=\boldsymbol{U}_d\boldsymbol{\Sigma}_d^{1/2}\boldsymbol{V}_d^{\mathsf{T}}$, $\widehat{\boldsymbol{S}}=\widehat{\boldsymbol{U}}_d\widehat{\boldsymbol{\Sigma}}_d^{1/2}\widehat{\boldsymbol{V}}_d^{\mathsf{T}}$ be the Singular Value Decomposition (SVD) of $\boldsymbol{S}$ and $\widehat{\boldsymbol{S}}$, respectively. Note that $\boldsymbol{\Phi}=\boldsymbol{V}_d\boldsymbol{\Sigma}_d\boldsymbol{V}_d^{\mathsf{T}}$, $\widehat{\boldsymbol{\Phi}}=\widehat{\boldsymbol{V}}_d\widehat{\boldsymbol{\Sigma}}_d\widehat{\boldsymbol{V}}_d^{\mathsf{T}}$. We can obtain that $\boldsymbol{\Gamma}$ is a square symmetric positive semi-definite matrix, since $\boldsymbol{\Gamma}$ can be decomposed into

$$\boldsymbol{\Gamma}=\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q},$$

where $\boldsymbol{P}_\gamma\boldsymbol{\Lambda}_\gamma\boldsymbol{P}_\gamma^{\mathsf{T}}$ is the SVD of $\boldsymbol{I}_d/\gamma+\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}$ and

$$\boldsymbol{Q}=\boldsymbol{\Lambda}_\gamma^{-1/2}\boldsymbol{P}_\gamma^{\mathsf{T}}\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}.$$

Thus, (22) holds, yielding that (20) also holds.

Finally, we prove that a larger imputation rate γ leads to a smaller variance term $\sqrt{\boldsymbol{s}^{\mathsf{T}}\left(\boldsymbol{\Psi}\right)^{-1}\boldsymbol{s}}$. From (21), the variance term can be represented as follows:

$$\sqrt{\boldsymbol{s}^{\mathsf{T}}\left(\boldsymbol{\Psi}\right)^{-1}\boldsymbol{s}}=\left[\boldsymbol{s}^{\mathsf{T}}\boldsymbol{\Theta}^{-1}\boldsymbol{s}-\boldsymbol{s}^{\mathsf{T}}\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}\boldsymbol{M}_\gamma^{-1}\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\boldsymbol{s}\right]^{1/2}, \tag{23}$$

where $\boldsymbol{M}_\gamma = \boldsymbol{I}_d/\gamma + \widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}$. Letting $\boldsymbol{M}_\gamma = \boldsymbol{U}_{\boldsymbol{M}_\gamma}\boldsymbol{\Lambda}_{\boldsymbol{M}_\gamma}\boldsymbol{U}_{\boldsymbol{M}_\gamma}^{\mathsf{T}}$ be the SVD of $\boldsymbol{M}_\gamma$, and $\boldsymbol{z} = \boldsymbol{U}_{\boldsymbol{M}_\gamma}^{\mathsf{T}}\widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\boldsymbol{s}$, from (23) we can written the variance term as follows:

$$\sqrt{\boldsymbol{s}^{\mathsf{T}}\left(\boldsymbol{\Psi}\right)^{-1}\boldsymbol{s}} = \left[\boldsymbol{s}^{\mathsf{T}}\boldsymbol{\Theta}^{-1}\boldsymbol{s} - \boldsymbol{z}^{\mathsf{T}}\boldsymbol{\Lambda}_{\boldsymbol{M}_\gamma}^{-1}\boldsymbol{z}\right]^{1/2}. \tag{24}$$

In (24), we can observed that

$$\boldsymbol{z}^{\mathsf{T}}\boldsymbol{\Lambda}_{\boldsymbol{M}_\gamma}\boldsymbol{z} = \|(\boldsymbol{\Lambda}_{\boldsymbol{M}_\gamma})^{-1/2}\boldsymbol{z}\|_2^2 \in \left[\frac{1}{\sigma_{\max}(\boldsymbol{M}) + 1/\gamma}\|\boldsymbol{z}\|_2^2,\ \frac{1}{\sigma_{\min}(\boldsymbol{M}) + 1/\gamma}\|\boldsymbol{z}\|_2^2\right],$$

where $\boldsymbol{M} = \widehat{\boldsymbol{S}}\boldsymbol{\Theta}^{-1}\widehat{\boldsymbol{S}}^{\mathsf{T}}$, which indicates that a larger imputation rate γ leads to a smaller variance term. □

Finally, we provide a deeper understanding of the additional bias in Theorem 2 in the manuscript.

Remark B.1 (Controllable Bias). *Our reward imputation approach incurs a bias term* $\gamma^{\frac{1}{2}}\eta^{-\frac{1}{2}}C_{\mathrm{Imp}}$ *in addition to the two bias terms* $\lambda\|\boldsymbol{\theta}_A^*\|_2$ *and* ν *that exist in every UCB-based policy. But this additional bias term is controllable due to the presence of imputation rate* γ *that can help controlling the additional bias. Moreover, from the proof of* (16)*, we can obtain that, the term* C_{Imp} *in the additional bias can be replaced by a function* $f_{\mathrm{Imp}}(n)$ *(defined in* (17)*), and the additional bias term turns out to be* $\gamma^{\frac{1}{2}}\eta^{-\frac{1}{2}}f_{\mathrm{Imp}}(n)$*. Since* $f_{\mathrm{Imp}}(n)$ *has the same functional form as the function* $f(n)$ *in Lemma B.1, we can find the conditions that* $f_{\mathrm{Imp}}(n)$ *is monotonic decreasing following* (8) *in Lemma B.1. Specifically, letting* CIR_i *be the cumulative instantaneous regret in the* i*-th episode defined in* (15)*,*

1) *when* $\sqrt{\eta} \neq 1/2$*, the condition of a monotonic decreasing function* $f_{\mathrm{Imp}}(\cdot)$ *is equivalent to, for any positive integer* $i \geq 2$*,*

$$\mathrm{CIR}_i \leq \frac{(1-\sqrt{\eta})\left[\sqrt{\eta}^{i-1} - (1-\sqrt{\eta})^{i-1}\right]}{2\sqrt{\eta}-1}\mathrm{CIR}_1,$$

indicating that the regret after N *episodes satisfies*

$$\begin{aligned}\sum_{2\leq i\leq N}\mathrm{CIR}_i &\leq \mathrm{CIR}_1\sum_{2\leq i\leq N}\frac{(1-\sqrt{\eta})\left[\sqrt{\eta}^{i-1} - (1-\sqrt{\eta})^{i-1}\right]}{2\sqrt{\eta}-1}\\ &= \mathrm{CIR}_1\frac{1-\sqrt{\eta}}{2\sqrt{\eta}-1}\sum_{2\leq i\leq N}\left[\sqrt{\eta}^{i-1} - (1-\sqrt{\eta})^{i-1}\right]\\ &= \mathrm{CIR}_1\frac{1}{\sqrt{\eta}(2\sqrt{\eta}-1)}\left[2\sqrt{\eta}-1+(1-\sqrt{\eta})^{N+1} - (\sqrt{\eta})^{N+1}\right]\\ &= \mathrm{CIR}_1\frac{1}{\sqrt{\eta}}\left[1+\frac{(1-\sqrt{\eta})^{N+1} - (\sqrt{\eta})^{N+1}}{2\sqrt{\eta}-1}\right].\end{aligned} \tag{25}$$

2) *for the case* $\sqrt{\eta} = 1/2$*, the condition of a monotonic decreasing function* $f_{\mathrm{Imp}}(\cdot)$ *is equivalent to* $\mathrm{CIR}_i \leq (i-1)(\sqrt{\eta})^{i-1}\mathrm{CIR}_1$ *for any positive integer* $i \geq 2$*, indicating that the regret after* N *episodes satisfies*

$$\begin{aligned}\sum_{2\leq i\leq N}\mathrm{CIR}_i &\leq \mathrm{CIR}_1\sum_{2\leq i\leq N}(i-1)(\sqrt{\eta})^{i-1}\\ &= \frac{\sqrt{\eta}}{(1-\sqrt{\eta})^2}\mathrm{CIR}_1 - \left[\frac{1}{(1-\sqrt{\eta})^2} + \frac{N-1}{1-\sqrt{\eta}}\right](\sqrt{\eta})^N\mathrm{CIR}_1\\ &= \left(2 - \frac{1+N}{2^{N-1}}\right)\mathrm{CIR}_1\\ &= \left[\frac{1}{\sqrt{\eta}} - (1+N)\sqrt{\eta}^{N-1}\right]\mathrm{CIR}_1.\end{aligned} \tag{26}$$

From (25) *and* (26), *we can conclude that a monotonic decreasing function* $f_{\mathrm{Imp}}(\cdot)$ *indicates the upper bound of regret after* N *episodes is of order* $O(\mathrm{CIR}_1/\sqrt{\eta})$. *The conclusion also indicates that setting the discount parameter as* $\sqrt{\eta} = \Theta(\mathrm{CIR}_1/N)$ *achieves a* $O(N)$ *regret bound (i.e., a* $\tilde{O}(\sqrt{dT})$ *regret bound following Remark 3). Note that setting the discount parameter as* $\sqrt{\eta} = \Theta(\mathrm{CIR}_1/N)$ *is a mild condition, since the cumulative instantaneous regret* CIR_1 *is typically of order* $O(B)$ *(*$B = O(\sqrt{T/d})$ *in Remark 3) yielding that* $\sqrt{\eta} = \Theta(d^{-1})$. *Overall, since a larger imputation rate* γ *leads to a smaller variance while increasing the bias (variance analysis can be found in Remark 2),* γ *controls a trade-off between the bias term and the variance term. When* f_{Imp} *is a monotonic decreasing function w.r.t. number of episodes* n*, the additional bias term* $\gamma^{\frac{1}{2}}\eta^{-\frac{1}{2}}f_{\mathrm{Imp}}(n)$ *can be easily controlled, e.g., gradually increasing* γ *with the number of episodes, avoiding the large bias from* $f_{\mathrm{Imp}}(n)$ *at the beginning of reward imputation. We design a rate-scheduled approach for choosing the imputation rate* γ *in Section 5.*

Remark B.2 (Relationship to Exploration and Exploitation Trade-off)**.** *Exploration-exploitation dilemma is the key challenge in online learning under bandit settings. In the full-information setting, agent (e.g., UCB policy) can observe the rewards from all the actions, and thus does not need to consider the problem of exploring the feedback mechanisms, and achieves a lower variance part in the regret upper bound (Theorem A.1). Along this line, our reward computation approach is proposed to approximate the setting of full-information feedback, which somewhat relaxes the explore/exploit dilemma and also brings a lower variance part and a controllable additional bias part in the regret. Extra information that pushes the policy towards exploitation and away from exploration comes from the estimated reward structures of the non-executed actions maintained in each episode, and the proposed reward imputation can be seen as an effective and efficient tool to capture this extra information.*

B.2 Approximation Properties of Sketching

Although some error bounds of approximation using SJLT have been proposed (Nelson and Nguyên, 2013; Kane and Nelson, 2014; Bourgain et al., 2015), it is still unknown what is the lower bound of the sketch size while applying SJLT to the sketched ridge regression problem in our SPUIR. To address this issue, we first prove two approximation properties of SJLT which are necessary to achieve approximation error bound of the sketched ridge regression using SJLT. For convenience, we drop all the superscripts and subscripts in these theoretical results.

Lemma B.2 ((Nelson and Nguyên, 2013))**.** *Let* $\boldsymbol{U} \in \mathbb{R}^{L\times d}$ *be a matrix with orthonormal columns,* $\boldsymbol{\Pi} \in \mathbb{R}^{c\times L}$ *the SJLT. Assuming that* $D = \Theta(\varepsilon_\sigma^{-1}\log^3(d\delta_0^{-1}))$ *for* $\boldsymbol{\Pi}$, $\varepsilon_\sigma \in (0,1)$ *and* $d \le c$, *with probability at least* $1-\delta_0$ *all singular values of* $\boldsymbol{\Pi U}$

$$\sigma_i(\boldsymbol{\Pi U}) = 1 \pm \varepsilon_\sigma, \quad i \in [d],$$

as long as

$$c \ge \frac{d\log^8\left(d\delta_0^{-1}\right)}{\varepsilon_\sigma^2}.$$

Further, this holds if the hash function h *and* σ *defining the* $\boldsymbol{\Pi}$ *is* $\Omega\left(\log(d\delta_0^{-1})\right)$*-wise independent.*

Theorem B.1 (Approximation Properties of SJLT)**.** *Let* $\boldsymbol{U} \in \mathbb{R}^{L\times d}$ *be a matrix with orthonormal columns, and* $\boldsymbol{A}$ *be a matrix of any proper size. If* $\boldsymbol{\Pi} \in \mathbb{R}^{c\times L}$ *is the SJLT satisfying the assumptions in Lemma B.2, and* $d \le c \le L$, *then* $\boldsymbol{\Pi}$ *has the following two properties:*

1) *Subspace embedding property: set* $c = \Omega\left(d\ \mathrm{polylog}\left(d\delta_{\mathrm{s}}^{-1}\right)/\varepsilon_{\mathrm{s}}^2\right)$, *for* $\varepsilon_{\mathrm{s}} \in (0,1)$, *with probability at least* $1-\delta_{\mathrm{s}}$,
$$\|\boldsymbol{U}^\intercal\boldsymbol{\Pi}^\intercal\boldsymbol{\Pi U} - \boldsymbol{I}_d\|_2 \le \varepsilon_{\mathrm{s}};$$

2) *Matrix multiplication property: set* $c = \Omega(d/(\varepsilon_{\mathrm{m}}\delta_{\mathrm{m}}))$, *for* $\varepsilon_{\mathrm{m}} \in (0,1)$, *with probability at least* $1-\delta_{\mathrm{m}}$,
$$\|\boldsymbol{U}^\intercal\boldsymbol{\Pi}^\intercal\boldsymbol{\Pi A} - \boldsymbol{U}^\intercal\boldsymbol{A}\|_{\mathrm{F}}^2 \le \varepsilon_{\mathrm{m}}\|\boldsymbol{A}\|_{\mathrm{F}}^2.$$

Proof of Theorem B.1. 1) From Lemma B.2, by setting $c = \Omega\left(d\ \mathrm{polylog}\left(d\delta_{\mathrm{s}}^{-1}\right)/\varepsilon_0^2\right)$, we can obtain the upper bounds of eigenvalues: with probability at least $1-\delta_{\mathrm{s}}$,

$$\lambda_i\left(\boldsymbol{U}^\intercal\boldsymbol{\Pi}^\intercal\boldsymbol{\Pi U}\right) = \sigma_i^2(\boldsymbol{\Pi U}) \in [(1-\varepsilon_\sigma)^2, (1+\varepsilon_\sigma)^2] \subseteq [1-2\varepsilon_\sigma, 1+3\varepsilon_\sigma], \tag{27}$$

which yields that

$$|\lambda_i\left(\boldsymbol{U}^\intercal\boldsymbol{\Pi}^\intercal\boldsymbol{\Pi}\boldsymbol{U}-\boldsymbol{I}_d\right)|\leq 3\varepsilon_\sigma. \tag{28}$$

(29) is equivalent to

$$\left\|\boldsymbol{U}^\intercal\boldsymbol{\Pi}^\intercal\boldsymbol{\Pi}\boldsymbol{U}-\boldsymbol{I}_d\right\|_2\leq 3\varepsilon_\sigma.$$

Letting $\varepsilon_{\mathrm{s}}=3\varepsilon_\sigma$ and $\varepsilon_\sigma\in(0,1/3)$ yields the subspace embedding property.

2) From Lemma 1 in (Zhang and Liao, 2019), we have

$$\mathbb{E}\left[\|\boldsymbol{U}^\intercal\boldsymbol{\Pi}^\intercal\boldsymbol{\Pi}\boldsymbol{A}-\boldsymbol{U}^\intercal\boldsymbol{A}\|_{\mathrm{F}}^2\right]\leq\frac{2}{c}\|\boldsymbol{U}\|_{\mathrm{F}}^2\|\boldsymbol{A}\|_{\mathrm{F}}^2=\frac{2d}{c}\|\boldsymbol{A}\|_{\mathrm{F}}^2. \tag{29}$$

Combining (29) with the Markov's inequality, we obtain that, with probability at least $1-\delta_{\mathrm{m}}$,

$$\|\boldsymbol{U}^\intercal\boldsymbol{\Pi}^\intercal\boldsymbol{\Pi}\boldsymbol{A}-\boldsymbol{U}^\intercal\boldsymbol{A}\|_{\mathrm{F}}^2\leq\frac{2d}{\delta_{\mathrm{m}}c}\|\boldsymbol{A}\|_{\mathrm{F}}^2.$$

Letting $\varepsilon_{\mathrm{m}}=\dfrac{2d}{\delta_{\mathrm{m}}c}$ yields the matrix multiplication property.

□

B.3 Proof of Theorem 3

Next, using the approximation properties of SJLT in Theorem B.1, we prove that the objective function value of the imputation regularized ridge regression problem for reward imputation can be approximated well with a relative-error bound. Moreover, we prove that the solution solving the sketched ridge regression problem for reward imputation is also a good approximation of the solution solving the imputation regularized ridge regression. The following theorem is a detailed version of Theorem 3.

Theorem B.2 (Approximation Error Bound of Imputation using Sketching, Detailed Version of Theorem 3)**.** *Let* $\gamma\in[0,1]$ *be the imputation rate,* $\lambda>0$ *the regularization parameter,* $\boldsymbol{\Pi}\in\mathbb{R}^{c\times L}$ *and* $\widehat{\boldsymbol{\Pi}}\in\mathbb{R}^{c\times\widehat{L}}$ *be the SJLT, and* $\boldsymbol{L}\in\mathbb{R}^{L\times d},\widehat{\boldsymbol{L}}\in\mathbb{R}^{\widehat{L}\times d},\boldsymbol{T}\in\mathbb{R}^{L},\widehat{\boldsymbol{T}}\in\mathbb{R}^{\widehat{L}},\boldsymbol{\theta}\in\mathbb{R}^d$*. Denote the imputation regularized ridge regression function* F *and sketched ridge regression function* F^{S} *for reward imputation by*

$$F(\boldsymbol{\theta})=\|\boldsymbol{L}\boldsymbol{\theta}-\boldsymbol{T}\|_2^2+\gamma\left\|\widehat{\boldsymbol{L}}\boldsymbol{\theta}-\widehat{\boldsymbol{T}}\right\|_2^2+\lambda\|\boldsymbol{\theta}\|_2^2,$$

$$F^{\mathrm{S}}(\boldsymbol{\theta})=\|\boldsymbol{\Pi}\left(\boldsymbol{L}\boldsymbol{\theta}-\boldsymbol{T}\right)\|_2^2+\gamma\left\|\widehat{\boldsymbol{\Pi}}\left(\widehat{\boldsymbol{L}}\boldsymbol{\theta}-\widehat{\boldsymbol{T}}\right)\right\|_2^2+\lambda\|\boldsymbol{\theta}\|_2^2,$$

and the solutions of these regression problems by

$$\bar{\boldsymbol{\theta}}=\operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d}F(\boldsymbol{\theta})\quad\textit{and}\quad\tilde{\boldsymbol{\theta}}=\operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d}F^{\mathrm{S}}(\boldsymbol{\theta}).$$

Let $\delta\in(0,0.1]$*,* $\varepsilon\in(0,1)$*,* $\rho_\lambda=\|\boldsymbol{L}_{\mathrm{all}}\|_2^2/(\|\boldsymbol{L}_{\mathrm{all}}\|_2^2+\lambda)$*. For* $\boldsymbol{\Pi}$ *and* $\widehat{\boldsymbol{\Pi}}$*, assuming that* $D=\Theta(\varepsilon^{-1}\log^3(d\delta^{-1}))$ *and*

$$c=\Omega\left(d\ \mathrm{polylog}\left(d\delta^{-1}\right)/\varepsilon^2\right),$$

with probability at least $1-\delta$*,*

$$F(\tilde{\boldsymbol{\theta}})\leq(1+\rho_\lambda\varepsilon)F(\bar{\boldsymbol{\theta}}), \tag{30}$$

$$\|\tilde{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}\|_2\leq\frac{\sqrt{\rho_\lambda\varepsilon F\left(\bar{\boldsymbol{\theta}}\right)}}{\sigma_{\min}\left(\boldsymbol{L}_{\mathrm{all}}^\lambda\right)}, \tag{31}$$

where $\boldsymbol{L}_{\mathrm{all}}^\lambda=\left[\boldsymbol{L};\sqrt{\gamma}\widehat{\boldsymbol{L}};\sqrt{\lambda}\boldsymbol{I}_d\right]\in\mathbb{R}^{(L+\widehat{L}+d)\times d}$*. Furthermore, if there is a constant fraction of the norm of* $\boldsymbol{T}_{\mathrm{all}}^0$ *lies in the column space of* L_{all}^λ*, then* (31) *can be strengthened. Formally, assuming that a mild structural assumption on the context matrix and the reward vector is satisfied, i.e.,* $\|\boldsymbol{U}_{\mathrm{all}}\boldsymbol{U}_{\mathrm{all}}^\intercal\boldsymbol{T}_{\mathrm{all}}^0\|_2\geq\xi\|\boldsymbol{T}_{\mathrm{all}}^0\|_2$ *with a constant* $\xi\in(0,1]$*, then with probability at least* $1-\delta$*,*

$$\|\tilde{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}\|_2\leq\left(\kappa(\boldsymbol{L}_{\mathrm{all}}^\lambda)\sqrt{\xi^{-2}-1}\right)\sqrt{\rho_\lambda\epsilon}\|\bar{\boldsymbol{\theta}}\|_2, \tag{32}$$

where $\kappa(\boldsymbol{A})$ *denotes the condition number of* $\boldsymbol{A}$*,* $\boldsymbol{T}_{\mathrm{all}}^0=[\boldsymbol{T};\widehat{\boldsymbol{T}};\boldsymbol{0}_d]\in\mathbb{R}^{(L+\widehat{L}+d)}$*, and* $\boldsymbol{L}_{\mathrm{all}}^\lambda=\boldsymbol{U}_{\mathrm{all}}\boldsymbol{\Sigma}_{\mathrm{all}}\boldsymbol{V}_{\mathrm{all}}^\intercal$ *is the SVD of* $\boldsymbol{L}_{\mathrm{all}}^\lambda$*.*

Proof of Theorem B.2. We first introduce some more notation of block matrices that will simplify the proof of the theorem:

$$\boldsymbol{\Pi}_{\text{all}} = \begin{pmatrix} \boldsymbol{\Pi} & \boldsymbol{O} \\ \boldsymbol{O} & \widehat{\boldsymbol{\Pi}} \end{pmatrix}, \quad \boldsymbol{L}_{\text{all}} = \begin{pmatrix} \boldsymbol{L} \\ \sqrt{\gamma}\widehat{\boldsymbol{L}} \end{pmatrix}, \quad \boldsymbol{T}_{\text{all}} = \begin{pmatrix} \boldsymbol{T} \\ \widehat{\boldsymbol{T}} \end{pmatrix}. \tag{33}$$

Then the regression functions can be rewritten as follows:

$$F(\boldsymbol{\theta}) = \|\boldsymbol{L}_{\text{all}}\boldsymbol{\theta} - \boldsymbol{T}_{\text{all}}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2, \qquad F^{\text{S}}(\boldsymbol{\theta}) = \|\boldsymbol{\Pi}_{\text{all}}\left(\boldsymbol{L}_{\text{all}}\boldsymbol{\theta} - \boldsymbol{T}_{\text{all}}\right)\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2.$$

Obviously, $\boldsymbol{\Pi}_{\text{all}}$ is still an SJLT. Combining Theorem B.1 with theorem 19 in (Wang et al., 2017), we can obtain, setting

$$c = \Omega\left(\max\{d \text{ polylog}\left(d\delta_{\text{s}}^{-1}\right)/\varepsilon_{\text{s}}^2,\ d/(\varepsilon_{\text{m}}\delta_{\text{m}})\}\right),$$

with probability at least $1 - (\delta_{\text{s}} + \delta_{\text{m}})$,

$$F(\tilde{\boldsymbol{\theta}}) - F(\bar{\boldsymbol{\theta}}) \le \rho_\lambda \tau F(\bar{\boldsymbol{\theta}}), \tag{34}$$

where $\rho_\lambda = \frac{\|\boldsymbol{L}_{\text{all}}\|_2^2}{\|\boldsymbol{L}_{\text{all}}\|_2^2 + \lambda}$ and $\tau = \frac{2\max\{\varepsilon_{\text{s}}^2, \varepsilon_{\text{m}}\}}{1-\varepsilon_{\text{s}}}$. Letting $\varepsilon_{\text{s}} = \varepsilon_{\text{m}} := \varepsilon_0$, (34) can be rewritten as

$$F(\tilde{\boldsymbol{\theta}}) - F(\bar{\boldsymbol{\theta}}) \le \frac{2\rho_\lambda\varepsilon_0}{1-\varepsilon_0} F(\bar{\boldsymbol{\theta}}), \tag{35}$$

Assuming that $\delta_{\text{s}} = \delta_{\text{m}} := \delta/2 \in (0, 0.1]$ and $\varepsilon_0 \in (0, 1/3)$, setting $\epsilon = \frac{2\varepsilon_0}{1-\varepsilon_0} \in (0,1)$, from (35) we obtain the upper bound (30).

Next, we bound the difference between the solutions solving the sketched ridge regression problem and the original regression problem. Since $\sigma_{\min}^2(\boldsymbol{A})\|\boldsymbol{x}\|_2^2 \le \|\boldsymbol{A}\boldsymbol{x}\|_2^2$ for any $\boldsymbol{A}$ and $\boldsymbol{x}$ with proper sizes, we have

$$\sigma_{\min}^2(\boldsymbol{L}_{\text{all}})\|\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}\|_2^2 \le \left\|\boldsymbol{L}_{\text{all}}(\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\right\|_2^2. \tag{36}$$

The key ingredient of bounding $\|\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}\|_2$ is to bound $\|\boldsymbol{L}_{\text{all}}(\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\|_2$. Let $\boldsymbol{L}_{\text{all}}^\lambda = \left[\boldsymbol{L}; \sqrt{\gamma}\widehat{\boldsymbol{L}}; \sqrt{\lambda}\boldsymbol{I}_d\right] \in \mathbb{R}^{(L+\widehat{L}+d)\times d}$, $\boldsymbol{T}_{\text{all}}^0 = [\boldsymbol{T}; \widehat{\boldsymbol{T}}; \boldsymbol{0}_d] \in \mathbb{R}^{(L+\widehat{L}+d)}$, $\boldsymbol{L}_{\text{all}}^\lambda = \boldsymbol{U}_{\text{all}}\boldsymbol{\Sigma}_{\text{all}}\boldsymbol{V}_{\text{all}}^\intercal$ be the SVD of $\boldsymbol{L}_{\text{all}}^\lambda$, and denote a matrix with orthonormal columns by $\boldsymbol{U}_{\text{all}}^\perp \in \mathbb{R}^{(L+\widehat{L}+d)\times(L+\widehat{L})}$ which satisfies

$$\boldsymbol{U}_{\text{all}}\boldsymbol{U}_{\text{all}}^\intercal + \boldsymbol{U}_{\text{all}}^\perp(\boldsymbol{U}_{\text{all}}^\perp)^\intercal = \boldsymbol{I}_{L+\widehat{L}+d} \quad \text{and} \quad \boldsymbol{U}_{\text{all}}^\intercal\boldsymbol{U}_{\text{all}}^\perp = \boldsymbol{O}.$$

Then, we can rewrite the solution $\bar{\boldsymbol{\theta}}$ as follows:

$$\begin{aligned} \bar{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} F(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\|\boldsymbol{L}_{\text{all}}^\lambda\boldsymbol{\theta} - \boldsymbol{T}_{\text{all}}^0\right\|_2^2 \\ &= (\boldsymbol{L}_{\text{all}}^\lambda)^\dagger\boldsymbol{T}_{\text{all}}^0 = \boldsymbol{V}_{\text{all}}\boldsymbol{\Sigma}_{\text{all}}^{-1}\boldsymbol{U}_{\text{all}}^\intercal\boldsymbol{T}_{\text{all}}^0, \end{aligned}$$

which yields that

$$\begin{aligned} \boldsymbol{T}_{\text{all}}^0 - \boldsymbol{L}_{\text{all}}^\lambda\bar{\boldsymbol{\theta}} &= \boldsymbol{T}_{\text{all}}^0 - \boldsymbol{L}_{\text{all}}^\lambda\boldsymbol{V}_{\text{all}}\boldsymbol{\Sigma}_{\text{all}}^{-1}\boldsymbol{U}_{\text{all}}^\intercal\boldsymbol{T}_{\text{all}}^0 \\ &= \boldsymbol{T}_{\text{all}}^0 - \boldsymbol{U}_{\text{all}}\boldsymbol{\Sigma}_{\text{all}}\boldsymbol{V}_{\text{all}}^\intercal\boldsymbol{V}_{\text{all}}\boldsymbol{\Sigma}_{\text{all}}^{-1}\boldsymbol{U}_{\text{all}}^\intercal\boldsymbol{T}_{\text{all}}^0 \\ &= \boldsymbol{T}_{\text{all}}^0 - \boldsymbol{U}_{\text{all}}\boldsymbol{U}_{\text{all}}^\intercal\boldsymbol{T}_{\text{all}}^0 \\ &= \boldsymbol{U}_{\text{all}}^\perp(\boldsymbol{U}_{\text{all}}^\perp)^\intercal\boldsymbol{T}_{\text{all}}^0. \end{aligned} \tag{37}$$

Thus, $\boldsymbol{T}_{\text{all}}^0 - \boldsymbol{L}_{\text{all}}^\lambda\bar{\boldsymbol{\theta}}$ is orthogonal to $\boldsymbol{U}_{\text{all}}$, and consequently to $\boldsymbol{L}_{\text{all}}^\lambda(\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})$, and we can obtain the following equality by Pythagoras's theorem:

$$\left\|\boldsymbol{L}_{\text{all}}^\lambda(\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\right\|_2^2 = \left\|\boldsymbol{L}_{\text{all}}^\lambda\tilde{\boldsymbol{\theta}} - \boldsymbol{T}_{\text{all}}^0\right\|_2^2 - \left\|\boldsymbol{L}_{\text{all}}^\lambda\bar{\boldsymbol{\theta}} - \boldsymbol{T}_{\text{all}}^0\right\|_2^2. \tag{38}$$

Combining (38) with (30) yields that

$$\left\|\boldsymbol{L}_{\text{all}}^\lambda(\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})\right\|_2^2 = F(\tilde{\boldsymbol{\theta}}) - F(\bar{\boldsymbol{\theta}}) \le \rho_\lambda\varepsilon F(\bar{\boldsymbol{\theta}}). \tag{39}$$

Substituting (39) into (36) concludes the proof of (31).

If we make a mild structural assumption on the context matrix and the reward vector, we can provide a stronger bound of $\|\tilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}\|_2$. Specifically, assuming that $\|\boldsymbol{U}_{\mathrm{all}}\boldsymbol{U}_{\mathrm{all}}^{\intercal}\boldsymbol{T}_{\mathrm{all}}^{0}\|_2 \geq \xi\|\boldsymbol{T}_{\mathrm{all}}^{0}\|_2$ with a constant $\xi \in (0,1]$, from (37) and Pythagoras's theorem we have

$$\begin{aligned} F(\bar{\boldsymbol{\theta}}) &= \|\boldsymbol{L}_{\mathrm{all}}^{\lambda}\bar{\boldsymbol{\theta}} - \boldsymbol{T}_{\mathrm{all}}^{0}\|_2^2 \\ &= \|\boldsymbol{T}_{\mathrm{all}}^{0}\|_2^2 - \|\boldsymbol{U}_{\mathrm{all}}\boldsymbol{U}_{\mathrm{all}}^{\intercal}\boldsymbol{T}_{\mathrm{all}}^{0}\|_2^2 \\ &\leq (\xi^{-2} - 1)\|\boldsymbol{U}_{\mathrm{all}}\boldsymbol{U}_{\mathrm{all}}^{\intercal}\boldsymbol{T}_{\mathrm{all}}^{0}\|_2^2 \\ &= (\xi^{-2} - 1)\|\boldsymbol{L}_{\mathrm{all}}^{\lambda}\bar{\boldsymbol{\theta}}\|_2^2 \\ &\leq (\xi^{-2} - 1)\|\boldsymbol{L}_{\mathrm{all}}^{\lambda}\|_2^2\ \|\bar{\boldsymbol{\theta}}\|_2^2 \\ &\leq (\xi^{-2} - 1)\sigma_{\max}^2(\boldsymbol{L}_{\mathrm{all}}^{\lambda})\|\bar{\boldsymbol{\theta}}\|_2^2. \end{aligned} \tag{40}$$

Combining (40) with (31) yields (32). $\square$

B.4 Proof of Theorem 4

Proof of Theorem 4. In our sketched policy, letting $C_{\boldsymbol{\theta}^*}^{\max} = \max_{A\in\mathcal{A}}\|\boldsymbol{\theta}_A^*\|_2$, $C_{\mathrm{Imp}} > 0$, $\nu = \sqrt{2C_R^2\log(2MB/\delta)}$, and

$$\omega = \lambda C_{\boldsymbol{\theta}^*}^{\max} + \nu + \gamma^{\frac{1}{2}}\eta^{-\frac{1}{2}}C_{\mathrm{Imp}},$$

from Theorem 2 we obtain that

$$\left|\langle\bar{\boldsymbol{\theta}}_A^n, \boldsymbol{s}_{n,b}\rangle - \langle\boldsymbol{\theta}_A^*, \boldsymbol{s}_{n,b}\rangle\right| \leq \omega\sqrt{\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\boldsymbol{s}_{n,b}}. \tag{41}$$

Before proving the upper bound of $\left|\left\langle\tilde{\boldsymbol{\theta}}_A^n, \boldsymbol{s}_{n,b}\right\rangle - \langle\boldsymbol{\theta}_A^*, \boldsymbol{s}_{n,b}\rangle\right|$, we need to provide a technical tool as follows. For convenience, we also drop all the superscripts and subscripts. The goal is to find a constant C_α such that

$$\sqrt{\boldsymbol{s}^{\intercal}\boldsymbol{\Psi}^{-1}\boldsymbol{s}} \leq C_\alpha\sqrt{\boldsymbol{s}^{\intercal}\boldsymbol{W}^{-1}\boldsymbol{s}}, \tag{42}$$

which is equivalent to the condition that the matrix $C_\alpha^2\boldsymbol{W}^{-1} - \boldsymbol{\Psi}^{-1}$ is positive semidefinite. Let $\boldsymbol{L}_{\mathrm{all}}$ and $\boldsymbol{\Pi}_{\mathrm{all}}$ be the matrices defined in (33), $\boldsymbol{L}_{\mathrm{all}} = \widetilde{\boldsymbol{U}}_{\mathrm{all}}\widetilde{\boldsymbol{\Sigma}}_{\mathrm{all}}\widetilde{\boldsymbol{V}}_{\mathrm{all}}^{\intercal}$ be the SVD of $\boldsymbol{L}_{\mathrm{all}}$, and $\tilde{\sigma}_1 \geq \tilde{\sigma}_2\cdots \geq \tilde{\sigma}_d$ be the singular values of $\boldsymbol{L}_{\mathrm{all}}$. Then the i-th eigenvalue of $\boldsymbol{\Psi}^{-1} = (\lambda\boldsymbol{I}_d + \boldsymbol{L}_{\mathrm{all}}^{\intercal}\boldsymbol{L}_{\mathrm{all}})^{-1}$ can be represented as $\lambda_i(\boldsymbol{\Psi}^{-1}) = 1/(\tilde{\sigma}_i^2 + \lambda)$, and the i-th eigenvalue of $\boldsymbol{W}^{-1} = (\lambda\boldsymbol{I}_d + \boldsymbol{L}_{\mathrm{all}}^{\intercal}\boldsymbol{\Pi}_{\mathrm{all}}^{\intercal}\boldsymbol{\Pi}_{\mathrm{all}}\boldsymbol{L}_{\mathrm{all}})^{-1}$ is $\lambda_i(\boldsymbol{W}^{-1}) = 1/(\hat{\lambda}_i + \lambda)$, where $\hat{\lambda}_i$ is the i-th eigenvalue of $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{all}}\widetilde{\boldsymbol{U}}_{\mathrm{all}}^{\intercal}\boldsymbol{\Pi}_{\mathrm{all}}^{\intercal}\boldsymbol{\Pi}_{\mathrm{all}}\widetilde{\boldsymbol{U}}_{\mathrm{all}}\widetilde{\boldsymbol{\Sigma}}_{\mathrm{all}}$.

From the Lidskii's theorem and (27), we have

$$\hat{\lambda}_i \in [\tilde{\sigma}_d^2(1 - 2\varepsilon_\sigma), \tilde{\sigma}_1^2(1 + 3\varepsilon_\sigma)]. \tag{43}$$

Assuming that the positive semi-definiteness of $C_\alpha^2\boldsymbol{W}^{-1} - \boldsymbol{\Psi}^{-1}$ is satisfied, we obtain that $C_\alpha^2\lambda_i(\boldsymbol{W}^{-1}) - \lambda_i(\boldsymbol{\Psi}^{-1}) \geq 0$ for $i \in [d]$, and combining this inequality with (43) yields that

$$C_\alpha = \sqrt{[\tilde{\sigma}_1^2(1 + 3\varepsilon_\sigma) + \lambda]/(\tilde{\sigma}_d^2 + \lambda)}.$$

From the proof of Theorem B.1 and Theorem 3, we can obtain that $\varepsilon_\sigma = \varepsilon/(6 + 3\varepsilon)$, yielding that

$$C_\alpha = \sqrt{\frac{\tilde{\sigma}_1^2[1 + \varepsilon/(2 + \varepsilon)] + \lambda}{\tilde{\sigma}_d^2 + \lambda}},$$

which decreases with increase of $1/\varepsilon$. Similarly to the proof of C_α satisfying (42), we can obtain that

$$\sqrt{\boldsymbol{s}^{\intercal}\boldsymbol{W}^{-1}\boldsymbol{s}} \leq C_{\mathrm{reg}}\sqrt{\boldsymbol{s}^{\intercal}\boldsymbol{\Psi}^{-1}\boldsymbol{s}}, \tag{44}$$

provided that

$$C_{\mathrm{reg}} = \sqrt{\frac{\tilde{\sigma}_1^2 + \lambda}{\tilde{\sigma}_d^2[1 - 2\varepsilon/(6 + 3\varepsilon)] + \lambda}}.$$

Obviously, C_{reg} also decreases with increase of $1/\varepsilon$.

Then, letting $\alpha = \omega C_\alpha$, from (41) and (42) we have

$$\begin{aligned}
\left|\left\langle \tilde{\boldsymbol{\theta}}^n_A, \boldsymbol{s}_{n,b}\right\rangle - \left\langle \boldsymbol{\theta}^*_A, \boldsymbol{s}_{n,b}\right\rangle\right| &\le \left|\left\langle \tilde{\boldsymbol{\theta}}^n_A, \boldsymbol{s}_{n,b}\right\rangle - \left\langle \bar{\boldsymbol{\theta}}^n_A, \boldsymbol{s}_{n,b}\right\rangle\right| + \left|\left\langle \bar{\boldsymbol{\theta}}^n_A, \boldsymbol{s}_{n,b}\right\rangle - \left\langle \boldsymbol{\theta}^*_A, \boldsymbol{s}_{n,b}\right\rangle\right| \\
&\le \left|\left\langle \tilde{\boldsymbol{\theta}}^n_A - \bar{\boldsymbol{\theta}}^n_A, \boldsymbol{s}_{n,b}\right\rangle\right| + \omega\sqrt{\boldsymbol{s}^\intercal_{n,b}\left(\boldsymbol{\Psi}^n_A\right)^{-1}\boldsymbol{s}_{n,b}} \\
&\le Y_{n,b} + \alpha\sqrt{\boldsymbol{s}^\intercal_{n,b}\left(\boldsymbol{W}^n_A\right)^{-1}\boldsymbol{s}_{n,b}},
\end{aligned} \tag{45}$$

where $Y_{n,b}$ denotes the upper bound of $\left|\left\langle \tilde{\boldsymbol{\theta}}^n_A - \bar{\boldsymbol{\theta}}^n_A, \boldsymbol{s}_{n,b}\right\rangle\right|$ for any $A \in \mathcal{A}$.

Next, using the compatibility of norm, we give a specific representation of the sum of $Y_{n,b}$ as follows:

$$\sum_{b\in[B]} Y_{n,b} = \max_{A\in\mathcal{A}} \|\boldsymbol{S}^n_A(\tilde{\boldsymbol{\theta}}^n_A - \bar{\boldsymbol{\theta}}^n_A)\|_1 \le \max_{A\in\mathcal{A}} \|\boldsymbol{S}^n_A\|_1 \|\tilde{\boldsymbol{\theta}}^n_A - \bar{\boldsymbol{\theta}}^n_A\|_1 \le \max_{A\in\mathcal{A}} \|\boldsymbol{S}^n_A\|_1 \sqrt{d} \|\tilde{\boldsymbol{\theta}}^n_A - \bar{\boldsymbol{\theta}}^n_A\|_2. \tag{46}$$

Further, we give a more specific upper bound in (46) under mild structural assumption in Theorem 3. Let $\kappa^{\max}_{\text{all}}$ denote the maximum of the condition numbers of $\boldsymbol{L}^\lambda_{\text{all}}(A,n)$ for $A \in \mathcal{A}, n \in [N]$, and $\boldsymbol{L}^\lambda_{\text{all}}(A,n) = \left[\boldsymbol{L}^n_A; \sqrt{\gamma}\widehat{\boldsymbol{L}}^n_A; \sqrt{\lambda}\boldsymbol{I}_d\right]$, and $\boldsymbol{U}_{\text{all}}(A,n)$ be the left singular matrix of $\boldsymbol{L}^\lambda_{\text{all}}(A,n)$. Letting $\boldsymbol{T}^0_{\text{all}}(A,n) = [\boldsymbol{T}^n_A; \widehat{\boldsymbol{T}}^n_A; \boldsymbol{0}_d]$, assuming that $\|\boldsymbol{U}_{\text{all}}(A,n)\boldsymbol{U}_{\text{all}}(A,n)^\intercal \boldsymbol{T}^0_{\text{all}}(A,n)\|_2 \ge \xi\|\boldsymbol{T}^0_{\text{all}}(A,n)\|_2$ with a constant $\xi \in (0,1]$, substituting the upper bound (32) in Theorem 3 into (46) yields that

$$\sum_{b\in[B]} Y_{n,b} \le \left(\kappa^{\max}_{\text{all}}\sqrt{\xi^{-2}-1}\right) C_{\boldsymbol{S}} C^{\max}_{\boldsymbol{\theta}} \sqrt{\rho_\lambda \epsilon d}, \tag{47}$$

where $C_{\boldsymbol{S}} = \max_{n\in[N],A\in\mathcal{A}} \|\boldsymbol{S}^n_A\|_1$, $C^{\max}_{\boldsymbol{\theta}} = \max_{A\in\mathcal{A},n\in[N]} \|\bar{\boldsymbol{\theta}}^n_A\|_2$.

From (44), (45), (47) and the definition of our sketched policy, letting $C_Y = \left(\kappa^{\max}_{\text{all}}\sqrt{\xi^{-2}-1}\right) C_{\boldsymbol{S}} C^{\max}_{\boldsymbol{\theta}}$, we obtain that

$$\begin{aligned}
\text{Reg}(N,B) &= \sum_{n\in[N],b\in[B]} \left[\max_{A\in\mathcal{A}} \left\langle \boldsymbol{\theta}^*_A, \boldsymbol{s}_{n,b}\right\rangle - \left\langle \boldsymbol{\theta}^*_{A_{I_{n,b}}}, \boldsymbol{s}_{n,b}\right\rangle\right] \\
&\le \sum_{n\in[N],b\in[B]} \left[\max_{A\in\mathcal{A}} \left(\left\langle \tilde{\boldsymbol{\theta}}^n_A, \boldsymbol{s}_{n,b}\right\rangle + \alpha\sqrt{\boldsymbol{s}^\intercal_{n,b}\left(\boldsymbol{W}^n_A\right)^{-1}\boldsymbol{s}_{n,b}}\right) + Y_{n,b} - \left\langle \boldsymbol{\theta}^*_{A_{I_{n,b}}}, \boldsymbol{s}_{n,b}\right\rangle\right] \\
&= \sum_{n\in[N],b\in[B]} \left[\left\langle \tilde{\boldsymbol{\theta}}^n_{A_{I_{n,b}}}, \boldsymbol{s}_{n,b}\right\rangle + \alpha\sqrt{\boldsymbol{s}^\intercal_{n,b}\left(\boldsymbol{W}^n_{A_{I_{n,b}}}\right)^{-1}\boldsymbol{s}_{n,b}} + Y_{n,b} - \left\langle \boldsymbol{\theta}^*_{A_{I_{n,b}}}, \boldsymbol{s}_{n,b}\right\rangle\right] \\
&\le 2\alpha \sum_{n\in[N],b\in[B]} \sqrt{\boldsymbol{s}^\intercal_{n,b}\left(\boldsymbol{W}^n_{A_{I_{n,b}}}\right)^{-1}\boldsymbol{s}_{n,b}} + 2\sum_{n\in[N],b\in[B]} Y_{n,b} \\
&\le 2\alpha C_{\text{reg}}\sqrt{B} \sum_{n\in[N]} \sqrt{\sum_{b\in[B]} \boldsymbol{s}^\intercal_{n,b}\left(\boldsymbol{\Psi}^n_{A_{I_{n,b}}}\right)^{-1}\boldsymbol{s}_{n,b}} + 2NC_Y\sqrt{\rho_\lambda \epsilon d} \\
&= 2\alpha C_{\text{reg}}\sqrt{B} \sum_{n\in[N]} \sqrt{\sum_{b\in[B]} \left\langle \boldsymbol{s}_{n,b}\,\boldsymbol{s}^\intercal_{n,b}, \left(\boldsymbol{\Psi}^n_{A_{I_{n,b}}}\right)^{-1}\right\rangle} + 2NC_Y\sqrt{\rho_\lambda \epsilon d} \\
&= 2\alpha C_{\text{reg}}\sqrt{B} \sum_{n\in[N]} \sqrt{\sum_{A\in\mathcal{A}} \left\langle \boldsymbol{S}^{n\intercal}_A \boldsymbol{S}^n_A, \left(\boldsymbol{\Psi}^n_A\right)^{-1}\right\rangle} + 2NC_Y\sqrt{\rho_\lambda \epsilon d} \\
&= 2\alpha C_{\text{reg}}\sqrt{B} \sum_{n\in[N]} \sqrt{\sum_{A\in\mathcal{A}} \operatorname{tr}\left(\boldsymbol{S}^{n\intercal}_A \boldsymbol{S}^n_A \left(\boldsymbol{\Psi}^n_A\right)^{-1}\right)} + O(N\sqrt{\rho_\lambda \epsilon d}), \\
&\le 2\alpha C_{\text{reg}}\sqrt{BM} \sum_{n\in[N]} \sqrt{\max_{A\in\mathcal{A}}\left\{\operatorname{tr}\left(\boldsymbol{S}^{n\intercal}_A \boldsymbol{S}^n_A \left(\boldsymbol{\Psi}^n_A\right)^{-1}\right)\right\}} + O(N\sqrt{\rho_\lambda \epsilon d}).
\end{aligned} \tag{48}$$

When the structural assumption in Theorem 3 is not satisfied, from (31), we can obtain that the second term in (48) is also of order $O(\sqrt{\rho_{\lambda}\epsilon d})$, which does not influence the order of the final regret bound. Finally, combining (48) with lemma 3 in (Han et al., 2020) gives the final regret bound. ☐

Remark B.3 (Lower Bound of Regret). *Han et al. (2020) demonstrated a lower bound of regret for contextual batched bandit, which is of order $\Omega(\sqrt{dT})$. But this lower bound assumes that there are only two actions and both the actions share the same true reward model, it can not be directly applied to our CBB setting where each action corresponds to a different reward model. Despite the lack of the lower bound in CBB setting, if all M actions share the same true reward model, our regret upper bound of order $O(\sqrt{MdT})$ could reduce to a bound of order $O(\sqrt{dT})$ of magnitude, indicating the optimality of our regret upper bound. We leave the lower bound in CBB setting for further work.*

B.5 Regret Analyses of Our Extensions

In this section, we proof the regret bounds of the extensions of SPUIR in Section "Extensions of Our Approach".

Corollary B.1 (Regret Bounds of SPUIR-Exp, SPUIR-Poly, SPUIR-Kernel). *Assuming that the conditions in Theorem 4 holds and δ_1, δ_2 are the probabilities defined in Theorem 4, then:*

1) *With probability at least $1 - N(\delta_1 + \delta_2)$, SPUIR-Exp (for an exponential expected reward) and SPUIR-Poly (for a polynomial expected reward) enjoy the regret upper bound of the same order as that of SPUIR (shown in Theorem 4).*

2) *Comparing with the regret upper bound of SPUIR in Theorem 4, with probability at least $1-N(\delta_1+\delta_2+\delta_3)$, SPUIR-Kernel enjoys a regret bound with an additional error of order $O(\sqrt{N/\delta_3 d_{\mathrm{r}}}B)$ against the optimal policy in Gaussian RKHS, where d_{r} is the dimension of the random features and $\delta_3 \in (0,1)$. Setting $B = O(\sqrt{T/d})$ and $d_{\mathrm{r}} = O(N)$ yields an additional error of order $O(\sqrt{T/(d\delta_3)})$, and SPUIR-Kernel also enjoys the regret upper bound of the same order as that of SPUIR (shown in Theorem 4).*

Proof of Corollary B.1. 1) For an exponential expected reward, the regret of SPUIR-Exp can be written as follows:

$$\mathrm{Reg}_{\mathrm{E}}(N,B) = \sum_{n\in[N],b\in[B]} \left[\max_{A\in\mathcal{A}} G_{\mathrm{E}}\left(\boldsymbol{\theta}_A^*, \boldsymbol{s}_{n,b}\right) - G_{\mathrm{E}}\left(\boldsymbol{\theta}_{A_{I_{n,b}}}^*, \boldsymbol{s}_{n,b}\right)\right],$$

where $G_{\mathrm{E}}(\boldsymbol{\theta}, \boldsymbol{s}) = \exp\left(\boldsymbol{\theta}^{\intercal}\boldsymbol{s}\right)$. Using the linearization trick of convex functions (Shalev-Shwartz, 2011), the regret upper bound of SPUIR-Exp can be expressed using the inner products

$$\begin{aligned}&\mathrm{Reg}_{\mathrm{E}}(N,B)\\ \le &\sum_{n\in[N],b\in[B]} \left[\max_{A\in\mathcal{A}} \left\langle \boldsymbol{\theta}_A^*, \nabla_{\boldsymbol{\theta}_A^*} G_{\mathrm{E}}(\boldsymbol{\theta}_A^*, \boldsymbol{s}_{n,b})\right\rangle - \left\langle \boldsymbol{\theta}_{A_{I_{n,b}}}^*, \nabla_{\boldsymbol{\theta}_A^*} G_{\mathrm{E}}(\boldsymbol{\theta}_A^*, \boldsymbol{s}_{n,b})\right\rangle\right],\end{aligned}$$

where $\nabla_{\boldsymbol{\theta}} G_{\mathrm{E}}(\boldsymbol{\theta}, \boldsymbol{s}) = \exp\left(\boldsymbol{\theta}^{\intercal}\boldsymbol{s}\right)\boldsymbol{s}$. Then, the gradient $\nabla_{\boldsymbol{\theta}} G_{\mathrm{E}}(\boldsymbol{\theta}, \boldsymbol{s})$ can be treated as the example received at each step and applying Theorem 4 we obtain that SPUIR-Exp enjoys the regret bound of the same order. Similarly, for a polynomial expected reward, the regret upper bound of SPUIR-Poly can be expressed using the gradient $\nabla_{\boldsymbol{\theta}} G_{\mathrm{P}}(\boldsymbol{\theta}, \boldsymbol{s}) = 2\left(\boldsymbol{\theta}^{\intercal}\boldsymbol{s}\right)\boldsymbol{s}$, and applying Theorem 4 also yields the regret bound of the same order.

2) For SPUIR-Kernel, instead of the linear reward $\langle \boldsymbol{\theta}, \boldsymbol{s}_{n,b}\rangle$ in Euclidean space, we assume that the expected reward lies in a reproducing kernel Hilbert space (RKHS). More specifically, for action $A \in \mathcal{A}$, the expected reward can be formulated as $G_{\mathrm{K}}(\boldsymbol{\alpha}_A^*, \boldsymbol{s}) = \sum_{n\in[N],b\in[B]} \alpha_{n,b,A}^* \kappa(\boldsymbol{s}, \boldsymbol{s}_{n,b})$, where κ denotes a Gaussian kernel function with a kernel width σ_{R}. For fast implementation, we use $\mathcal{T}_{n,b}(\boldsymbol{\theta}, A) = \langle \boldsymbol{\theta}, \phi(\boldsymbol{s}_{n,b})\rangle$ in random feature space as an approximation of G_{K}, where the random feature mapping ϕ can be explicitly defined as in (Rahimi and Recht, 2007).

$$\phi(\boldsymbol{s}) = \frac{1}{\sqrt{d_{\mathrm{r}}}}\left[\cos(\boldsymbol{u}_1^{\intercal}\boldsymbol{s}), \cos(\boldsymbol{u}_2^{\intercal}\boldsymbol{s}), \ldots, \cos(\boldsymbol{u}_{d_{\mathrm{r}}}^{\intercal}\boldsymbol{s}), \sin(\boldsymbol{u}_1^{\intercal}\boldsymbol{s}), \sin(\boldsymbol{u}_2^{\intercal}\boldsymbol{s}), \ldots, \sin(\boldsymbol{u}_{d_{\mathrm{r}}}^{\intercal}\boldsymbol{s})\right]^{\intercal},$$

$\boldsymbol{u}_i \in \mathbb{R}^d, i \in [d_\mathrm{r}]$, are random parameter vectors sampled independently according to the Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma_\mathrm{R}^{-2}\boldsymbol{I}_d)$. Then, the regret of SPUIR-Kernel can be defined as follows:

$$\mathrm{Reg}_\mathrm{K}(N, B) = \sum_{n\in[N],b\in[B]} \left[\max_{A\in\mathcal{A}} G_\mathrm{K}\left(\boldsymbol{\alpha}_A^*, \boldsymbol{s}_{n,b}\right) - G_\mathrm{K}\left(\boldsymbol{\alpha}_{A_{I_{n,b}}}^*, \boldsymbol{s}_{n,b}\right)\right].$$

Letting

$$\begin{aligned}
\mathrm{Reg}_\mathrm{RF}(N, B) &= \sum_{n\in[N],b\in[B]} \left[\max_{A\in\mathcal{A}} \langle\boldsymbol{\theta}_A^*, \phi(\boldsymbol{s}_{n,b})\rangle - \left\langle\boldsymbol{\theta}_{A_{I_{n,b}}}^*, \phi(\boldsymbol{s}_{n,b})\right\rangle\right], \\
\mathrm{Diff}_1 &= \sum_{n\in[N],b\in[B]} \max_{A\in\mathcal{A}} \left|G_\mathrm{K}\left(\boldsymbol{\alpha}_A^*, \boldsymbol{s}_{n,b}\right) - \langle\boldsymbol{\theta}_A^*, \phi(\boldsymbol{s}_{n,b})\rangle\right|, \\
\mathrm{Diff}_2 &= \sum_{n\in[N],b\in[B]} \left|G_\mathrm{K}\left(\boldsymbol{\alpha}_{A_{I_{n,b}}}^*, \boldsymbol{s}_{n,b}\right) - \left\langle\boldsymbol{\theta}_{A_{I_{n,b}}}^*, \phi(\boldsymbol{s}_{n,b})\right\rangle\right|,
\end{aligned}$$

we can obtain the regret upper bound of SPUIR-Kernel as follows:

$$\mathrm{Reg}_\mathrm{K}(N, B) \le \mathrm{Reg}_\mathrm{RF}(N, B) + \mathrm{Diff}_1 + \mathrm{Diff}_2,$$

where the upper bound of $\mathrm{Reg}_\mathrm{RF}(N, B)$ can be obtained by applying Theorem 4 to the random feature space which is of the same order as that of SPUIR. Furthermore, the key of the proof is bounding Diff_1 and Diff_2. Next, we bound the term $\left|G_\mathrm{K}\left(\boldsymbol{\alpha}_A^*, \boldsymbol{s}\right) - \langle\boldsymbol{\theta}_A^*, \phi(\boldsymbol{s})\rangle\right|$ that is the key part of Diff_1 and Diff_2. According to the representer theorem, the reward parameter vector in the random feature space can be expressed by $\boldsymbol{\theta}_A^* = \sum_{n\in[N],b\in[B]} \alpha_{n,b,A}^* \phi(\boldsymbol{s}_{n,b})$, yielding that

$$\begin{aligned}
&\left|G_\mathrm{K}\left(\boldsymbol{\alpha}_A^*, \boldsymbol{s}\right) - \langle\boldsymbol{\theta}_A^*, \phi(\boldsymbol{s})\rangle\right| \\
=& \left|\sum_{n\in[N],b\in[B]} \alpha_{n,b,A}^* \kappa(\boldsymbol{s}, \boldsymbol{s}_{n,b}) - \langle\boldsymbol{\theta}_A^*, \phi(\boldsymbol{s})\rangle\right| \\
\le& \sum_{n\in[N],b\in[B]} \left|\alpha_{n,b,A}^*\right| \left|\kappa(\boldsymbol{s}, \boldsymbol{s}_{n,b}) - \langle\phi(\boldsymbol{s}_{n,b}), \phi(\boldsymbol{s})\rangle\right| \\
\le& \varepsilon_\mathrm{RF}\|G_\mathrm{K}\|_1,
\end{aligned}$$

where $\|G_\mathrm{K}\|_1 = \sum_{n\in[N],b\in[B]} |\alpha_{n,b,A}^*|$ denotes the ℓ_1-norm of G_K, and ε_RF denotes the approximation error bound of random features, i.e., $|\kappa(\boldsymbol{s}, \boldsymbol{s}_{n,b}) - \langle\phi(\boldsymbol{s}_{n,b}), \phi(\boldsymbol{s})\rangle| \le \varepsilon_\mathrm{RF}$. According to the probabilistic error bound in (Rahimi and Recht, 2007, 2008; Feng et al., 2015), we have that, with probability at least $1 - N\delta_3$,

$$\varepsilon_\mathrm{RF} = \frac{1}{\sqrt{2N\delta_3 d_\mathrm{r}}},$$

yielding that

$$\mathrm{Diff}_1 + \mathrm{Diff}_2 \le \sqrt{\frac{2N}{\delta_3 d_\mathrm{r}}}\|G_\mathrm{K}\|_1 B.$$

Finally, comparing with the regret upper bound of SPUIR in Theorem 4, SPUIR-Kernel enjoys a regret bound with an additional error term $\sqrt{2N/\delta_3 d_\mathrm{r}}\|G_\mathrm{K}\|_1 B$ against the optimal policy in Gaussian RKHS.

□

C Detailed Experimental Settings and More Experimental Results

In this section, we provide more details and results in the experiments.

Table 1: Description of datasets in the experiments (T: number of instances; B: batch size; N: number of episodes; d: dimensionality of context; M: number of actions; C_B satisfying $B = C_B^2 N/d$)

Dataset	T	B	N	d	M	C_B
`synthetic data`	126,000	1,400	90	40	5	25.00
`Criteo-recent`	75,000	1,000	75	50	5	25.82
`Criteo-all`	1,276,000	4,000	319	50	15	25.04
`commercial product`	216,568	1,700	128	50	5	25.83

C.1 Description of Datasets

Table 1 summarizes the description of datasets used in the experiments.

Next, we provide more details about the three datasets.

Synthetic Data. Inspired by the experiments in (Saito et al., 2020), the synthetic data generation procedure was formulated as follows, which simulates the streaming recommendation environment.

- Context $\boldsymbol{s}_i \in \mathbb{R}^d$: we drew elements of $\boldsymbol{s}_i$ independently from a Gaussian distribution $\mathcal{N}(0.1, 0.2^2)$, where $d = 40$;
- Click-Through-Rate (CTR): the CTRs for the 5 actions were respectively set as $\{10\%, 15\%, 25\%, 20\%, 30\%\}$;
- The indicator variables of click events:

$$C_i = \begin{cases} 1 & \text{a click occurs in context } \boldsymbol{s}_i, \\ 0 & \text{otherwise.} \end{cases}$$

  We sampled the click index set according to the uniform distribution.
- Conversion rate (CVR) in context $\boldsymbol{s}_i$: when $C_i = 1$,

$$\mathrm{CVR}(\boldsymbol{s}_i) := \mathrm{sigmoid}(\langle \boldsymbol{w}_\mathrm{c}, \boldsymbol{s}_i \rangle), = \frac{1}{1 + \exp(-\langle \boldsymbol{w}_\mathrm{c}, \boldsymbol{s}_i \rangle)},$$

  where the coefficient vector $\boldsymbol{w}_\mathrm{c} \in \mathbb{R}^d$ is sampled according to a Gaussian distribution as $\boldsymbol{w}_\mathrm{c} \sim \mathcal{N}(\kappa_\mathrm{c} \mathbf{1}_d, \sigma_\mathrm{c}^2 \boldsymbol{I}_d)$, and we set different means and standard deviations for different action with $\kappa_\mathrm{c} \in [0 : -0.2 : -0.8]$ and $\sigma_\mathrm{c} \in [0.01 : +0.01 : 0.05]$;

Criteo Data. We used the publicly available Criteo dataset[3], consisting of Criteo's traffic on display ads over a period of two months (Chapelle, 2014), where each context consists of 8 integer features and 9 categorical features. Following the experiments in (Yoshikawa and Imai, 2018), the categorical features were represented as one-hot vectors and then concatenated to the integer features. We reduced the dimensionality of the feature vectors to 50 using principal component analysis (PCA). All of the algorithms were tested in a simulated online environment that was trained on users' logs in the Criteo dataset. Specifically, we chose several campaigns from the Criteo dataset, where each campaign represents a category of items and corresponds to an action. This online environment contains a prediction model for the CVR, which was well trained by applying DFM (Chapelle, 2014) using the true user feedbacks. This environment model was trained for each chosen campaign, whose AUCs are ranging from 70% to 90%, assuring that the online environment can provide nearly realistic feedbacks. To simulate the uncertainty of user behaviors, Gaussian noises with zero-mean were added to the model parameters. At each step, the online environment randomly selected a campaign and samples one context from this campaign, and revealed the context to the agent with a preset CTR. To generate a reasonable sequence of instances, the environment kept the order of timestamps of the contexts in each campaign. We tested our algorithms and the baselines with the following two online environments on the Criteo dataset: `Criteo-recent` contains 5 campaigns ($75,000$ instances) chosen from the recent campaigns, corresponding to 5 actions; `Criteo-all` contains 15 campaigns ($1,276,000$ instances) chosen from all the campaigns, corresponding to 15 actions.

---

[3] https://labs.criteo.com/2013/12/conversion-logs-dataset/

Data Collected from a Real Commercial App for Coupon Recommendation. To verify the effectiveness and efficiency of our algorithms on real products, we conducted experiments on a real dataset collected from a Tencent's WeChat app. We call this dataset `commercial product`, where the data were collected after the users gave consent, and did not contain any personally identifiable information or offensive content. Since this dataset from a commercial app is proprietary, we did not provide a URL. We will release this dataset after the publication of this paper. In this commercial app, after clicking a recommended coupon, a user may convert the coupon after some time, or just leave it there. The dataset was collected during a 1-month period with a subsampling, and consists of $216,568$ instances from 5 categories of coupons (including food, drink, clothing, travel, and electronics), where each context is described by 86 numerical features and 16 categorical features, including user profiles and item features in users' browsing history. Each context vector $\boldsymbol{s}$ can be seen as a user embedding summarizing her preferences in different aspects. We make each action correspond to one coupon category, representing by a reward parameter vector $\boldsymbol{\theta}_A$ for action A. The reason for this setup is that, the effects of every component in user embedding on user feedback differ for different coupon categories (food, drink, travel, clothing, electronics). Then, reward parameter vector $\boldsymbol{\theta}_A$ can be seen as an embedding representation of coupons from category A, which is online learnt using feedbacks on category A from different users. Technically, through the inner product between $\boldsymbol{\theta}_A$ and the user embedding (context) $\boldsymbol{s}$, the bandit algorithm will set larger weights for components in user embedding that are more important to category A. Overall, bandit algorithms in CBB setting are suitable choices for streaming recommendation with multiple feedback mechanisms.

The timestamps of clicks and conversions were also recorded. Following the settings on the Criteo data, we also represented the categorical features as a one-hot vector, reduced the dimensionality of the feature vectors to 50 by PCA. The action space contains 5 actions, where each corresponding to one coupon category. Due to the limitation of real online experiments, in this experiment, we still trained DFM using the true user feedbacks as the online environment, where AUCs range from 75% to 90%.

To simulate the real environment under partial-information feedback, The experiments were conducted in environments where the distribution of the initialization data is atypical. Specifically, in the experiments, we set different numbers of the initialization instances for each action. In the synthetic environment, we set the number of the initial instances as $140, 210, 350, 280, 420$ for the 5 actions, respectively. In `Criteo-recent`, we set the proportion of the initial instances as $0.1, 0.15, 0.25, 0.2, 0.3$ for the 5 actions, and set the number of the initial instances as $[100 : 23 : 423]$ for the 15 actions in `Criteo-all`.

C.2 Detailed Specification of Hyperparameters

The average reward was used to evaluate the accuracy of algorithms, which is computed by $\frac{1}{nB}\sum_{k=1}^{n}\sum_{b=1}^{B} R_{k,b,A}^{\mathrm{true}}$ for the first n episodes, where $R_{k,b,A}^{\mathrm{true}}$ is the true reward of action A at step b in the k-th episode. In these experiments, the true reward is defined by $R_{k,b,A}^{\mathrm{true}} = \lambda_{\mathrm{c}} C_{k,b,A} + (1-\lambda_{\mathrm{c}}) V_{k,b,A}$ ($C_{k,b,A}$ and $V_{k,b,A}$ denote true binary variables of user click and conversion when executing action A given context $\boldsymbol{s}_{k,b}$), where $\lambda_{\mathrm{c}} = 0.01$ on the synthetic data, Criteo Data, and commercial product data, respectively. As in most contextual bandit literature (Li et al., 2010; Chu et al., 2011), we set the regularization parameter $\lambda = 1$ in the Euclidean regularization. According to theoretical analysis in Remark 3, we set the batch size as $B = C_B^2 N/d$, set the constant $C_B \approx 25$ and the sketch size $c = 150$ on all the datasets ($B = 1400, 1000, 4000, 1700$ for `synthetic data`, `Criteo-recent`, `Criteo-all`, and `commercial product`). The regularization parameters ω, α in our policy and that in the batch UCB policy were tuned in $[0.2 : +0.2 : 1.2]$. For the SJLT in SPUIR and its variants, sketch size was set as $c = 150$ and the number of block D was selected in $\{1, 2, 4, 6\}$. Except for the rate-scheduled variants of our approaches, the imputation rate γ was selected in $[0.1 : +0.2 : 0.9]$. Besides, the discount parameter η was tuned in $[0.1 : +0.2 : 0.9]$. In the nonlinear variant of our approach SPUIR-Kernel, we selected the dimension of the random features d_{r} in $\{50, 100, 200\}$ and the kernel width of Gaussian kernel in $\{2^{-(i+1)/2}, i = [-12 : 2 : 12]\}$.

We provide the detailed specification of experimental setups in different datasets, as shown in Table 2.

Rate-Scheduled Approach. We equip PUIR and SPUIR with a rate-scheduled approach, called PUIR-RS and SPUIR-RS, respectively. We design a rate-scheduled approach following the theoretical

Table 2: The detailed specification of experimental setups in our SPUIR and its extensions

Item	Notation	`synthetic`	`Criteo`	`commercial`
weight in reward	λ_{c}	0.01	0.01	0.01
sketch size	c	150	150	150
batch size	B	1400	1000 (recent), 4000 (all)	1700
regularization parameters	ω	0.2	0.6 (recent), 0.8 (all)	0.6
regularization parameters	α	0.6	0.2 (recent), 0.8 (all)	0.8
imputation rate	γ	0.7	0.1 (recent), 0.3 (all)	0.5
discount parameter	η	0.9	0.9 (recent), 0.1 (all)	0.3
number of block	D	1	1	1
dimension of random features	d_{r}	–	50	–
kernel width	σ_{R}	–	$2^{-(i+1)/2}, i = -4$	–

results about the imputation rate γ. From Remark 1&2, we can obtain that a larger imputation rate γ leads to a smaller variance while increasing the bias. From Remark B.1, we conclude that the additional bias term includes a monotonic decreasing function w.r.t. number of episodes under mild conditions. Therefore, instead of using a fixed imputation rate, we can gradually increase γ with the number of episodes, avoiding the large bias at the beginning of the reward imputation while achieving a small variance. Specifically, we set $\gamma = X\%$ for episodes from $(X - 10)\% \times N$ to $X\% \times N$, where $X \in [10, 100]$.

C.3 More Experimental Results

For better illustration, in Figure 3 of the manuscript, we omitted the curves of algorithms whose average rewards are 5% lower than the highest reward. Now we provide the curves of all the algorithms in Figure 2. In Table 3, we present the average reward results (mean ± std) on synthetic data and Criteo dataset. From the results in Table 3, we have the following observations: (1) The proposed PUIR, SPUIR and their rate-scheduled versions persistently achieved higher average rewards than other baselines; (2) The proposed imputation approaches achieved lower variances than the baselines (SBUCB, BLTS-B) that have comparable accuracies.

Table 4 presents the running time of all algorithms on both synthetic and Criteo datasets. The results in the table further validate that our reward imputation approaches outperformed DFM-S and BLTS-B in terms of efficiency. The variants of our algorithms utilizing sketching (SPUIR, SPUIR-RS) significantly reduced the time costs of reward imputation, taking less than twice as long to execute compared to the baselines without reward imputation (SBUCB, BEXP3, BEXP3-IPW).

Figure 3 and Figure 4 present experimental curves showcasing bias, variance, and regret. Based on the experimental results, the following conclusions can be drawn:

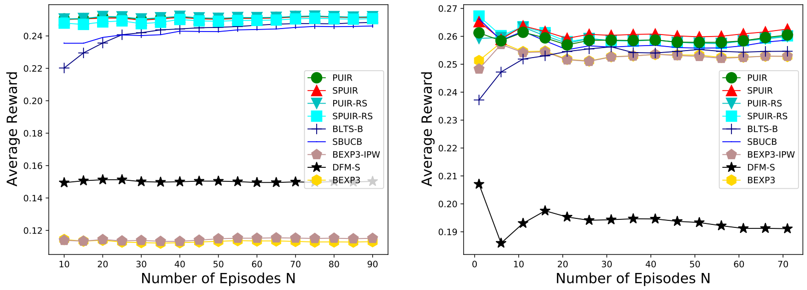
(1) The additional bias introduced by our approaches gradually diminishes with increasing episodes, as depicted in Figure 3(a). Moreover, the additional bias in comparison to the inherent bias (bias of SBUCB without reward imputation) is only a marginal fraction, approximately , as illustrated in Figure 3(b).

(2) The proposed reward imputation plays a pivotal role in significantly reducing variance. This achievement is evidenced by a nearly reduction in variance when compared to the variance of SBUCB, as shown in Figure 4(b).

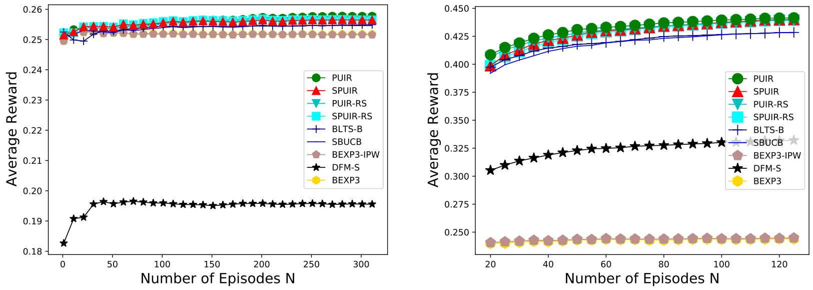
(3) The regret associated with PUIR and SPUIR is notably smaller compared to the baseline SBUCB without reward imputation. This distinction is illustrated in Figure 4(a).

D Potential Social Impact and Limitations

This is a technical work proposing provable algorithms with low variances and controllable biases, which do not learn any private information of input data. We do not foresee any potential negative societal impacts due to our work. Researchers working on online learning theory and application could benefit from this paper. In the long run, we expect that the proposed reward imputation approach



(a) `synthetic data` (b) `Criteo-recent`



(c) `Criteo-all` (d) `commercial product`

Figure 2: Average rewards of the compared algorithms, the proposed SPUIR and its variants on synthetic dataset, Criteo dataset, and the real commercial product data

Table 3: Average rewards (mean ± std) on synthetic dataset and Criteo dataset

Algorithm	`synthetic data`	`Criteo-recent`	`Criteo-all`
DFM-S	0.1503 ± 0.0092	0.1913 ± 0.0135	0.1955 ± 0.0074
SBUCB	0.2461 ± 0.0143	0.2587 ± 0.0156	0.2546 ± 0.0085
BEXP3	0.1131 ± 0.0075	0.2530 ± 0.0129	0.2518 ± 0.0078
BEXP3-IPW	0.1151 ± 0.0076	0.2530 ± 0.0167	0.2516 ± 0.0087
BLTS-B	0.2483 ± 0.0266	0.2547 ± 0.0169	0.2549 ± 0.0093
PUIR	0.2517 ± 0.0139	0.2605 ± 0.0148	0.2577 ± 0.0076
SPUIR	0.2514 ± 0.0134	0.2596 ± 0.0147	0.2565 ± 0.0076
PUIR-RS	0.2519 ± 0.0137	0.2600 ± 0.0147	0.2572 ± 0.0079
SPUIR-RS	0.2507 ± 0.0127	0.2595 ± 0.0148	0.2565 ± 0.0075

has the potential to contribute to the fair decision-making that may eliminate the decision bias due to the existence of unobserved reward feedbacks. Besides, we must emphasize that the CBB setting we consider is based on the premise that the agent is purely reward-driven. Thus, for different decision tasks, a suitable linear/nonlinear reward model needs to be selected for better performances.

References

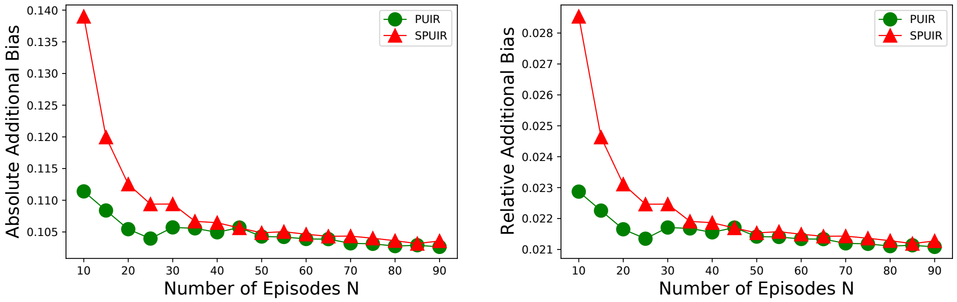
Bourgain, J., Dirksen, S., and Nelson, J. (2015). Toward a unified theory of sparse dimensionality reduction in Euclidean space. In *Proceedings of the 47th Annual ACM on Symposium on Theory of Computing*, pages 499–508.

Chapelle, O. (2014). Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1097–1105.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Table 4: Running time (second) on synthetic dataset and Criteo dataset

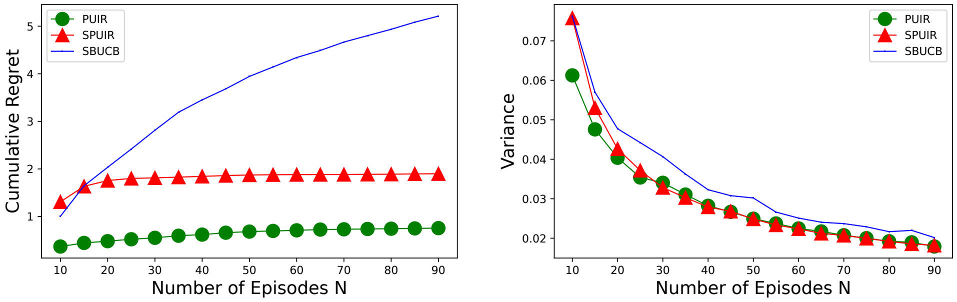
Algorithm	`synthetic data`	`Criteo-recent`	`Criteo-all`
DFM-S	120.0137 ± 6.6161	90.8571 ± 9.8601	914.0387 ± 12.2150
SBUCB	18.5502 ± 1.0452	12.6610 ± 0.7432	133.1098 ± 1.0266
BEXP3	21.2660 ± 1.1942	16.5152 ± 0.8955	157.3661 ± 1.8741
BEXP3-IPW	23.0734 ± 1.6932	16.3281 ± 1.2863	165.9273 ± 2.3430
BLTS-B	82.8094 ± 2.4060	62.1428 ± 2.0063	661.7370 ± 2.8947
PUIR	65.3308 ± 1.6082	51.1678 ± 1.2380	533.0501 ± 1.9183
SPUIR	24.2262 ± 2.0536	20.1062 ± 1.5879	202.7450 ± 2.6634
PUIR-RS	62.4110 ± 1.6723	48.1473 ± 1.3022	521.3862 ± 2.0280
SPUIR-RS	22.7946 ± 2.2360	19.0760 ± 1.9124	197.2114 ± 2.6012



(a) Absolute Additional Bias

(b) Relative Additional Bias

Figure 3: The curves of biases introduced by our proposed SPUIR and PUIR on the synthetic dataset. The bias introduced by SBUCB can be estimated as $\lambda\|\boldsymbol{\theta}_A^*\|_2 + \nu$ as in Theorem 2. The absolute additional bias introduced by PUIR and SPUIR, as depicted in Figure (a), can be estimeted by $\gamma^{\frac{1}{2}}\eta^{-\frac{1}{2}} f_{\text{Imp}}(n)$ as in Theorem 2, where $f_{\text{Imp}}(n)$ is upper bounded by C_{Imp} in Eq. (9) and is a refined upper bound of bias defined in Equation (17) within the Appendix B.1. Figure (b) depicts the relative proportion of the additional bias introduced by our proposed approaches compared to the bias of the SBUCB, called relative additional bias.



(a) Cumulative Regret

(b) Variance

Figure 4: The cumulative regret and variance of SBUCB and our proposed SPUIR as well as PUIR on the synthetic dataset. The variance is calculated by $[\boldsymbol{s}_{n,b}^{\intercal}\left(\boldsymbol{\Psi}_A^n\right)^{-1}\boldsymbol{s}_{n,b}]^{\frac{1}{2}}$ as in Theorem 2.

Feng, C., Hu, Q., and Liao, S. (2015). Random feature mapping with signed circulant matrix projection. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 3490–3496.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J. H., Glynn, P. W., and Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *CoRR*, abs/2004.06321.

Kane, D. M. and Nelson, J. (2014). Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4:1–4:23.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670.

Nelson, J. and Nguyên, H. L. (2013). OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pages 117–126.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184.

Rahimi, A. and Recht, B. (2008). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21*, pages 1313–1320.

Saito, Y., Morishita, G., and Yasui, S. (2020). Dual learning algorithm for delayed conversions. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1849–1852.

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Wang, S., Gittens, A., and Mahoney, M. W. (2017). Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3608–3616.

Yoshikawa, Y. and Imai, Y. (2018). A nonparametric delayed feedback model for conversion rate prediction. *arXiv:1802.00255v1*.

Zhang, X. and Liao, S. (2019). Incremental randomized sketching for online kernel learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7394–7403.