
CONTEXT-AWARE ONLINE RECOMMENDATION WITH BAYESIAN INCENTIVE COMPATIBILITY

Anonymous authors

Paper under double-blind review

ABSTRACT

Recommender systems play a crucial role in internet economies by connecting users with relevant products or services. However, designing effective recommender systems faces two key challenges: (1) the exploration-exploitation trade-off in balancing new product exploration against exploiting known preferences, and (2) context-aware Bayesian incentive compatibility in accounting for users' heterogeneous preferences and self-interested behaviors. This paper formalizes these challenges into a Context-aware Bayesian Incentive-Compatible Recommendation Problem (CBICRP). To address the CBICRP, we propose a two-stage algorithm (RCB) that integrates incentivized exploration with an efficient offline learning component for exploitation. In the first stage, our algorithm explores available products while maintaining context-aware Bayesian incentive compatibility to determine sufficient sample sizes. The second stage employs inverse proportional gap sampling integrated with arbitrary efficient machine learning method to ensure sublinear regret. Theoretically, we prove that RCB achieves $O(\sqrt{KdT})$ regret and satisfies Bayesian incentive compatibility (BIC). Empirically, we validate RCB's strong incentive gain, sublinear regret, and robustness through simulations and a real-world application on personalized warfarin dosing. Our work provides a principled approach for incentive-aware recommendation in online preference learning settings.

1 INTRODUCTION

In the current era of the internet economy, recommender systems have been widely adopted across various domains such as advertising, consumer goods, music, videos, news, job markets, and travel routes (Koren et al., 2009; Li et al., 2010; Covington et al., 2016; Wang et al., 2017; Zheng et al., 2018; McInerney et al., 2018; Naumov et al., 2019; Lewis et al., 2020; Bao et al., 2023). Modern recommendation markets typically involve three key stakeholders: products, users, and the platform (which acts as a *principal*). The platform collects and analyzes user data to enhance future distribution services and to respond effectively and promptly to user feedback. In these context-aware markets, the platform serves as the planner and fulfills a dual role: recommending the best available product (i.e., exploitation) and experimenting with lesser-known products to gather more information (i.e., exploration). This exploration is crucial because users often have heterogeneous preferences, and many products may initially seem unappealing. However, exploration feedback can be valuable as it provides critical insights into the products and helps determine whether they might be worthwhile for future users with similar interests. Unlike in service-oriented scenarios, these are marketplaces where choices are ultimately made by users rather than imposed by the platform.

The key challenge arises from the fact that heterogeneous users exhibit various interests to exploration and are usually lack *incentives* to adhere to the platform's recommendations due to varying interests. A myopic user is likely to choose products based solely on immediate benefits, demonstrating a bias toward exploitation over exploration. How can the platform effectively keep a balance between exploration and exploitation while taking individualized incentive compatibility into account? In other words, recommender systems commonly face two significant obstacles: (1) *exploration-exploitation tradeoff*: How can the platform design recommender systems that maximize rewards but also consider that failing to sufficiently explore all available products initially may lead to sub-optimal decisions? (2) *context-aware incentive compatibility*: How can we strategically address the tendency of heterogeneous users to behave myopically?

In this paper, we first formalize those challenges into a *Context-aware Bayesian Incentive-Compatible Recommendation Problem* (CBICRP). This protocol assumes that the platform can communicate directly with users, for example, by sending individualized product recommendations, and then observing the user’s actions and the outcomes. The key difference between this protocol and standard bandit algorithm is that user’s actions incorporate not only their personalized interests and a common public prior over all products but also the individualized message sent by the platform. That is, users will continuously evaluate the difference between products after receiving the message/recommendation sent by the platform which is formalized in a Bayesian way.

The basic multi-armed bandit (MAB) model of incentivized exploration has been examined in (Kremer et al., 2014; Che & Hörner, 2018; Mansour et al., 2020; Sellke & Slivkins, 2023), which model the recommendation policy within the framework of MAB problems and incorporating incentive compatibility constraints by agents’ Bayesian priors, but these models assume independent prior preference over products but in reality, these products share correlated prior beliefs. Subsequently, Hu et al. (2022); Sellke (2023) propose BIC recommendation policies for customers with dependent priors with Thompson sampling algorithm. However, Hu et al. (2022) considered the combinatorial semi-bandits which didn’t consider the users’ contextual information and corresponding personalized preferences over products. Similarly, Sellke (2023); Kalvit et al. (2024) considered the fixed design setting where feature x_i are product-owned and fixed rather than our setting that feature $x_{i,t}$ is user-possessed and online sampled which introduces more technical difficulty since fixed design of x can be transformed into the MAB setting and randomized design of x can not. In addition, their settings only need to learn one parameter and our setting needs to learn K arms’ parameters (Lattimore & Szepesvári, 2020; Bastani & Bayati, 2020). Besides, our framework can easily incorporate any efficient offline marching learning methods, which greatly strengthens its applicability.

Recommendation context bandit algorithm (RCB) is composed of a two-stage design’s algorithm. In the first stage, the platform explores all available products, taking into account context-aware incentive compatibility, and determines the minimal amount of information (sample size) that needs be collected for the subsequent stage. The second stage employs an *inverse proportional gap sampling bandit* integrated with any efficient plug-in offline machine learning method. This approach aims to simultaneously ensure sublinear regret and maintain context-aware BIC.

Our main contributions can be delineated into three parts:

1. We formalize the context-aware online recommendation problem under BIC constraints in §3. This formulation accommodates context-aware user preferences and incorporates BIC constraints.
2. We introduce a two-stage context-aware BIC bandit algorithm (RCB) for addressing CBICRP (see Algorithms 1 and 2). This algorithm adapts to *any efficient offline machine learning algorithm* as a component of the exploitation stage. RCB is also a decision length T -free algorithm, as long as T is greater than a constant. Moreover, we demonstrate that RCB achieves an $\mathcal{O}(\sqrt{KdT})$ regret bound (Theorem 2), where K is the number of products and d is the feature dimension. It also maintains the BIC constraints (Theorem 1).
3. Lastly, we validate the effectiveness of RCB through its performance in terms of incentive gain and sublinear regret, and its robustness across various environmental and hyperparameter settings in §6.1. Additionally, we apply our algorithm to real-world data (personalized warfarin dose allocation) and compare it with other methods to demonstrate its efficacy in §6.2.

In §2, we provide related works. In §3, we introduce the heterogeneous recommendation protocol featuring BIC and the associated challenges. §4 details the design of our algorithm. In §5, we demonstrate that RCB upholds the BIC constraint and suffers sublinear regret. §6 showcases the effectiveness and robustness of RCB through simulations and real-data studies.

Notations. We denote $[N] = [1, 2, \dots, N]$ where N is a positive integer. Define $x \in \mathbb{R}^d$ be a d -dimensional random vector. The capital $X \in \mathbb{R}^{d \times d}$ represents a $d \times d$ real-valued matrix. Let I_d represent a $d \times d$ diagonal identity matrix. We use $\mathcal{O}(\cdot)$ to denote the asymptotic complexity. We denote T as the time horizon.

2 RELATED WORKS

Incentivized Exploration. There is a growing literature about a three-way interplay of exploration, exploitation, and incentives, comprising a variety of scenarios. The study of mechanisms to incentivized exploration has been initiated by (Kremer et al., 2014). They mainly focus on deriving the Bayesian-optimal policy for the case of only two actions and deterministic rewards, where Che & Horner (2015) also propose a model to derive a BIC policy to this setting. Frazier et al. (2014) considers a different setting with monetary transfers between the platform and agents. Later, exploration-exploitation problems with multiple self-interested agents have also been studied: multiple agents engaging in exploration without a planner to coordinate them e.g., (Keller et al., 2005), context-aware pricing with model uncertainty e.g., (Besbes & Zeevi, 2009; Badanidiyuru et al., 2018), dynamic auctions e.g., (Ostrovsky & Schwarz, 2023; Han & Dai, 2023), pay-per-click ad auctions with unknown click probabilities e.g., (Babaioff et al., 2015), as well as human computation e.g., (Ho et al., 2014).

Bandit Algorithms. There are various strategies and algorithms to solve the sequential decision making problem (Bubeck et al., 2012; Slivkins et al., 2019; Maillard, 2019; Lattimore & Szepesvári, 2020), such as the ϵ -greedy (Auer et al., 2002; Chen et al., 2021; Han et al., 2022; Shi et al., 2022), explore-then-commit (Robbins, 1952; Abbasi-Yadkori et al., 2009; Li et al., 2022), upper confidence bound (UCB) (Lai & Robbins, 1985; Auer, 2002; Li et al., 2021; Wang et al., 2023), Thompson sampling (Thompson, 1933; Russo & Van Roy, 2014; Li et al., 2023), bootstrap sampling (Kveton et al., 2019; Wang et al., 2020; Wu et al., 2022; Ramprasad et al., 2023), information directed sampling (Russo & Van Roy, 2014; Hao & Lattimore, 2022), inversely proportional to the gap sampling (Abe & Long, 1999; Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2022), and betting (Waudby-Smith et al., 2022; Li et al., 2024). Additional related works can be found in Appendix §A.

3 RECOMMENDATION PROTOCOL

We first illustrate the basic *Context-aware Bayesian Incentive-Compatible Recommendation Problem* (CBICRP). Assume a sequence of T streaming users arrive sequentially to the platform and each user p_t with covariates (features) x_t such as age, race, and location where these observed covariates $\{x_t\}_{t \geq 1}$ are drawn independently from distribution \mathcal{D}_X over a deterministic set $\mathcal{X} \subset \mathbb{R}^d$. The platform has a set of products \mathcal{A} , e.g., ads/music/video/medicine, where $|\mathcal{A}| = K$. Each product (also called as arms in bandit literature) is represented as the *unknown* vector $\beta_i \in \mathbb{R}^d$. At time t , user p_t arrives at the platform and the platform need to recommend arms to the user which follows the following protocol:

1. The platform recommends the user with a best arm I_t based on user’s covariates x_t .
2. User myopically chooses an action $a_t \in \mathcal{A}$ and receives a stochastic reward $y_t(a_t) \in \mathcal{Y}$ where $\mathcal{Y} \in [0, 1]$, and leaves.
3. We assume the user provides reward $y_t(a_t)$ following the linear model $y_t(a_t) = \mu(x_t, a_t) + \eta_{t,a_t}$, where $\mu(x_t, a_t) = x_t^\top \beta_{a_t}$.¹

and $\{\eta_{t,a_t}\}_{t \geq 1}$ are σ -subgaussian random variables if $\mathbb{E}[e^{t\eta}] \leq e^{\sigma^2 t^2/2}$ for every $t \in \mathbb{R}$, and independent of the covariates $\{x_t\}_{t \geq 1}$. Besides, for notation simplicity, let y_t denote the vector potential reward in $[0, 1]^K$, $\mu(x_t)$ as the vector true personalized reward in $[0, 1]^K$, and η_t as the vector noise in \mathbb{R}^d . Without loss of generality, we assume \mathcal{X} and β are bounded which means that it exists positive constants x_{\max} and b such that $\|x\|_2 \leq L, \forall x_t \in \mathcal{X}$ and $\|\beta_i\|_2 \leq b$ for all $i \in [K]$, which is a common assumption in literature (Abbasi-Yadkori et al., 2011; Bastani & Bayati, 2020; Li et al., 2021) and usually assume $L = b = 1$. It’s important to note that the reward function contains two stochastic sources: the covariate vector x_t and the noise η_t , which is general harder than the fixed design $\{x_t\}_{t \geq 1}$ in bandit (Lattimore & Szepesvári, 2020). Besides, we define the data domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and denote $\mathcal{D}_{\mathcal{Z}}$ as the probability distribution over set \mathcal{Z} .

The key difference between the above recommendation protocol with previous literature in sequential decision making (Sutton & Barto, 2018; Lattimore & Szepesvári, 2020) is that the user p_t may not follow the (best) recommendation arm I_t , that is, $I_t \neq a_t$. Users can switch to other recommended options rather than simply click or not click the best recommended product provided by

¹The discussion of the nonlinear is available in Appendix §F.

the algorithm. However, in CBICRP, the platform performs as a principal to recommend I_t and the decision a_t is made by the user based on prior knowledge over arms, and the user have the option to other products recommended by the platform. We assume the platform and all users share a prior belief over arms $\mathcal{P}_0 = \mathcal{P}_{1,0} \times \dots \times \mathcal{P}_{K,0}$ where product prior parameter $\beta_i \sim \mathcal{P}_{i,0}$ with the mean $\beta_{i,0} = \mathbb{E}[\beta_i]$ and covariance matrix $\text{var}(\beta_i) = \Sigma_{i,0}$. Additionally, given covariate x_t , denote $\mu_0(x_t, i) = \mathbb{E}[\mu(x_t, i)]$ as the prior mean reward for arm i . It's important to note that this setting is different from the bandit setup whose arm parameter β_i is unknown and fixed.

Ideally, we hope users follow the (best) recommended arm I_t even it is not the greedy option for them given that the goal of each user is to maximize her expected reward conditional on her priors over products. Here we define the event that best recommendations have been followed in the past before time t with prior knowledge \mathcal{P}_0 as $\Gamma_{t-1} = \{I_s = a_s : s \in [t-1]\} \cup \mathcal{P}_0$, which works as a public information. Then we can formally define the ϵ -Context-aware Bayesian-Incentive Compatible (CBIC) for users as follows.

Definition 1 (ϵ -CBIC). Given an *incentive budget* $\epsilon \geq 0$, a recommendation algorithm is ϵ -Context-aware Bayesian Incentive-Compatible (ϵ -CBIC) if

$$\mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | I_t = i, \Gamma_{t-1}] \geq -\epsilon, \quad \forall t \in [T], i \in [K]. \quad (1)$$

If $\epsilon = 0$, we call it Context-aware Bayesian Incentive-Compatible (CBIC). For brevity, we use the term CBIC to denote both CBIC and ϵ -CBIC throughout the following paper, unless emphasized.

This definition implies that after receiving additional information, such as the recommended arm I_t and the historical information Γ_{t-1} , the user always follow the recommended arm or at most with expected reward (informally speaking, utility) loss less than ϵ . Specifically, the user selects the arm i that maximizes the posterior mean reward, which is either the best recommended arm I_t or another arm whose posterior mean reward is within an ϵ budget of the maximum. From the perspective of the principal, it needs to *contextually* determine which arm to be recommended based on the current covariate x_t and all historical feedback $\mathfrak{S}_{1:t-1}$ at time t , where $\mathfrak{S}_{1:t} = \{(x_t, y_t, a_t)\}_{1:t}$ denotes the sigma-algebra generated by the history up to round t . The objective for the platform is to design a sequential decision-making policy $\pi = \{\pi_t(\cdot)\}_{t \geq 1}$ that maximizes the expected reward for each user while adhering to the CBIC constraint, where $\pi_t(x_t | \mathfrak{S}_{1:t-1}) : \mathcal{X} \rightarrow \mathcal{A}$ denote the arm chosen at time t . Finally, let's define the regret with respect to CBIC constraint when following the policy π . The regret $\lambda_{[T]}(\pi)$ is defined as follows:

$$\text{Reg}_{[T]}(\pi) = \sum_{t=1}^T \mathbb{E}[\mu(x_t, \pi_t^*(x_t)) - \mu(x_t, \pi_t(x_t))] \quad (2)$$

where $\pi_t^*(x_t)$ is the posterior optimal arm given all information up to $t-1$, covariate x_t , and prior knowledge \mathcal{P}_0 . The $\text{Reg}_{[T]}(\pi)$ is taken over the randomness in the realized rewards and the randomness inherent in the algorithm. Finally, we summarize the key challenge in the CBICRP:

Key challenge:

In CBICRP, users exhibit context-aware prior preferences over arms, requiring that recommended products be more valuable than those selected myopically even still within an ϵ margin of the maximum reward. Concurrently, the platform aims to maximize long-term expected rewards. Therefore, the principal challenge lies in designing an algorithm that can **simultaneously** balance the users' incentive, the platform's requirement for maximizing expected rewards, and the exploration.

4 ALGORITHMS

In this section, we introduce the Recommendation Contextual Bandit (RCB) algorithm, which is structure into two stages, the *cold start* stage and the *exploitation* stage. The objective during the cold start stage is to develop an algorithm that not only maintains CBIC for users to gain trust but also fulfills the minimal sample size requirement necessary for the subsequent algorithm requirement for the platform with minimal budget and time cost. In the second stage, the design of RCB focuses on constructing a sampling bandit algorithm that incorporate any efficient offline machine learning methods for the long term goal of the balance of freshness and exploitation. This goal is fulfilled by balancing the ϵ -budget allocation strategically and a carefully designed of sequential spread parameter $\{\gamma_m\}_m$ over algorithm's batches m .

4.1 COLD START STAGE

During the cold start stage, it needs to determine two important quantities, *minimum sample size* N for each arm and *exploration probability* L . In addition, denote $N_i(t)$ as the current number of pulls of arm i at time t , and $B_t = \{i \mid N_i(t) = N, \forall i \in [K]\}$ as the set of arms that have been pulled N times. Additionally, S_i represents the set collecting historical rewards and covariates for arm i , and $S = \{S_k\}_{k \in [K]}$ encompasses the historical information for all arms.

The cold start stage's process comprises two steps: (1) identify the most popular arm based on the context-aware preference priors, and (2) recommend the remaining arms in a manner that economically allocates the incentive budget.

(1) The Most Popular Arm's Sample Collection (MPASC). If no arm has collected N samples, meaning B_t is empty, the platform recommends arm i to agent p_t , where arm i has the highest prior mean reward with respect to agent p_t . Subsequently, agent p_t provides feedback $y_{t,i}$ according to reward model. Afterwards, the platform updates the number of pulls $N_i(t)$ and the data S_i respectively: $N_i(t) = N_i(t-1) + 1, S_i = S_i \cup (x_t, y_{t,i})$. Once an arm has been pulled N times, it is removed from further consideration and added to B_t . The principle initially verifies whether any arm has accumulated N samples. This step determines which arm is prior optimal, indicating the most popular among heterogeneous users.

(2) Rest Arm Sample's Collection (RASC). The platform initially samples a Bernoulli random variable $q_t \sim \text{Ber}(1/L)$ to determine the recommendation strategy for the current user. With a probability of $1/L$, the platform recommends exploring promoted (sample-poor) products, while with an exploitation probability of $1 - 1/L$, it suggests exploiting organic (sample-efficient) products. The optimal value of L is determined based on prior information and the incentive budget ϵ , as specified in Theorem 1 in §5.

a) Promoted Recommendation. If $q_t = 1$, the platform recommends agent p_t to explore with a promoted arm which is the highest prior mean reward arm within the set of $[K]/B_t$, representing that arms have not been pulled N times,

$$\tilde{a}_t = \underset{i \in [K]/B_t}{\operatorname{argmax}} \mathbb{E}[\mu(x_t, i)]. \quad (3)$$

Then agent p_t receives reward y_{t,\tilde{a}_t} and the platform updates the number of pulls and samples of pair of the covariate and reward respectively: $N_{\tilde{a}_t}(t) \leftarrow N_{\tilde{a}_t}(t-1) + 1, S_{\tilde{a}_t} \leftarrow S_{\tilde{a}_t} \cup (x_t, y_{t,\tilde{a}_t})$. When arm \tilde{a}_t has been pulled N times, arm \tilde{a}_t is added to set B_t .

b) Organic Recommendation. If $q_t = 0$, the platform recommends the agent p_t to exploit with the organic arm a_t^* , which is the highest expected mean reward arm conditional on S_{B_t} .

$$a_t^* = \underset{i \in [K]}{\operatorname{argmax}} \mathbb{E}[\mu(x_t, i) | S_{B_t}]. \quad (4)$$

That is, arms in B_t 's expected rewards are evaluated through posterior mean rewards and arms not in B_t 's expected rewards are evaluated through prior mean rewards. Then the agent p_t receives reward y_{t,a_t^*} , but in this case, the principal will not update $N_{a_t^*}(t)$ and $S_{a_t^*}$.

4.2 EXPLOITATION STAGE

Given the data S (defined in §4.1) collected during the cold start stage, where each arm accumulates N samples, the platform's objective in the exploitation stage is to recommend arms with higher posterior means while satisfying the CBIC constraint. Thus, the key challenge of the bandit algorithm's design lies in balancing exploitation efficiency with the allocation of the incentive budget ϵ . The general principle of the bandit algorithm involves first strategically dividing the decision points into a series of epochs of increasing length. At the beginning of each epoch, samples collected in the previous epoch are used to update the *spread parameter* γ_m to control the balance of exploration and exploitation tradeoff at epoch m , thereby informing the decisions for the current epoch. Here we first denote the m th epoch's rounds as $\mathcal{T}_m = \{t \in [2^{m-1}, 2^m), m \geq m_0\}$ and $m(t)$ representing the epoch where the current t belongs to. The cold start stage's epoch is demoted as $m_0 = \lceil 2 + \log_2 N \rceil$ and the final stage is denoted as m_1 . The principal collected data at the m th epoch denoted as $W_{\mathcal{T}_m} = \{x_t, a_t, y_t(a_t)\}_{t \in \mathcal{T}_m}$.

At epoch $m \in [m_0, m_1]$, the platform then obtains the posterior mean estimator $\hat{\beta}_i = \mathbb{E}_{\beta_i \sim p(\beta_i | W_{\mathcal{T}_{m-1}})}[\beta_i]$, where $p(\beta_i | W_{\mathcal{T}_{m-1}})$ represents the posterior distribution based on data from

Algorithm 1: Cold Start Stage

Input : $K, N, L, B, S, \{N_i(t)\}_{i \in [K]}, t = 1$.

```

1 STEP 1 - THE MOST POPULAR ARM SAMPLE COLLECTION (MPASC)
2 while there is no arm been pulled N times do
3   Agent  $p_t$  is recommended with arm  $i = \operatorname{argmax}_{j \in [K]} \mathbb{E}[\mu(x_t, j)]$  and receives reward  $y_{t,i}$ .
4   The platform updates pulls and rewards:  $N_i(t) \leftarrow N_i(t-1) + 1, S_i \leftarrow S_i \cup (x_t, y_{t,i})$ .
5   If  $N_i(t) = N$ , add  $i$  to  $B_t$ .  $t \leftarrow t + 1$ . STEP 1 stopped.
6   Update  $t \leftarrow t + 1$ .
7 STEP 2 - REST ARM SAMPLE COLLECTION (RASC)
8 while there exists an arm  $i$  such that the number of pulled  $N_i(t)$  has not reached N do
9   Samples  $q_t \sim \operatorname{Ber}(1/L)$ .
10  if  $q_t = 1$  then
11     $p_t$  is recommended to explore with the arm  $\tilde{a}_t$  based on Eq.3 and receives  $y_{t,\tilde{a}_t}$ .
12    Updates  $N_{\tilde{a}_t}(t) \leftarrow N_{\tilde{a}_t}(t-1) + 1$  and dataset  $S_{\tilde{a}_t} \leftarrow S_{\tilde{a}_t} \cup (x_t, y_{t,\tilde{a}_t})$ .
13    If  $N_{\tilde{a}_t}(t) = N$ , add  $\tilde{a}_t$  to  $B_t$ .
14  else
15     $p_t$  is recommended to exploit with the arm  $a_t^*$  based on Eq.4 and receives  $y_{t,a_t^*}$ .
16  Update  $t \leftarrow t + 1$ .

```

Algorithm 2: Exploitation Stage

Input : S , epochs m_0, m_1 , function class \mathcal{F} , learning algorithm $\text{Off}_{\mathcal{F}}$, confidence level δ .

```

1 for epoch  $m \in [m_0, m_1]$  do
2   Set  $\gamma_m = 4\sqrt{K/\mathcal{E}_{\mathcal{F},\delta}(|\mathcal{T}_{m-1}|)}$ .
3   Feed  $m-1$  epoch's data  $W_{\mathcal{T}_{m-1}}$  into the  $\text{Off}_{\mathcal{F}}$  and get  $\{\hat{\beta}_{m,i}\}_{i \in [K]}$ .
4   for  $t \in \mathcal{T}_m$  do
5     Agent  $p_t$  arrives with covariate  $x_t$ . Compute estimate  $\hat{\mu}_{m(t)}(x_t, i) = x_t^\top \hat{\beta}_{m,i}, \forall i \in [K]$ .
6     Obtain the optimal arm  $b_t = \operatorname{argmax}_{i \in [K]} \hat{\mu}_{m(t)}(x_t, i)$ .
7     Sample  $a_t \sim p_m(i)$  according to Eq.5 and observe reward  $y_t(a_t)$ .

```

$W_{\mathcal{T}_{m-1}}$). Subsequently, the platform computes the *predictive estimate reward* $\hat{\mu}_t(x_t, i) = x_t^\top \hat{\beta}_i$ for all arms. We denote $b_t = \operatorname{argmax}_{i \in [K]} \hat{\mu}_t(x_t, i)$ as the *best predictive arm*. The platform then randomly selects arm a_t according to the distribution $p_t(i)$, for $t \in \mathcal{T}_m$:

$$p_m(i) = \begin{cases} 1 - \sum_{i \neq b_t} p_t(i), & \text{if } i = b_t. \\ 1/[K + \gamma_m(\hat{\mu}_t(x_t, b_t) - \hat{\mu}_t(x_t, i))], & \text{if } i \neq b_t. \end{cases} \quad (5)$$

where the spread parameter $\gamma_m = 4\sqrt{K/\mathcal{E}_{\mathcal{F},\delta}(|\mathcal{T}_{m-1}|)}$ regulates the balance between exploration and exploitation, and $\mathcal{E}_{\mathcal{F},\delta}(|\mathcal{T}_{m-1}|)$ denotes the mean squared prediction error (MSPE) at epoch $m-1$. A smaller γ_m results in a more dispersed p_t , enhancing exploration. Conversely, a larger γ_m leads to a more concentrated p_t , focusing recommendations on the best predictive arm b_t . As the epoch progresses, γ_m increases and is inversely proportional to the square root of the MSPE. The MSPE is typically derived via cross-validation using an efficient offline statistical learning method. Below, we present the formal definition of $\mathcal{E}_{\mathcal{F},\delta}(n)$ with n i.i.d. training samples.

Definition 2. Let p be an arbitrary action selection kernel. Given a sample size of n data of the format (x_i, a_i, y_{i,a_i}) , which are i.i.d. according to $(x_i, y_i) \sim \mathcal{D}, a_i \sim p(\cdot|x_i)$, the offline learning algorithm $\text{Off}_{\mathcal{F}}$ based on the data and a general function class \mathcal{F} returns a predictor $\hat{\mu}_t(x, a) : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$. For any $\delta > 0$, with probability at least $1 - \delta$, we have $\mathbb{E}_{x \sim \mathcal{P}_X, a \sim p(\cdot|x)} [\hat{\mu}_t(x, a) - \mu(x_t, a_t)]^2 \leq \mathcal{E}_{\mathcal{F},\delta}(n)$.

Computational Cost: The cold start stage's computational cost is $\mathcal{O}(KLN)$ in expectation and the exploitation stage's computational cost are mainly based on the offline sample efficient machine learning method. Usually it needs $\mathcal{O}(K/\epsilon'^2)$ samples in expectation for non-parametric methods and $\mathcal{O}(Kd/\epsilon')$ samples in expectation for parametric methods to get the desired offline error ϵ' .

5 THEORY

In this section, we first provide necessary assumptions in §5.1 to get the N, L , and the analytical regret upper bound. Then we demonstrate that RCB simultaneously satisfies the CBIC constraints in the whole decision process in §5.2 when sample size N and probability L are well designed. In §5.3, we show RCB achieves a $\mathcal{O}(\sqrt{KdT})$ regret.

5.1 REGULARITY CONDITIONS

In order to satisfy the CBIC constraint, we list two assumptions over the prior distribution.

Assumption 1 (Prior-Posterior Distribution Assumption). Denote $G_t(i) = \min_{j \in B_t, i \in [K]/B_t} \mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | S_{B_t}]$ as the minimum prior-posterior gap when we have N samples of arm $j \in B_t$ and zero sample of arm i in the cold start stage. There exists time-independent prior constants $n_{\mathcal{P}}, \tau_{\mathcal{P}}, \rho_{\mathcal{P}} > 0$ such that $\forall n \geq n_{\mathcal{P}_0}, i \in [K]$, then $\Pr(G_t(i) \geq \tau_{\mathcal{P}_0}) \geq \rho_{\mathcal{P}_0}$.

Any given arm i can be a posteriori best arm by margin $\tau_{\mathcal{P}_0}$ with probability at least $\rho_{\mathcal{P}_0}$ after seeing sufficiently many samples from B_t . The platform provides a fighting chance for those arms from $[K]/B_t$ with a low prior mean, which means after seeing sufficiently many samples of arm $j \in B_t$ there is a positive probability that arm $i \in [K]/B_t$ (zero sample collected) is better. What's more, we assume the gap between arms are at least greater than $\tau_{\mathcal{P}_*}$ with at least probability $\rho_{\mathcal{P}_*}$ after we have $n_{\mathcal{P}_*}$ data.

Assumption 2 (Posterior Distribution Assumption). Denote $G_t(b_t) = \min_{j \neq b_t} \mathbb{E}[\mu(x_t, b_t) - \mu(x_t, j) | S]$ as the minimum posterior gap when we have N samples of each arms in the exploitation stage. There exist a uniform time-independent posterior constants $n_{\mathcal{P}_*}, \tau_{\mathcal{P}_*}, \rho_{\mathcal{P}_*} > 0$ such that $\forall n \geq n_{\mathcal{P}_*}, i \in [K]$, then $\Pr(G_t(b_t) \geq \tau_{\mathcal{P}_*}) \geq \rho_{\mathcal{P}_*}$.

The we provide the regularity conditions over covariates \mathcal{P}_X as follows to avoid the singularity.

Assumption 3 (Minimum Eigenvalue of Σ). Define the minimum eigenvalue of the covariance matrix of X as $\lambda_{\min}(\Sigma) = \lambda_{\min}(\mathbb{E}_{x \sim \mathcal{P}_X}[xx^\top])$. There exists such a $\phi_0 > 0$ satisfying that $\lambda_{\min}(\Sigma) \geq \phi_0$.

Assumption 4 (Prior Covariance Matrix Minimum Eigenvalue Assumption). For each arm i , the minimum eigenvalue of prior covariance matrix $\Sigma_{i,0}$ satisfying: (1) $\Sigma_{i,0} \succeq \lambda_{i,0} \mathbf{I}_d$. (2) $\{\lambda_{i,t}\}_{t \geq 0}$ is increasing with order $\mathcal{O}(t)$.

This assumption assumes that with more interaction and feedback occurred in the platform, users have a context-aware prior belief and this prior becomes weaker and weaker since users tend to trust the platform's recommendation rather than have strong belief for specific arms. And these minimum eigenvalues of the covariance matrix become larger which means that users are more open to those products rather than with strong opinion towards specific products. We also explore when this assumption is violated in Appendix §E.

5.2 CONTEXT-AWARE BAYESIAN INCENTIVE COMPATIBLE CONSTRAINT

Next we provide the requirements for the minimum sample size $N(\epsilon)$ and the exploration probability L to efficiently allocate the budget ϵ and effectively recommend the optimal arms to users.

Theorem 1. With Assumptions 1 - 3, and the prior follows the normal distribution, if the parameters N, L are larger than some prior-dependent constant and the platform follows the RCB algorithm, then it preserves the ϵ -CBIC property with probability at least $\rho_{\mathcal{P}_0} \rho_{\mathcal{P}_*}$. More precisely, it suffices to take

$$N(\epsilon) \geq \frac{(\sigma^2 d + 1)K^3}{\phi_0(\tau_{\mathcal{P}_*} + \epsilon)^2} \text{ and } L \geq 1 + \frac{1 - \epsilon}{\tau_{\mathcal{P}_0} \rho_{\mathcal{P}_0} + \epsilon}. \quad (6)$$

And the exploitation stage starts at $m_0(\epsilon) \geq \lceil 2 + \log_2 N(\epsilon) \rceil$.

This theorem demonstrates that RCB maintains ϵ -CBIC throughout the entire recommendation process given the lower bound of N and L . We provide that the minimum sample size $N(\epsilon)$ is cubic with respect to the number of arms K , linear in relation to the covariate dimension d , inversely quadratic to the sum of budget ϵ and the minimal optimal posterior gap $\tau_{\mathcal{P}_*}$, and inversely linear to the minimum eigenvalue of the covariance matrix of our features ϕ_0 . This critically shows the tradeoff that a relatively larger budget ϵ significantly reduces the minimal sample size needed. Additionally, the

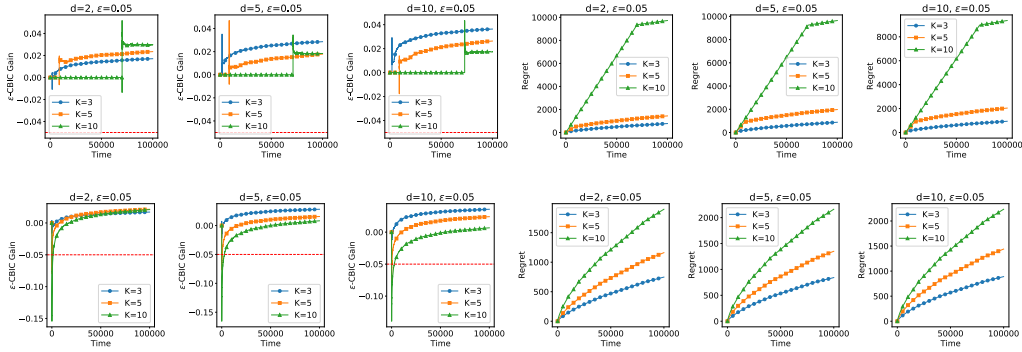


Figure 1: Incentive gain (left) and cumulative regret (right) of Setting 1 (upper) and Setting 2 (lower).

determination of the spread parameter γ_m is based on the pivot of the functionality of ϵ in $N(\epsilon)$. In RCB, given $N(\epsilon)$ in the cold start stage, γ_m for each epoch is entirely determined by the offline learning method and is independent of ϵ due to the increasing length of the epochs.

5.3 REGRET UPPER BOUND

In the following theorem, we show the regret upper bound of RCB.

Theorem 2. Given $N(\epsilon)$ and L from Theorem 1, and Assumption 4, for any $T \geq \tau_{m_0-1} + 1$, with probability at least $1 - \delta$, the regret upper bound of RCB is $\tau_{m_0-1}(\epsilon) + \mathcal{O}(\sqrt{Kd(T - \tau_{m_0-1}(\epsilon))})$.

The total regret is partitioned into two components: the cold start stage's regret τ_{m_0-1} and the exploitation stage $\mathcal{O}(\sqrt{KdT})$, where the latter depends only on the square root of the number of arms K , the covariate dimension d , and the decision horizon T . This square root dependency on T , d , and K underscores the efficiency of the approach, as detailed in (Lattimore & Szepesvári, 2020). Moreover, the effect of the ϵ budget is predominantly observed in the regret of the cold start stage, especially when T is small.

6 EXPERIMENTS

In this section, we apply RCB to synthetic data (§6.1) and real data (§6.2) to demonstrate its effectiveness by illustrating how RCB ensures sublinear regret, maintains CBIC, and exhibits robustness across various hyperparameters. Our code is available to ensure reproducibility of the results.

6.1 SIMULATION STUDIES

The goal of this section is to demonstrate that RCB algorithm can satisfy the ϵ -CBIC constraint and simultaneously secure the sublinear regret. For all settings, the following parameters need to be specified (a) *environment parameters*: time horizon T , number of arms K , feature dimension d , and noise level σ ; (b) *ϵ -CBIC parameters*: budget ϵ , prior-posterior minimum gap constants τ_{P_0} and ρ_{P_0} ; (c) *prior belief parameters*: prior P_0 , where we assume the prior follows the normal distribution.

Setting 1 (Environment Effects): We consider RCB's robustness in terms of different $K = [2, 5, 10]$, $d = [3, 5, 10]$. For rest parameters, we set $T = 10^5$, $\sigma = 0.05$, $\epsilon = 0.05$, $\tau_{P_0} = 0.01$, and $\rho_{P_0} = 0.95$. The prior are set to be $\beta_{i,0} = \mathbf{0}_d$ and $\Sigma_{i,0} = 1/5\mathbf{I}_d$.

Setting 2 (Ad-hoc Design): This scenario demonstrates the results when the platform adopts an ad-hoc approach to $N(\epsilon)$ without following the guidelines of Theorem 1. Here, N is set to $\{10, 100, 1000\}$. All other parameters remain consistent with those specified in Setting 1.

Analysis of Setting 1 (Upper part of Figure 1): Different columns in the figure represent various dimensions d , with the first three columns illustrating the ϵ -CBIC gain and the last three columns detailing the regrets observed. Our findings indicate that RCB satisfies the ϵ -CBIC property, as evidenced by the gain consistently exceeding -0.05 (dashed line), or budget not been used up. During the exploitation stage, there is an observable upward trend in the instantaneous ϵ -CBIC gain, suggesting that the recommendation system increasingly gains trust from customers (larger ϵ gain). The

right segment of the figure explores the relationship between regret, d , and K . It was observed that the regret for $K = 10$ significantly exceeds that for $K = 3$ and $K = 5$. This discrepancy arises because, to maintain the ϵ -CBIC property, the duration of the cold start stage increases cubically with K , representing a substantial cost during this initial phase. In contrast, the impact of d on cost is relatively minimal, as articulated in Theorem 1.

Analysis of Setting 2 (Lower part of Figure 1): This setting mirrors Setting 1 in terms of overall configuration. However, in this scenario, the platform does not adhere to the sample size requirements needed to satisfy the ϵ -CBIC property, opting instead for an arbitrary fixed cold start length of $N(\epsilon) = \{10, 100, 1000\}$. The simulation results for $N(\epsilon) = \{100, 1000\}$ are detailed in Appendix §E. When compared with the regret observed in Setting 1, which is at the level of 10^5 , the regret in Setting 2 is considerably lower, at approximately 10^3 . However, in terms of ϵ -CBIC gain, Setting 1 consistently shows positive gains, fully complying with the ϵ -CBIC property, whereas Setting 2 experiences periods of negative gains, particularly when the number of arms is high ($K = 10$). This negative trend is more pronounced as d increases, making it increasingly challenging to estimate an appropriate cold start length, as further discussed in Appendix §E. Notably, even with $N(\epsilon) = 1000$, the ϵ -CBIC gain remains negative for most instances when $d = 5$ or 10.

6.2 REAL DATA

We utilize a publicly available dataset from the Pharmacogenomics Knowledge Base (PharmGKB) that includes medical records of 5,700 patients treated with warfarin across various global research groups (Consortium, 2009). In the U.S., inappropriate warfarin dosing leads to about 43,000 emergency department visits annually. Traditional fixed-dose strategies can result in severe adverse effects due to initial dosing inaccuracies. Our study aims to optimize initial dosages by leveraging patient-specific factors from the cleaned data of 5,528 patients. Detailed data information and preproc are provided in Appendix E.2.

Arms Construction: We follow the arm construction as it in (Bastani & Bayati, 2020) and formulate the problem as a K -armed bandit with covariates ($K = 3$). We bucket the optimal dosages using the “clinically relevant” dosage differences: (1) Low: under 3mg/day (33% of cases), (2) Medium: 3-7mg/day (54% of cases), and (3) High: over 7mg/day (13% of cases). In particular, patients who require a low (high) dose would be at risk for excessive (inadequate) anti-coagulation under the physicians medium starting dose.

Reward Construction: For each patient, the reward is set to 1 if the dosing algorithm selects the arm corresponding to the patient’s true optimal dose; otherwise, the reward is 0. This straightforward reward function allows the regret to directly quantify the number of incorrect dosing decisions. Additionally, it is important to note that while we employ a binary reward for simplicity, we model the reward as a linear function. Despite this, RCB demonstrates robust performance in this setting, indicating its applicability for scenarios involving discrete outcomes.

Ground Truth: We estimate the true arm parameters β_i using the linear regression with the entire dataset for specific group. Besides, we scale the optimal warfarin dosing into $[0, 1]$ with minimum dosing as 0, and maximum dosing as 1. The true mean warfarin dosage is obtained from the inner production of β_i (based on the optimal arm) multiplies the covariate of this patient. Besides, for the counterfactual arm, the true mean dosage are set to be 0.

RCB Setup: The total number of trials is set at $T = 5528$, with reward noise $\hat{\sigma} = 0.054$ estimated from the true optimal dosing of warfarin after scaling. To create an online decision-making scenario, we simulate the process across 10 random permutations of patient arrivals, averaging the results over these permutations. The exploration budget ϵ is varied among $[0.025, 0.035, 0.045]$. The minimum gap τ_{p_0} is set at 0.005. The prior variance is defined as $\Sigma = [0.4, 0.6, 0.8]\mathbf{I}_d$, and the prior means are $\beta_{2,0} = 0.05 \times \mathbf{I}_d$, $\beta_{1,0} = \beta_{3,0} = \mathbf{0}_d$. Further details on hyperparameters are available in §E.2.

Evaluation Criteria: We apply four criteria to evaluate the warfarin dose decision. (1) *Regret*: The regret is optimal mean dose minus 0. (2) *ϵ -CBIC Gain*. (3) *Fraction of Incorrect Decision*: the fraction of incorrect decision. (4) *Weighted Risk Score*: the correct decision deserves 1 point and incorrect decision loss 1 point and multiple the true dosage sample proportion, which is newly proposed by us.

Result Analysis: In Table 1, we exhibits the RCB’s true dosage correction ratio and physician assigned dosage correction ratio (always choose medium) and the weighted risk score.

Table 1: Comparison RCB and physician algorithm and distribution of patients

		RCB Algo Assigned Dosage			Physician Algo Assigned Dosage			% of Patients
		Low	Medium	High	Low	Medium	High	
True Dosage	Low	50%	48%	2%	0%	100%	0%	27%
	Medium	14%	84%	2%	0%	100%	0%	60%
	High	2%	93%	5%	0%	100%	0%	13%

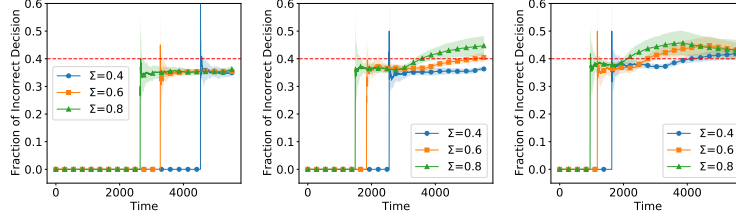


Figure 2: Left to right: fraction of incorrect decision under different setup of budgets (ϵ) $[2.5, 3.5, 4.5] \times 10^{-2}$. Dotted line represents the lasso bandit's error rate (Bastani & Bayati, 2020).

Fraction of Incorrect Decision: In Figure 2, we present the fraction of incorrect decisions, a newly metric, which is particularly relevant in the field that the non-optimal arm has high cost and the optimal arm often remains unknown and difficult to ascertain. Our findings indicate varying levels of incorrect decisions based on the size of ϵ and different prior variances. At $\epsilon = 0.025$, three prior variances show a similar fraction of incorrect decisions, with all variations approximately at a 0.35 decision error rate, which is considered state of the art when compared to the lasso bandit described in (Bastani & Bayati, 2020), which utilizes prior knowledge of non-zero feature counts. At $\epsilon = 0.035$, only $\Sigma = 0.4\mathbf{I}$ achieves the lowest fraction of incorrect decisions, approximately 0.37. When ϵ is increased to 0.045, the fraction of incorrect decisions for all three beliefs exceeds 0.4. These observations suggest that with strong prior knowledge of the optimal dosage, a smaller ϵ improves correction rates. This highlights that RCB may require an extended cold start phase to reach optimal performance and build sufficient confidence in its recommendations.

Weighted Risk Score: In Table 1, we present the dosages assigned by RCB, the true dosages, the dosages assigned by a typical physician, and the true percentage of patients for each dosage. Notably, 60% of patients require a medium dosage, while 27% should receive a low dosage, and 13% a high dosage. We use blue percentages to indicate the correction rate of dosages assigned by RCB within each true dosage, and red percentages to denote extremely incorrect decisions across these levels. The physician algorithm, which consistently prescribes a medium level dosage, achieves a 100% correctness rate at the low dosage level. Conversely, RCB attains correction rates of 50%, 84%, and 5% for the low, medium, and high dosage levels, respectively, with an *extremely* incorrect rate of 2% for the low and high levels. With respect to the weighted risk score, we find that at $\epsilon = 0.025$, the three prior beliefs achieve scores of 0.291, 0.289, and 0.274, respectively, indicating higher scores are better. When $\epsilon = 0.035$ and $\Sigma = 0.4\mathbf{I}$, the score is 0.265. The physician policy, evaluated under the metric of the weighted risk score, calculates as $-1 \times 0.27 + 1 \times 0.60 - 1 \times 0.13 = 0.20$, significantly lower than the scores provided by RCB (0.291).

7 CONCLUSION

We propose a new RCB framework to address the context-aware BIC problem, where the information about the arms needs to be learned. This approach can leverage any sample-efficient machine learning method. We theoretically prove that RCB is regret-optimal in terms of the number of arms K , dimension d , and horizon length T , all in square root order, and satisfies the ϵ -BIC constraints. Furthermore, we experimentally demonstrate that our algorithm achieves sublinear regret, is robust to different priors, dimensions, and budgets, and outperforms the state-of-the-art bandit algorithms.

REFERENCES

- Yasin Abbasi-Yadkori, András Antos, and Csaba Szepesvári. Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*, volume 92, pp. 236, 2009.
- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.
- Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pp. 3–11. Citeseer, 1999.
- Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pp. 19–26. PMLR, 2012.
- Jaime Arango, Tina Chuck, Susan S Ellenberg, Bridget Foltz, Colleen Gorman, Heidi Hinrichs, Susan McHale, Kunal Merchant, Jonathan Seltzer, Stephanie Shapley, et al. Good clinical practice training: identifying key elements and strategies for increasing training efficiency. *Therapeutic innovation & regulatory science*, 50(4):480–486, 2016.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.
- Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply, 2015.
- Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.
- Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1007–1014, 2023.
- Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.
- Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations research*, 57(6):1407–1420, 2009.
- Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- Yeon-Koo Che and Johannes Horner. Optimal design for social learning. 2015.
- Yeon-Koo Che and Johannes Hörner. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925, 2018.
- Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021.
- International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.
- Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pp. 191–198, 2016.
- Jeffrey Ely, Alexander Frankel, and Emir Kamenica. Suspense and surprise. *Journal of Political Economy*, 123(1):215–260, 2015.

594 Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with
595 regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR,
596 2020.

597 Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In
598 *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 5–22, 2014.

600 Boris Freidlin, Edward L Korn, Robert Gray, and Alison Martin. Multi-arm clinical trials of new
601 agents: some design considerations. *Clinical Cancer Research*, 14(14):4368–4371, 2008.

602 Jiale Han and Xiaowu Dai. Robust multi-item auction design using statistical learning: Overcoming
603 uncertainty in bidders’ types distributions. *arXiv preprint arXiv:2302.00941*, 2023.

604 Qiyu Han, Will Wei Sun, and Yichen Zhang. Online statistical inference for matrix contextual
605 bandit. *arXiv preprint arXiv:2212.11385*, 2022.

606 Botao Hao and Tor Lattimore. Regret bounds for information-directed reinforcement learning. *Ad-
607 vances in Neural Information Processing Systems*, 35:28575–28587, 2022.

608 Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of
609 statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

610 Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for
611 crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings
612 of the fifteenth ACM conference on Economics and computation*, pp. 359–376, 2014.

613 Johannes Hörner and Andrzej Skrzypacz. Selling information. *Journal of Political Economy*, 124
614 (6):1515–1562, 2016.

615 Xinyan Hu, Dung Ngo, Aleksandrs Slivkins, and Steven Z Wu. Incentivizing combinatorial bandit
616 exploration. *Advances in Neural Information Processing Systems*, 35:37173–37183, 2022.

617 Anand Kalvit, Aleksandrs Slivkins, and Yonatan Gur. Incentivized exploration via filtered posterior
618 sampling. *arXiv preprint arXiv:2402.13338*, 2024.

619 Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101
620 (6):2590–2615, 2011.

621 Godfrey Keller, Sven Rady, and Martin Cripps. Strategic experimentation with exponential bandits.
622 *Econometrica*, 73(1):39–68, 2005.

623 Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender
624 systems. *Computer*, 42(8):30–37, 2009.

625 Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the wisdom of the crowd. *Journal of
626 Political Economy*, 122(5):988–1012, 2014.

627 Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad
628 Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In
629 *International Conference on Machine Learning*, pp. 3601–3610. PMLR, 2019.

630 Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances
631 in applied mathematics*, 6(1):4–22, 1985.

632 Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

633 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
634 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented genera-
635 tion for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:
636 9459–9474, 2020.

637 Jiayi Li, Yuantong Li, and Xiaowu Dai. Jiayi li, yuantong li and xiaowu dai’s contribution to the
638 discussion of estimating means of bounded random variables by betting by waudby-smith and
639 ramdas. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):41–43,
640 2024.

-
- Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.
- Yuantong Li, Chi-Hua Wang, and Guang Cheng. Online forgetting process for linear regression models. In *International Conference on Artificial Intelligence and Statistics*, pp. 217–225. PMLR, 2021.
- Yuantong Li, Chi-hua Wang, Guang Cheng, and Will Wei Sun. Rate-optimal contextual online matching bandit. *arXiv preprint arXiv:2205.03699*, 2022.
- Yuantong Li, Guang Cheng, and Xiaowu Dai. Double matching under complementary preferences. *arXiv preprint arXiv:2301.10230*, 2023.
- Odalric-Ambrym Maillard. *Mathematics of statistical sequential decision making*. PhD thesis, Université de Lille, Sciences et Technologies, 2019.
- Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. *Operations Research*, 68(4):1132–1161, 2020.
- James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 31–39, 2018.
- Jaouad Mourtada and Lorenzo Rosasco. An elementary analysis of ridge regression with random design. *Comptes Rendus. Mathématique*, 360(G9):1055–1063, 2022.
- Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*, 2019.
- Michael Ostrovsky and Michael Schwarz. Reserve prices in internet advertising auctions: A field experiment. *Journal of Political Economy*, 131(12):3352–3376, 2023.
- Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544):2901–2914, 2023.
- Luis Rayo and Ilya Segal. Optimal information disclosure. *Journal of political Economy*, 118(5): 949–987, 2010.
- Herbert Robbins. Some aspects of the sequential design of experiments. 1952.
- Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.
- Mark Sellke. Incentivizing exploration with linear contexts and combinatorial actions. In *International Conference on Machine Learning*, pp. 30570–30583. PMLR, 2023.
- Mark Sellke and Aleksandrs Slivkins. The price of incentivizing exploration: A characterization via thompson sampling and sample complexity. *Operations Research*, 71(5):1706–1732, 2023.
- Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793, 2022.
- David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.
- Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

-
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.
- Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.
- Chi-Hua Wang, Yang Yu, Botao Hao, and Guang Cheng. Residual bootstrap exploration for bandit algorithms. *arXiv preprint arXiv:2002.08436*, 2020.
- Chi-Hua Wang, Zhanyu Wang, Will Wei Sun, and Guang Cheng. Online regularization toward always-valid high-dimensional dynamic pricing. *Journal of the American Statistical Association*, pp. 1–13, 2023.
- Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD’17*, pp. 1–7. 2017.
- Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/JMS Journal of Data Science*, 2022.
- Shuang Wu, Chi-Hua Wang, Yuantong Li, and Guang Cheng. Residual bootstrap exploration for stochastic linear bandit. In *Uncertainty in Artificial Intelligence*, pp. 2117–2127. PMLR, 2022.
- Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. Drn: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 world wide web conference*, pp. 167–176, 2018.

Appendix

A ADDITIONAL RELATED WORKS

A.1 INFORMATION DESIGN

Another related work is Bayesian persuasion as introduced by [Kamenica & Gentzkow \(2011\)](#), focusing on a single round where the planner’s signal is informed by the "history" of previous interactions. In exploring strategic information disclosure, [Rayo & Segal \(2010\)](#) investigate how planners can encourage better decision-making among agents by controlling information flows. The temporal aspect of information release is addressed by ([Ely et al., 2015](#); [Hörner & Skrzypacz, 2016](#)), who study the optimization of suspense and the commercial strategy of selling information over time, respectively. These contributions highlight different facets of information design.

A.2 APPLICATIONS IN MEDICAL FIELDS

Patients’ incentives are a significant barrier to conducting medical trials, especially large-scale ones for affordable treatments. BIC exploration represents a theoretical effort to overcome this challenge. Medical trials initially motivated the study of multi-arm bandits (MABs) and exploration-exploitation tradeoffs ([Villar et al., 2015](#)). However, disclosing information about the medical trial is necessary to meet the “informed consent” standards set by various regulations ([Arango et al., 2016](#)). In addition, medical trials, particularly those involving multiple treatments, underscore the relevance of BIC bandit exploration with multiple actions where traditional trials typically compare a new treatment against a placebo, but the designs incorporating multiple treatments are gaining practical importance and have been explored in biostatistics literature ([Freidlin et al., 2008](#)). BIC bandit exploration with contexts consideration is increasingly applied in adaptive trial designs, leveraging patients’ "background information" to tailor treatments.

B CBIC PROPERTY

B.1 PROOF OF THEOREM 1 - COLD START STAGE

Proof. To guarantee the CBIC property for the cold start of RCB, it suffices to have a lower bound on parameter L to avoid too many samples wasted in the cold start stage.

The cold start stage can be split into K phases and each phase last LN round in expectation based on the algorithm design except the most popular arm. Although the first phase (most popular arm) last unknown rounds, it usually lasts a pretty short period. So in the following analysis, we ignore the CBIC property in the initial sample collection stage (MPASC stage).

Due to the design of cold start stage, agents are unaware which phase they belong to, they are only aware they have $1/L$ probability to be chosen in the cold start stage. We first argue that for each agent p_t in phase $l \in [2, K]$ (except the MPASC), she has no incentive not to follow the recommended arm.

(1). If agent p_t is recommended with the arm $j \neq \tilde{a}_t$, then she knows since this arm j is the organic arm a_t^* and is not the promoted arm; so by the definition of the organic arm, it is CBIC for the agent to follow it.

(2). If agent p_t is recommended with the arm \tilde{a}_t and does not want to deviate to some other arms $j \neq \tilde{a}_t$. That is to say, we need to prove that when the platform recommend arm i , the agent p_t has no incentive to deviate the current recommendation arm i to other arm j in expected reward. From the user’s perspective, the platform needs to demonstrate this,

$$\mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | I_t = i] \Pr(I_t = i) \geq 0. \quad (\text{B.1})$$

Denote the time dependent posterior gap $G_{tij} := \mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | S_{B_t}]$ where arm i is the recommended arm by RCB and $j \neq i$, and the corresponding minimal posterior gap $G_t(i) = \min_{j \neq i} G_{tij}$. The G_{tij} represents the posterior gap between arm i and arm j at time t . The $G_t(i)$ represents the minimal gap given the current accumulative samples which is composed of two cases:

(1) $G_t(i) > 0$, that means arm i is the posterior best arm. (2) $G_t(i) \leq 0$, that means arm i is not the posterior best arm.

To satisfy the ϵ - CBIC property, we need the Eq.B.1 satisfied. By the law of iterated expectations $E[X] = E[E[X|Y]]$, we have

$$\begin{aligned} & \mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | I_t = i] \Pr(I_t = i) \\ &= \mathbb{E}[\mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | S_{B_t}] | I_t = i] \Pr(I_t = i) \\ &= \mathbb{E}[G_{tij} | I_t = i] \Pr(I_t = i) > -\epsilon. \end{aligned} \quad (\text{B.2})$$

Define two events $Q_{t,1} = \{q_t = 1\}$ and $Q_{t,0} = \{q_t = 0\}$, representing agent p_t is recommended with the promoted arm or organic arm respectively. Thus, there are two disjoint events under which agent p_t is recommended arm i , either $E_{t1} = \{G_t(i) > 0\}$ or $E_{t2} = \{G_t(i) \leq 0\} = \{G_t(i) \leq 0 \text{ and } p_t \in Q_{t,1}\}$. For notation simplicity, we denote $E_1 = E_{t1}$ and $E_2 = E_{t2}$. The reason $\{G_t(i) \leq 0\} = \{G_t(i) \leq 0 \text{ and } p_t \in Q_{t,1}\}$ is because $G_t(i) \leq 0$ happens only when $p_t \in Q_{t,1}$. So the above equation is equivalent to prove

$$\mathbb{E}[G_{tij} | I_{p_t} = i] \Pr(I_t = i) = \mathbb{E}[G_{tij} | E_1] \Pr(E_1) + \mathbb{E}[G_{tij} | E_2] \Pr(E_2) > 0. \quad (\text{B.3})$$

We observe that $\Pr(E_2) = \Pr(p_t \in Q_{t,1} | G_t(i) \leq 0) \Pr(G_t(i) \leq 0) = \Pr(G_t(i) \leq 0) / L_t$, where $q_t \sim \text{Ber}(1/L_t)$ and is time dependent and independent of other random variables. Since the event $p_t \in Q$ is independent of G_{tij} and agent p_t in $Q_{t,1}$ is randomly selected according to the Bernoulli distribution with expectation $1/L_t$. Therefore, we get:

$$\begin{aligned} & \mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | I_t = i] \Pr(I_t = i) \\ &= \mathbb{E}[G_{tij} | E_1] \Pr(E_1) + \mathbb{E}[G_{tij} | E_2] \Pr(E_2) \\ &= \mathbb{E}[G_{tij} | G_t(i) > 0] \Pr(G_t(i) > 0) + \mathbb{E}[G_{tij} | G_t(i) \leq 0 \text{ and } p_t \in Q_{t,1}] \frac{1}{L_t} \Pr(G_t(i) \leq 0) \\ &= \mathbb{E}[G_{tij} | G_t(i) > 0] \Pr(G_t(i) > 0) + \frac{1}{L_t} \mathbb{E}[G_{tij} | G_t(i) \leq 0] \Pr(G_t(i) \leq 0), \end{aligned} \quad (\text{B.4})$$

where the second equation holds by the independent property. By the fact that $\mathbb{E}[G_{tij}] = \mathbb{E}[G_{tij} | G_t(i) \leq 0] \Pr(G_t(i) \leq 0) + \mathbb{E}[G_{tij} | G_t(i) > 0] \Pr(G_t(i) > 0)$, so the above equation becomes

$$\begin{aligned} &= \mathbb{E}[G_{tij} | G_t(i) > 0] \Pr(G_t(i) > 0) + \frac{1}{L_t} \left(\mathbb{E}[G_{tij}] - \mathbb{E}[G_{tij} | G_t(i) > 0] \Pr(G_t(i) > 0) \right) \\ &= \left(1 - \frac{1}{L_t}\right) \mathbb{E}[G_{tij} | G_t(i) > 0] \Pr(G_t(i) > 0) + \frac{1}{L_t} \mathbb{E}[G_{tij}]. \end{aligned} \quad (\text{B.5})$$

We know $\mathbb{E}[G_{tij}] = \mathbb{E}[\mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | S_{B_t}]] = \mathbb{E}[\mu(x_t, i) - \mu(x_t, j)] = x_t^\top \beta_{i,0} - x_t^\top \beta_{j,0} = \mu_0(t, i) - \mu_0(t, j)$. Thus, the above equation will be

$$= \left(1 - \frac{1}{L_t}\right) \mathbb{E}[G_{tij} | G_t(i) > 0] \Pr(G_t(i) > 0) + \frac{1}{L_t} (\mu_0(t, i) - \mu_0(t, j)). \quad (\text{B.6})$$

To make the process be ϵ - CBIC, we need $\mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | I_t = i] \Pr(I_t = i) > -\epsilon$. Since we know $G_{tij} > G_t(i)$ by definition, so we have $\mathbb{E}[G_{tij} | G_t(i) > 0] > \mathbb{E}[G_t(i) | G_t(i) > 0]$. To combine them all, we get

$$\geq \left(1 - \frac{1}{L_t}\right) \mathbb{E}[G_t(i) | G_t(i) > 0] \Pr(G_t(i) > 0) + \frac{1}{L_t} (\mu_0(t, i) - \mu_0(t, j)) \geq -\epsilon. \quad (\text{B.7})$$

Thus, $\forall i, j \in [K]$, it suffices to pick L_t at time t such that:

$$\begin{aligned} L_t &\geq 1 - \frac{\mu_0(t, i) - \mu_0(t, j) + \epsilon}{\mathbb{E}[G_t(i) | G_t(i) > 0] \Pr(G_t(i) > 0) + \epsilon} \\ &= 1 + \frac{\mu_0(t, j) - \mu_0(t, i) - \epsilon}{\mathbb{E}[G_t(i) | G_t(i) > 0] \Pr(G_t(i) > 0) + \epsilon}, \end{aligned} \quad (\text{B.8})$$

Thus we need,

$$L_t \geq 1 + \frac{\overline{\Delta}_t^0 - \epsilon}{\tau_{\mathcal{P}_{0,t}} \rho_{\mathcal{P}_{0,t}} + \epsilon}, \quad (\text{B.9})$$

where $\bar{\Delta}_t^0 = \max_{i \neq j} [\mu_0(t, j) - \mu_0(t, i)]$, and $\mathbb{E}[G_t(i)|G_t(i) > 0]\Pr(G_t(i) > 0) \geq \tau_{\mathcal{P}_0, t} \rho_{\mathcal{P}_0, t}$.

By the design of the cold start stage, we know that arm i is the platform recommended arm and arm j is the arm agent p_t potentially wants to deviate to. Therefore, based on the prior knowledge, $\mu_0(t, j) \geq \mu_0(t, i)$. Since this L_t is time dependent, to get a time uniform L to let all agents have the ϵ -CBIC property, we need

$$\max_t L_t = 1 + \frac{\bar{\Delta}^0 - \epsilon}{\tau_{\mathcal{P}_0} \rho_{\mathcal{P}_0} + \epsilon}, \quad (\text{B.10})$$

where $\bar{\Delta}^0 = \max_t \bar{\Delta}_t^0$ and we know $\bar{\Delta}^0 \leq 1$, and $\tau_{\mathcal{P}_0} = \min_t \tau_{\mathcal{P}_0, t}$, $\rho_{\mathcal{P}_0} = \min_t \rho_{\mathcal{P}_0, t}$. So we have L needs to be at least

$$L \geq 1 + \frac{1 - \epsilon}{\tau_{\mathcal{P}_0} \rho_{\mathcal{P}_0} + \epsilon}. \quad (\text{B.11})$$

By selecting the time uniform L , we have the ϵ -CBIC property. □

B.2 PROOF OF THEOREM 1 - EXPLOITATION STAGE

Proof. To satisfy the CBIC property, which is any agent p_t who is recommended arm i ($I_t = i$) does not to want to switch to some other arm j in expectation. Besides, we assert that when the platform satisfies the CBIC property at the cold start stage and the CBIC property also holds when we have a minimum requirement of N , then in the following epochs, the RCB algorithm will automatically satisfy the CBIC in the exploitation stage. More formally, we need that

$$\mathbb{E}[\mu(x_t, i) - \mu(x_t, j)|I_t = i]\Pr(I_t = i) \geq -\epsilon/K, \forall t \in \text{exploitation stage}. \quad (\text{B.12})$$

Similarly to the construction of L in the previous analysis, we denote the time dependent posterior gap $G_{tij} := \mathbb{E}[\mu(x_t, i) - \mu(x_t, j)|S_*]$ where arm i is the recommended arm by RCB and $j \neq i$, where S_* is the dataset collected in the cold start stage. The corresponding minimal posterior gap $G_t(i) = \min_{j \neq i} G_{tij}$. The G_{tij} represents the posterior gap between arm i and arm j at time t . The $G_t(i)$ represents the minimal gap given the current accumulative samples which is composed of two cases: (1) If $G_t(i) > 0$, that means arm i is the best arm in terms of the posterior. (2) If $G_t(i) \leq 0$, that means arm i is not the posterior best arm. Recall the definition of $G_t(i)$, it suffices to show that

$$\begin{aligned} \mathbb{E}[G_t(i)|I_t = i] &= \mathbb{E}\left[\mathbb{E}[\mu(x_t, i) - \max_{j \in [K]/i} \mu(x_t, j)|S_*]|I_t = i\right] \\ &= \mathbb{E}\left[\mathbb{E}[\mu(x_t, i)|S_*] - \max_{j \in [K]/i} \mathbb{E}[\mu(x_t, j)|S_*]|I_t = i\right]. \end{aligned} \quad (\text{B.13})$$

Let S_* be the data set collected by the algorithm by the beginning of exploitation stage. The reward gap can be decomposed as

$$\begin{aligned} &\mathbb{E}\left[\mathbb{E}[\mu(x_t, i)|S_*] - \max_{j \in [K]/i} \mathbb{E}[\mu(x_t, j)|S_*]|I_t = i\right]\Pr(I_t = i) \\ &= \underbrace{\Pr(i = b_t) \mathbb{E}\left[\mathbb{E}[\mu(x_t, i)|S_*] - \max_{j \in [K]/i} \mathbb{E}[\mu(x_t, j)|S_*]|i = b_t\right]}_{\text{Part I Reward Gap}} \\ &\quad + \underbrace{\Pr(i \neq b_t) \mathbb{E}\left[\mathbb{E}[\mu(x_t, i)|S_*] - \max_{j \in [K]/i} \mathbb{E}[\mu(x_t, j)|S_*]|i \neq b_t\right]}_{\text{Part II Reward Gap}}, \end{aligned} \quad (\text{B.14})$$

where b_t is the highest posterior mean arm $b_t = \operatorname{argmax}_{j \in [K]} \mathbb{E}[\mu(x_t, j)|S_*]$.

Part I Reward Gap: The platform selects the highest posterior mean reward arm $b_t = \operatorname{argmax}_{j \in [K]} \mathbb{E}[\mu(x_t, j)|S_*] = \operatorname{argmax}_{j \in [K]} \hat{\mu}_m(x_t, j)$ according to the Algorithm 2's design with probability $\Pr(I_t = b_t) = 1 - \sum_{i \neq b_t} \frac{1}{K + \gamma_m u_i}$, where $u_i = \hat{\mu}_m(x_t, b_t) - \hat{\mu}_m(x_t, i)$. Denote $G_t(b_t)$ as the minimal optimal posterior gap $\mathbb{E}[\mu(x_t, b_t)|S_*] - \max_{j \in [K]/b_t} \mathbb{E}[\mu(x_t, j)|S_*]$, which is the

gap between the highest posterior mean utility and second highest posterior mean utility. By the sampling design of RCB and $\gamma_m > 0, \forall m \geq m_0$, we get that $p(b_t) \geq 1/K$, where $p(b_t)$ is the probability of selecting the highest posterior mean arm.

$$\text{Part I Reward Gap} \geq \frac{1}{K} G_t(b_t). \quad (\text{B.15})$$

Part II Reward Gap: According to the sampling structure, it has the probability that the platform recommended arm is not b_t , we have

$$\begin{aligned} \text{Part II Reward Gap} &= \Pr(i \neq b_t) \mathbb{E} \left[\mathbb{E}[\mu(x_t, i) | S_*] - \max_{j \neq i} \mathbb{E}[\mu(x_t, j) | S_*] | i \neq b_t \right] \\ &= \sum_{i \neq b_t} p_t(i) \left[\mathbb{E}[\mu(x_t, i) | S_*] - \mathbb{E}[\mu(x_t, b_t) | S_*] \right] \\ &= - \sum_{i \neq b_t} p_t(i) \left[\mathbb{E}[\mu(x_t, b_t) | S_*] - \mathbb{E}[\mu(x_t, i) | S_*] \right] \\ &= -r_t \end{aligned} \quad (\text{B.16})$$

where $r_t = \sum_{i \neq b_t} p_t(i) (\mathbb{E}[\mu(x_t, b_t) | S_*] - \mathbb{E}[\mu(x_t, i) | S_*])$. Therefore, to achieve CBIC property, we can lower bound the following term,

$$\mathbb{E}[\mu(x_t, i) - \mu(x_t, j) | I_t = i] \Pr(I_t = i) \geq \frac{G_t(b_t)}{K} - r_t \geq \frac{G_t(b_t)}{K} - \frac{K}{\gamma_m} \quad (\text{B.17})$$

The $G_t(b_t)/K$ is each step's expected gain and r_t is each step's expected loss, and by Lemma 7, we have $r_t \leq K/\gamma_m$ used in the last inequality. In order to satisfy the ϵ -CBIC property, we need

$$\frac{G_t(b_t)}{K} - \frac{K}{\gamma_m} > -\frac{\epsilon}{K} \quad (\text{B.18})$$

which is equivalent to need

$$\gamma_m(\epsilon) \geq \frac{K^2}{G_t(b_t) + \epsilon}. \quad (\text{B.19})$$

That is, in order to satisfy the ϵ -CBIC property, we need the spread parameter at each epoch $m(\geq m_0)$ is at least greater than $\gamma_m(\epsilon)$. Here $\tau_m = 2^m$ is the time step where epoch m stops. $\mathcal{E}_{\mathcal{F}, \delta}(m-1)$ represents the prediction error in the functional class \mathcal{F} when using training data collected in epoch $m-1$ that is in the time interval $(\tau_{m-2}, \tau_{m-1}]$. Based on the offline learning's result from Definition 2 given the epoch m , we have $\gamma_{m_0} = c\sqrt{K/\mathcal{E}_{\mathcal{F}, \delta}(\tau_{m_0-1} - \tau_{m_0-2})}$. So we can derive the requirement of the minimum prediction error at epoch m_0 . We need

$$\begin{aligned} \gamma_{m_0} &\geq \gamma_m(\epsilon) \\ c\sqrt{\frac{K}{\mathcal{E}_{\mathcal{F}, \frac{\delta}{2K^2}}(\tau_{m_0-1} - \tau_{m_0-2})}} &\geq \frac{K^2}{G_t(b_t) + \epsilon} \\ \mathcal{E}_{\mathcal{F}, \frac{\delta}{2K^2}}(\tau_{m_0-1} - \tau_{m_0-2}) &\leq \frac{c^2(G_t(b_t) + \epsilon)^2}{K^3} \\ \frac{c_3\sigma^2d}{\phi_0n} &\leq \frac{c^2(G_t(b_t) + \epsilon)^2}{K^3} \\ n &\geq \frac{(\sigma^2d + 1)K^3}{\phi_0(G_t(b_t) + \epsilon)^2} \end{aligned} \quad (\text{B.20})$$

where $\mathcal{E}_{\mathcal{F}, \frac{\delta}{2K^2}}(\tau_{m_0-1} - \tau_{m_0-2})$ is the prediction error with training sample size with $n = \tau_{m-1} - \tau_{m-2}$, which bounds the squared L_2 distance between $\hat{\mu}$ and μ on the test data sampled following the same data generation process as the training data. For the forth inequality, based on Corollary 1, we need the minimum sample size as $N(\epsilon) = \frac{(\sigma^2d+1)K^3}{\phi_0(G_t(b_t)+\epsilon)^2}$. We have $\tau_m = 2^m$, $\tau_{m-1} - \tau_{m-2} =$

972 $2^{m-1} - 2^{m-2} = 2^{m-2}$. By the minimum sample size requirement for the cold start stage's $N(\epsilon)$ for
 973 each arm, we know in exploitation stage, the starting epoch m_0 should be
 974

$$\begin{aligned} 975 \quad N &\leq \tau_{m-1} - \tau_{m-2}, \\ 976 \quad \log_2 N &\leq m - 2, \\ 977 \quad m &\geq m_0 = \lceil 2 + \log_2 \frac{(\sigma^2 d + 1)K^3}{\phi_0(\tau_{\mathcal{P}_*} + \epsilon)^2} \rceil. \end{aligned} \quad (\text{B.21})$$

978 where $\tau_{\mathcal{P}_*}$ is the minimum posterior mean gap based on Assumption 1. □
 979

980 C PREDICTION ERROR OF RIDGE REGRESSION WITH RANDOM DESIGN

981 From [Mourtada & Rosasco \(2022\)](#), we have the following lemmas of the prediction error of ridge
 982 regression with random design.

983 **Lemma 1.** *Assume the noise has gaussian distribution, then the excess risk bound is*

$$\mathbb{E} \left[\left\| \hat{\beta} - \beta \right\|_{\Sigma}^2 \right] \leq \left(1 + \frac{R^2}{\lambda n} \right)^2 \inf_{\beta \in \mathbb{R}^d} \{ L(\beta) + \lambda \|\beta\|^2 - L(\beta^*) \} + \left(1 + \frac{R^2}{\lambda n} \right) \frac{\sigma^2 \text{Tr}[(\Sigma + \lambda)^{-1} \Sigma]}{n} \quad (\text{C.1})$$

984 where $\|X\|_2 \leq R$ and risk $L(\beta) = \mathbb{E}[(Y - \langle \beta, X \rangle)^2]$.

985 **Lemma 2.** *For every $\lambda > 0$, we have*

$$\inf_{\beta \in \mathbb{R}^d} \{ L(\beta) + \lambda \|\beta\|^2 - L(\beta^*) \} = \lambda \left\| (\Sigma + \lambda)^{-1/2} \Sigma^{1/2} \beta^* \right\|^2 \leq \lambda \|\beta^*\|^2. \quad (\text{C.2})$$

986 **Corollary 1.** *The prediction error can be upper bounded by*

$$\mathbb{E} \left[(\hat{\beta}^\top X_t - \beta^\top X_t)^2 \right] \leq \mathbb{E} \left[\left\| \hat{\beta} - \beta \right\|_{\Sigma}^2 \right] \mathbb{E} \left[\|X_t\|_{\Sigma^{-1}}^2 \right] \leq \frac{R^2}{\lambda_{\min}(\Sigma)} \mathbb{E} \left[\left\| \hat{\beta} - \beta \right\|_{\Sigma}^2 \right] \leq \frac{c_3 \sigma^2 d}{\phi_0 n}.$$

987 *Proof.* By Lemma 1 and Lemma 2, we have

$$\begin{aligned} 988 \quad \mathbb{E} \left[\left\| \hat{\beta} - \beta \right\|_{\Sigma}^2 \right] &\leq \left(1 + \frac{R^2}{\lambda n} \right)^2 \lambda \|\beta^*\|^2 + \left(1 + \frac{R^2}{\lambda n} \right) \frac{\sigma^2 d}{n} \\ 989 \quad &\leq \left(1 + \frac{1}{c_1} \right)^2 \frac{c_1}{n} + \left(1 + \frac{1}{c_1} \right) \frac{\sigma^2 d}{n} \end{aligned} \quad (\text{C.3})$$

990 Assume that $\|\beta\|_2 \leq 1$, $R \leq 1$ and $\lambda = \frac{c_1}{n}$. So when $n \geq N = \frac{1}{c_1^2(c_2-1)^2}$ and denote $c_2 = (1 + \frac{1}{c_1})^2$.
 991 So when $n \geq N$, we have

$$\begin{aligned} 992 \quad \mathbb{E} \left[\left\| \hat{\beta} - \beta \right\|_{\Sigma}^2 \right] &\leq \frac{c_1 c_2}{n} + \frac{c_2 \sigma^2 d}{n} \\ 993 \quad &\leq \frac{c_1 c_2 + c_2 \sigma^2 d}{n} \\ 994 \quad &\leq c_3 \frac{\sigma^2 d}{n} \end{aligned} \quad (\text{C.4})$$

995 where we define $c_3 \sigma^2 d = c_1 c_2 + c_2 \sigma^2 d$ for $c_3 > 0$. Since we know $\lambda_{\min}(\Sigma) \geq \phi_0$, so the prediction
 996 error can be upper bounded by

$$\mathbb{E} \left[(\hat{\beta}^\top X_t - \beta^\top X_t)^2 \right] \leq \frac{c_3 \sigma^2 d}{\phi_0 n} \quad (\text{C.5})$$

1000 □

D PROOF OF NO REGRET LEARNING

We first denote $\Psi := \mathcal{A}^{\mathcal{X}}$ as the universal policy space, which contains all possible policies. Here we assume that $|\mathcal{X}| < \infty$ but allows $|\mathcal{X}|$ to be arbitrarily large. Focusing on such a setting enables us to highlight important ideas and key insights without the need to invoke measure theoretic arguments, which are necessary for infinite/uncountable \mathcal{X} . At epoch $m(t)$, $m = m(t)$ if t is clear, and $p_t(\cdot) = p_m(\cdot|x_t)$. We next analyze the following virtual process at round t in epoch $m(t)$. Here we use a novel virtual probability distribution $Q_m(\cdot)$ to analyze the $p_t(\cdot)$'s effect over the regret. There are three steps:

1. Algorithm samples $\pi_t \sim Q_m(\cdot)$, where $\pi_t : \mathcal{X} \rightarrow \mathcal{A}$ is a deterministic policy, and $Q_m(\cdot) : \mathcal{A}^{\mathcal{X}} \rightarrow \text{Probability Measure}$ (a probability distribution over all policies in $\mathcal{A}^{\mathcal{X}}$).
2. At time t , $x_t \sim \mathcal{P}_X$.
3. Algorithm selects $a_t = \pi_t(x_t)$.

Note that at round t , $Q_m(\cdot)$ is a stationary distribution which has already been determined at the beginning of epoch m . How to construct this $Q_m(\cdot)$? For any policy $p_m(\cdot|x)$, we can construct a unique product probability measure $Q_m(\cdot)$ on Ψ such that $Q_m(\pi) = \prod_{x \in \mathcal{P}_X} p_m(\pi(x)|x)$ for all $\pi \in \Psi$. This product measure $Q_m(\cdot)$ ensures that for every

$$p_m(a|x) = \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x) = a\} Q_m(\pi(x)). \quad (\text{D.1})$$

That is, for any arbitrary context $x \in \mathcal{X}$, the algorithm's recommended action generated by $p_m(\cdot|x)$ is probabilistically equivalent to the action generated by $Q_m(\cdot)$ through this virtual process. Since $Q_m(\cdot)$ is a dense distribution over all deterministic policies in the universal policy space, we refer to $Q_m(\cdot)$ as the "equivalent randomized policy" induced by $p_m(\cdot|x)$. Since $p_m(\cdot|x)$ is completely determined by γ_m and $\hat{\mu}_m$, we know that $Q_m(\cdot)$ is also completely determined by γ_m and $\hat{\mu}_m$. We emphasize that the exploitation stage does not actually compute $Q_m(\cdot)$, but implicit maintains $Q_m(\cdot)$ through spread parameter γ_m and estimated posterior mean $\hat{\mu}_m$, so called virtual process. That is important, as even when \mathcal{X} is known to the learner, computing the product measure $Q_m(\cdot)$ requires $\Omega(|\mathcal{X}|)$ computational cost which is intractable for large \mathcal{X} .

To get the regret upper bound, we need following notations. For any action selection kernel p and any policy π , let's define the following terms:

1. **Reward** $R_t(\pi)$: defines the expected reward in the measure of μ if it follows the policy π to select the action $\pi(x_t)$ with respect to distribution \mathcal{P}_X : $R_t(\pi) = \mathbb{E}_{x_t \sim \mathcal{D}_X} [\mu(x_t, \pi(x_t))]$
2. **Reward** \hat{R}_t : defines the expected reward in the measure of empirical $\hat{\mu}_{m(t)}$ if follows the policy π to select the action $\pi(x_t)$ with respect to distribution \mathcal{P}_X : $\hat{R}_t(\pi) = \mathbb{E}_{x_t \sim \mathcal{D}_X} [\hat{\mu}_{m(t)}(x_t, \pi(x_t))]$.
3. **Regret** $\lambda(\pi)$: defines the expected regret in the measure of μ if it follows the policy π to select the action $\pi(x_t)$ with respect to distribution \mathcal{P}_X : $\lambda(\pi) = R_t(\pi_\mu) - R_t(\pi)$.
4. **Regret** $\hat{\lambda}_t(\pi)$: defines the expected regret in the measure of empirical $\hat{\mu}_{m(t)}$ if it follow the policy π to select the action $\pi(x_t)$ with respect to distribution \mathcal{P}_X : $\hat{\lambda}_t(\pi) = \hat{R}_t(\pi_{\hat{\mu}_{m(t)}}) - \hat{R}_t(\pi)$.

where $\pi_{\hat{\mu}_{m(t)}}$ is the policy selects the action $b_t = \arg\max_{i \in [K]} \hat{\mu}_{m(t)}(x_t, i)$ according to Eq.5.

Besides, for any probability kernel p_m and any policy $\pi(\cdot)$, let $V(p_m, \pi)$ denote the expected inverse probability

$$V(p_m, \pi) = \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\frac{1}{p_m(\pi(x_t)|x_t)} \right] \quad (\text{D.2})$$

and define $\mathcal{V}_t(\pi)$ as the maximum expected inverse probability over the exploitation stage,

$$\mathcal{V}_t(\pi) = \max_{m_0 \leq m \leq m(t)-1} V(p_m, \pi) \quad (\text{D.3})$$

D.1 KEY LEMMAS

Lemma 3 (Azuma-Hoeffding Inequality). *Let $\{D_k, \mathcal{F}_k\}_{k=1}^\infty$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^n$, such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, 2, \dots, n$. Then, for all $t \geq 0$, $\Pr[|\sum_{k=1}^n D_k| \geq t] \leq 2 \exp[-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}]$.*

Lemma 4. $\forall t \in [\tau_{m-1} + 1, \tau_m]$, with probability at least $1 - \delta/2m^2$, we have

$$\mathbb{E}_{x_t, a_t} \left[\left(\hat{\mu}_{m(t)}(x_t, a_t) - \mu(x_t, a_t) \right)^2 | \mathfrak{S}_{t-1} \right] \leq \mathcal{E}_{\mathcal{F}, \delta/(2m^2)}(\tau_{m-1} - \tau_{m-2}) = \frac{16K}{\gamma_m^2} \quad (\text{D.4})$$

where $\tau_m = 2^m$. Therefore, the following event Λ_2 holds with probability at least $1 - \delta/2$:

$$\Lambda_2 := \left\{ \forall t \geq \tau_{m_0}, \mathbb{E}_{x_t, a_t} \left[\left(\hat{\mu}_{m(t)}(x_t, a_t) - \mu(x_t, a_t) \right)^2 | \mathfrak{S}_{t-1} \right] \leq \frac{16K}{\gamma_m^2} \right\}. \quad (\text{D.5})$$

Proof. Note that Algorithm 2 always collects $(x_t, a_t; y_t(a_t))$ -type data used for `OFFPOS` algorithm to conduct offline training, where $(x_t, y_t) \sim \mathcal{D}$ and $a_t \sim p_{m(t)-1}(\cdot | x_t)$ based on epoch $m(t) - 1$ collected data. Based on the prediction error of the `OFFPOS` algorithm provided in 2, we have $\forall t \in [\tau_{m-1} + 1, \tau_m]$,

$$\begin{aligned} \mathbb{E}_{x_t, a_t} \left[\left(\hat{\mu}_{m(t)}(x_t, a_t) - \mu(x_t, a_t) \right)^2 | \mathfrak{S}_{t-1} \right] &= \mathbb{E}_{x_t \sim \mathcal{P}_X, a_t \sim p_{m(t)-1}(\cdot | x_t)} \left[\left(\hat{\mu}_{m(t)}(x_t, a_t) - \mu(x_t, a_t) \right)^2 | p_{m(t)-1} \right] \\ &\leq \mathcal{E}_{\mathcal{F}, \delta/(2m^2)}(\tau_{m-1} + 1 - \tau_{m-2} - 1) = \frac{16K}{\gamma_m^2}, \end{aligned} \quad (\text{D.6})$$

where last the inequality simply follows from Lemma 4.1 and Lemma 4.2 from (Agarwal et al., 2012). \square

As we mentioned in previous, a starting point of our proof of regret upper bound is to translate the action selection kernel $p_m(\cdot | \cdot)$ into an equivalent distribution over policies $Q_m(\cdot)$. The following lemma provides a justification of such translation by showing the existence of an equivalent $Q_m(\cdot)$ for every $p_m(\cdot | \cdot)$. Here we refer Lemma 3 from (Simchi-Levi & Xu, 2022) in the following Lemma.

Lemma 5. *Fix any epoch $m \geq m_0$. The action selection scheme $p_m(\cdot | \cdot)$ is a valid probability kernel $\mathcal{B}(\mathcal{A}) \times \mathcal{X} \rightarrow [0, 1]$ over epoch m . There exists a probability measure Q_m on Ψ such that*

$$\forall a_t \in \mathcal{A}, \forall x_t \in \mathcal{X}, p_m(a_t | x_t) = \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x_t) = a_t\} Q_m(\pi) \quad (\text{D.7})$$

The following Lemma demonstrates $y_t(\pi_\mu) - y_t(a_t) - \sum_{\pi \in \Psi} Q_m(\pi) \lambda(\pi)$ is a martingale difference sequence with respect to \mathfrak{S}_t .

Lemma 6. *Fix any epoch $m \geq m_0 \in \mathbb{N}$, for any round t in epoch m , we have:*

$$\mathbb{E}_{x_t, y_t, a_t} \left[y_t(\pi_\mu) - y_t(a_t) | \mathfrak{S}_{t-1} \right] = \sum_{\pi \in \Psi} Q_m(\pi) \lambda(\pi) \quad (\text{D.8})$$

Proof. By the definition of $\mathbb{E}[y_t(a_t)]$, we have

$$\begin{aligned} &\mathbb{E}_{x_t, y_t, a_t} \left[y_t(\pi_\mu(x_t)) - y_t(a_t) | \mathfrak{S}_{t-1} \right] \\ &= \mathbb{E}_{x_t, a_t} \left[\mu(x_t, \pi_\mu(x_t)) - \mu(x_t, a_t) | \mathfrak{S}_{t-1} \right] \\ &= \mathbb{E}_{x_t \sim \mathcal{P}_X, a_t \sim p_{m(t)}(\cdot | x)} \left[\mu(x_t, \pi_\mu(x_t)) - \mu(x_t, a_t) \right] \\ &= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{a_t \in \mathcal{A}} p_{m(t)}(a_t | x_t) \left(\mu(x_t, \pi_\mu(x_t)) - \mu(x_t, a_t) \right) \right] \end{aligned} \quad (\text{D.9})$$

By Lemma 5, we have

$$\begin{aligned}
& \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{a_t \in \mathcal{A}} p_{m(t)}(a_t | x) \left(\mu(x_t, \pi_\mu(x_t)) - \mu(x_t, a_t) \right) \right] \\
&= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{a_t \in \mathcal{A}} \sum_{\pi \in \Psi} \mathbb{I}\{\pi(x_t) = a_t\} Q_m(\pi) \left(\mu(x_t, \pi_\mu(x_t)) - \mu(x_t, a_t) \right) \right] \\
&= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{\pi \in \Psi} Q_m(\pi) \left(\mu(x_t, \pi_\mu(x_t)) - \mu(x_t, \pi(x_t)) \right) \right] \tag{D.10} \\
&= \sum_{\pi \in \Psi} Q_m(\pi) \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\mu(x_t, \pi_\mu(x_t)) - \mu(x_t, \pi(x_t)) \right] \\
&= \sum_{\pi \in \Psi} Q_m(\pi) \lambda(\pi)
\end{aligned}$$

where the last equality is from the definition of the expected regret in the measure μ . \square

Lemma 7. Fix any epoch $m \geq m_0 \in \mathbb{N}$ and any round t in epoch m , we have:

$$\sum_{\pi \in \Psi} Q_m(\pi) \widehat{\text{Reg}}_t(\pi) < \frac{K}{\gamma_m}. \tag{D.11}$$

Proof. For any t in epoch m , based on the definition of $\widehat{\text{Reg}}_t(\pi) = \mathbb{E}_{x_t \sim \mathcal{D}_X} [\widehat{\mu}_{m(t)}(x_t, \pi_{\widehat{\mu}_{m(t)}}) - \widehat{\mu}_{m(t)}(x_t, \pi(x_t))]$ where $b_t = \pi_{\widehat{\mu}_{m(t)}}(x_t) = \arg\max_{i \in [K]} \widehat{\mu}_{m(t)}(x_t, i)$, we have

$$\begin{aligned}
& \sum_{\pi \in \Psi} Q_m(\pi) \widehat{\text{Reg}}_t(\pi) \\
&= \sum_{\pi \in \Psi} Q_m(\pi) \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, \pi(x_t)) \right] \\
&= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{\pi \in \Psi} Q_m(\pi) \left(\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, \pi(x_t)) \right) \right] \\
&= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{a_t \in \mathcal{A}} \sum_{\pi \in \Psi} Q_m(\pi) \mathbb{I}\{\pi(x_t) = a_t\} \left(\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t) \right) \right] \\
&= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{a_t \in \mathcal{A}} p_{m(t)}(a_t | x_t) \left(\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t) \right) \right] \\
&= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{a_t \in \mathcal{A}} \frac{1}{K + \gamma_m (\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t))} \left(\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t) \right) \right] \\
&= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\sum_{a_t \in \mathcal{A} \setminus \{b_t\}} \frac{1}{\gamma_m} \frac{\gamma_m (\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t))}{K + \gamma_m (\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t))} \right] \\
&< \frac{K - 1}{\gamma_m}. \tag{D.12}
\end{aligned}$$

where the last inequality holds by the $\frac{\gamma_m (\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t))}{K + \gamma_m (\widehat{\mu}_{m(t)}(x_t, b_t) - \widehat{\mu}_{m(t)}(x_t, a_t))} < 1$. \square

The next lemma establishes the relationship between the predicted implicit regret and the true implicit regret of any policy at round t . This lemma ensures that the predicted implicit regret of good policies are becoming more and more accurate, while the predicted implicit regret of bad policies do not need to have such property.

Lemma 8. Suppose the event Λ_2 in Lemma 4 holds, let $C_0 = 204$. For all policies π and epoch $m \geq m_0$, we have:

$$\begin{aligned}\lambda(\pi) &\leq 2\widehat{\text{Reg}}_t(\pi) + \frac{C_0 K}{\gamma_m} \\ \widehat{\text{Reg}}_t(\pi) &\leq 2\lambda(\pi) + \frac{C_0 K}{\gamma_m}\end{aligned}\tag{D.13}$$

That is, for any policy, Lemma 8 bounds the prediction error of the implicit regret estimate.

Proof. We prove it via induction on epoch m . We first consider the base case when $m = 1$ and $1 \leq t \leq \tau_1$. In this case, since $\gamma_1 = 1$, we know that $\forall \pi \in \Psi, \lambda(\pi) \leq \sqrt{K} \leq C_0 K / \gamma_1, \widehat{\text{Reg}}_t(\pi) = 0 \leq C_0 K / \gamma_1$. Note that we use condition $\mathbb{E}_{x \sim \mathcal{P}_X} [\sup_{a, a' \in \mathcal{A}} (\mu(x, a) - \mu(x, a'))] \leq \sqrt{K}$, which is very weak - in the special case of multi-armed bandits, it means "the gap between mean rewards of two actions is no greater than \sqrt{K} ". Thus the claim holds in the base case.

For the induction step, fix some epoch $m > 1$. We assume that for all epochs $m' \leq m$, all rounds t' in epoch m' , and all $\pi \in \Psi$,

$$\begin{aligned}\lambda(\pi) &\leq 2\widehat{\text{Reg}}_{t'}(\pi) + C_0 \frac{K}{\gamma_{m'}}, \\ \widehat{\text{Reg}}_{t'}(\pi) &\leq 2\lambda(\pi) + C_0 \frac{K}{\gamma_{m'}}.\end{aligned}\tag{D.14}$$

Step 1. For all rounds t in epoch m and all $\pi \in \Psi$, we first show that

$$\lambda(\pi) \leq 2\widehat{\text{Reg}}_t(\pi) + C_0 \frac{K}{\gamma_m}.$$

Based on the definition of $\lambda(\pi)$ and $\widehat{\text{Reg}}_t$, we have

$$\begin{aligned}\lambda(\pi) - \widehat{\text{Reg}}_t(\pi) &= (R_t(\pi_\mu) - R_t(\pi)) - (\widehat{R}_t(\pi_{\widehat{\mu}_{m(t)}}) - \widehat{R}_t(\pi)) \\ &\leq (R_t(\pi_\mu) - R_t(\pi)) - (\widehat{R}_t(\pi_\mu) - \widehat{R}_t(\pi)) \\ &\leq |\widehat{R}_t(\pi) - R_t(\pi)| + |R_t(\pi_\mu) - \widehat{R}_t(\pi_\mu)| \\ &\leq \frac{4\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{\gamma_m} + \frac{4\sqrt{\mathcal{V}_t(\pi_\mu)}\sqrt{K}}{\gamma_m} \\ &\leq \frac{\mathcal{V}_t(\pi)}{5\gamma_m} + \frac{\mathcal{V}_t(\pi_\mu)}{5\gamma_m} + \frac{40K}{\gamma_m}\end{aligned}\tag{D.15}$$

where the third inequality holds by Lemma 9, and the last inequality holds by the AM-GM inequality. Based on the definition of $\mathcal{V}_t(\pi)$, $\mathcal{V}_t(\pi_\mu)$ and the upper bound of the expected inverse probability from Lemma 10, there exist epochs at least one $i, j \leq m$ and $t \leq \tau_m$ such that

$$\begin{aligned}\mathcal{V}_t(\pi) &= V_t(p_i, \pi) = \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\frac{1}{p_i(\pi(x_t)|x_t)} \right] \leq K + \gamma_i \widehat{\text{Reg}}_{\tau_i}(\pi) \\ \mathcal{V}_t(\pi_\mu) &= V_t(p_j, \pi_\mu) = \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\frac{1}{p_j(\pi_\mu(x_t)|x_t)} \right] \leq K + \gamma_j \widehat{\text{Reg}}_{\tau_j}(\pi_\mu)\end{aligned}$$

Combining above two inequalities with Eq.D.15 of induction and $\gamma_i, \gamma_j \leq \gamma_m$, we have

$$\begin{aligned}\frac{\mathcal{V}_t(\pi)}{5\gamma_m} &\leq \frac{K + \gamma_i \widehat{\text{Reg}}_{\tau_i}(\pi)}{5\gamma_m} \leq \frac{K + \gamma_i(2\lambda(\pi) + C_0 \frac{K}{\gamma_i})}{5\gamma_m} \leq \frac{(1 + C_0)K}{5\gamma_m} + \frac{2}{5}\lambda(\pi) \\ \frac{\mathcal{V}_t(\pi_\mu)}{5\gamma_m} &\leq \frac{K + \gamma_j \widehat{\text{Reg}}_{\tau_j}(\pi_\mu)}{5\gamma_m} \leq \frac{K + \gamma_j(2\lambda(\pi_\mu) + C_0 \frac{K}{\gamma_j})}{5\gamma_m} = \frac{(1 + C_0)K}{5\gamma_m}\end{aligned}$$

where the last equality by $\lambda(\pi_\mu) = 0$. Combining all above, we have

$$\lambda(\pi) - \widehat{\text{Reg}}_t(\pi) \leq \frac{2}{5}\lambda(\pi) + \frac{2(1 + C_0)K}{5\gamma_m} + \frac{40K}{\gamma_m}$$

which is equivalent to

$$\lambda(\pi) \leq \frac{5}{3} \widehat{\text{Reg}}_t(\pi) + \frac{2C_0K}{3\gamma_m} + \frac{68K}{\gamma_m} \leq 2\widehat{\text{Reg}}_t(\pi) + \frac{C_0K}{\gamma_m},$$

by $C_0 \leq 204$.

Step 2. We then show for all rounds t in epoch m and all $\pi \in \Psi$,

$$\widehat{\text{Reg}}_t(\pi) \leq 2\lambda(\pi) + \frac{C_0K}{\gamma_m}. \quad (\text{D.16})$$

Similar to step 2, we can get the similar result. Thus we complete the inductive step, and the claim proves to be true for all $m \in \mathbb{N}$. \square

This following lemma is a key step to provide the relationship of $\widehat{\text{Reg}}_t(\pi)$ and $\lambda(\pi)$ in Lemma 8.

Lemma 9. For any round $t \geq \tau_{m_0} + 1$, for any policy $\pi \in \Psi$, we have

$$|\widehat{R}_t(\pi) - R_t(\pi)| \leq \frac{4\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{\gamma_{m(t)}} \quad (\text{D.17})$$

Proof. Fix any policy $\pi \in \Psi$, and any round $t > \tau_{m_0-1}$. By the definition of $\widehat{R}_t(\pi)$ and $R_t(\pi)$, we have

$$\widehat{R}_t(\pi) - R_t(\pi) = \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\widehat{\mu}_{m(t)}(x_t, \pi(x_t)) - \mu(x_t, \pi(x_t)) \right] \quad (\text{D.18})$$

Given a context x_t , define $\Delta_{x_t} = \widehat{\mu}_{m(t)}(x_t, \pi(x_t)) - \mu(x_t, \pi(x_t))$, then we have the equality $\mathbb{E}_{x_t \sim \mathcal{D}_X}[\Delta_{x_t}] = \widehat{R}_t(\pi) - R_t(\pi)$. For all $s = \tau_{m_0-1} + 1, \dots, \tau_{m(t)-1}$, we have

$$\begin{aligned} & \mathbb{E}_{a_s | x_s} \left[\left(\widehat{\mu}_{m(t)}(x_s, a_s) - \mu(x_s, a_s) \right)^2 | \mathfrak{E}_{t-1} \right] \\ &= \sum_{a_s \in \mathcal{A}} p_{m(s)}(a_s | x_s) \left[\widehat{\mu}_{m(t)}(x_s, a_s) - \mu(x_s, a_s) \right]^2 \\ &\geq p_{m(s)}(\pi(x_s) | x_s) \left[\widehat{\mu}_{m(t)}(x_s, \pi(x_s)) - \mu(x_s, \pi(x_s)) \right]^2 \\ &= p_{m(s)}(\pi(x_s) | x_s) \Delta_{x_s}^2 \end{aligned} \quad (\text{D.19})$$

where the first inequality holds by the kernel and squared terms both positive and ignoring other actions $a_s \neq \pi(x_s)$. Then we can take a sum of regret difference over the epoch m and multiply it by the maximum expected inverse probability $\mathcal{V}_t(\pi)$, defined in Eq.D.3. For the start of the epoch $m(t)$, we define $s_0 = \tau_{m(t)-1} + 1$ and assume $m(t) > m_0$, we have

$$\begin{aligned} & \mathcal{V}_t(\pi) \sum_{s=s_0}^{\tau_{m(t)}-1} \mathbb{E}_{x_s, a_s} \left[\left(\widehat{\mu}_{m(t)}(x_s, a_s) - \mu(x_s, a_s) \right)^2 | \mathfrak{E}_{t-1} \right] \\ &\geq \sum_{s=s_0}^{\tau_{m(t)}-1} V(p_{m(s)}, \pi) \mathbb{E}_{x_s, a_s} \left[\left(\widehat{\mu}_{m(t)}(x_s, a_s) - \mu(x_s, a_s) \right)^2 | \mathfrak{E}_{t-1} \right] \\ &= \sum_{s=s_0}^{\tau_{m(t)}-1} \mathbb{E}_{x_s} \left[\frac{1}{p_{m(s)}(\pi(x_s) | x_s)} \right] \mathbb{E}_{a_s | x_s} \left[\left(\widehat{\mu}_{m(t)}(x_s, a_s) - \mu(x_s, a_s) \right)^2 | \mathfrak{E}_{t-1} \right] \\ &\geq \sum_{s=s_0}^{\tau_{m(t)}-1} \left(\mathbb{E}_{x_s} \left[\sqrt{\frac{1}{p_{m(s)}(\pi(x_s) | x_s)}} \mathbb{E}_{a_s | x_s} \left[\left(\widehat{\mu}_{m(t)}(x_s, a_s) - \mu(x_s, a_s) \right)^2 | \mathfrak{E}_{t-1} \right] \right] \right)^2 \end{aligned} \quad (\text{D.20})$$

where the first inequality from the definition of $\mathcal{V}_t(\pi)$ and the second follows the Cauchy-Schwarz inequality. By the above inequality from Eq.D.19, we have the following

$$\begin{aligned}
&\geq \sum_{s=s_0}^{\tau_{m(t)}-1} \left(\mathbb{E}_{x_s} \left[\sqrt{\frac{1}{p_{m(s)}(\pi(x_s)|x_s)}} p_{m(s)}(\pi(x_s)|x_s) \Delta_{x_s}^2 \right] \right)^2 \\
&= \sum_{s=s_0}^{\tau_{m(t)}-1} \left(\mathbb{E}_{x_s} [|\Delta_{x_s}|] \right)^2 \\
&\geq \sum_{s=s_0}^{\tau_{m(t)}-1} |\hat{R}_t(\pi) - R_t(\pi)|^2 \\
&= (\tau_{m(t)} - s_0) |\hat{R}_t(\pi) - R_t(\pi)|^2
\end{aligned} \tag{D.21}$$

and the last inequality follows from the convexity of the l_1 norm and last equality holds by the definition of $\hat{R}_t(\pi)$ and $R_t(\pi)$. So we have

$$\begin{aligned}
|\hat{R}_t(\pi) - R_t(\pi)| &\leq \sqrt{\mathcal{V}_t(\pi)} \sqrt{\frac{\sum_{s=s_0}^{\tau_{m(t)}-1} \mathbb{E}_{x_s, a_s} \left[\left(\hat{\mu}_{m(t)}(x_s, a_s) - \mu(x_s, a_s) \right)^2 \middle| \mathfrak{E}_{t-1} \right]}{\tau_{m(t)} - \tau_{m(t)-1}}} \\
&\leq \frac{4\sqrt{\mathcal{V}_t(\pi)}\sqrt{K}}{\gamma_{m(t)}}
\end{aligned} \tag{D.22}$$

where the last inequality holds by the definition of the exploitation rate of $\gamma_{m(t)}$. \square

The following Lemma is a key step to control the expected inverse probability $V(p_{m(t)}, \pi)$.

Lemma 10. Fix any epoch $m \geq m_0 \in \mathbb{N}$, we have:

$$V(p_{m(t)}, \pi) \leq K + \gamma_m \widehat{\text{Reg}}_t(\pi) \tag{D.23}$$

Proof. For any policy $\pi \in \Psi$, given any context $x_t \in \mathcal{X}$, we have

$$\frac{1}{p_{m(t)}(\pi(x_t)|x_t)} \begin{cases} = K + \gamma_m (\hat{\mu}_{m(t)}(x_t, b_t) - \hat{\mu}_{m(t)}(x_t, \pi(x_t))), & \text{if } \pi(x_t) \neq b_t; \\ \leq \frac{1}{1/K} = K = K + \gamma_m (\hat{\mu}_{m(t)}(x_t, b_t) - \hat{\mu}_{m(t)}(x_t, \pi(x_t))), & \text{if } \pi(x_t) = b_t. \end{cases} \tag{D.24}$$

Based on the definition of the expected inverse probability in Eq.D.2, we have

$$\begin{aligned}
V(p_{m(t)}, \pi) &= \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\frac{1}{p_{m(t)}(\pi(x_t)|x_t)} \right] \\
&\leq K + \gamma_m \mathbb{E}_{x_t \sim \mathcal{D}_X} \left[\hat{\mu}_{m(t)}(x_t, b_t) - \hat{\mu}_{m(t)}(x_t, \pi(x_t)) \right] \\
&= K + \gamma_m \widehat{\text{Reg}}_t(\pi),
\end{aligned} \tag{D.25}$$

where the inequality follows by the condition if $\pi(x_t) = b_t$ and the last equation is followed by the definition of the expected regret in the measure of empirical $\hat{\mu}_{m(t)}$. \square

The following lemma provides the key step to provide the regret upper bound.

Lemma 11. For any $T \in \mathbb{N}$, with probability at least $1 - \delta$, the expected regret of RCB after T rounds is at most $\tau_{m_0-1} + 206K \sum_{t=\tau_{m_0-1}+1}^T 1/\gamma_{m(t)} + \sqrt{8(T - \tau_{m_0-1}) \log(2/\delta)}$.

Proof. For each round $t \geq \tau_{m_0-1} + 1$, define $M_t := y_t(\pi_\mu) - y_t(a_t) - \sum_{\pi \in \Psi} Q_m(\pi) \lambda(\pi)$ and M_t is a martingale difference sequence since $\mathbb{E}_{x_t, y_t, a_t} [M_t | \mathfrak{E}_{t-1}] = 0$ provided by Lemma 6. So we have

$$\mathbb{E}_{x_t, y_t, a_t} \left[y_t(\pi_\mu) - y_t(a_t) | \mathfrak{E}_{t-1} \right] = \sum_{\pi \in \Psi} Q_m(\pi) \lambda(\pi), \tag{D.26}$$

Since $|M_t| \leq 2$ by $y_t \in [0, 1]$, by the Azuma-Hoeffding's inequality from Lemma 3,

$$\sum_{t=\tau_{m_0-1}+1}^T M_t \leq \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \quad (\text{D.27})$$

with probability at least $1 - \delta/2$. By Lemma 4, we can upper bound the regret the with probability at least $1 - \delta/2$,

$$\begin{aligned} & \sum_{t=\tau_{m_0-1}+1}^T \mathbb{E} \left[y_t(\pi_\mu) - y_t(a_t) | \mathfrak{S}_{t-1} \right] \\ & \leq \sum_{t=\tau_{m_0-1}+1}^T \sum_{\pi \in \Psi} Q_m(\pi) \lambda(\pi) + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & \leq \sum_{t=\tau_{m_0-1}+1}^T \sum_{\pi \in \Psi} Q_m(\pi) (2\widehat{\text{Reg}}_t(\pi) + \frac{C_0 K}{\gamma_m}) + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & = \sum_{t=\tau_{m_0-1}+1}^T \sum_{\pi \in \Psi} [2Q_m(\pi) \widehat{\text{Reg}}_t(\pi) + Q_m(\pi) \frac{C_0 K}{\gamma_m}] + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & \leq \sum_{t=\tau_{m_0-1}+1}^T [\frac{2K}{\gamma_m} + \frac{C_0 K}{\gamma_m}] + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & \leq 206 \sum_{t=\tau_{m_0-1}+1}^T \frac{K}{\gamma_{m(t)}} + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \end{aligned} \quad (\text{D.28})$$

where the second inequality holds by Lemma 8 to control the implicit expected regret in π and the third inequality holds by Lemma 7 controlling the empirical regret. \square

D.2 PROOF OF THEOREM 2

Proof. By Lemma 11, with probability $1 - \delta$, we have

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}_{x_t \sim \mathcal{D}_X} (y_t(\pi_\mu) - y_t(a_t)) \\ & \leq \tau_{m_0-1} + \sum_{t=\tau_{m_0-1}+1}^T \frac{206K}{\gamma_{m(t)}} + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & \leq \tau_{m_0-1} + 52 \sum_{m=m_0}^{m_1} \sqrt{K \mathcal{E}_F(\tau_{m-2}, \tau_{m-1})} (\tau_m - \tau_{m-1}) + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})}. \end{aligned} \quad (\text{D.29})$$

With the assumption that the prior distribution \mathcal{P}_0 is normal and the variance is increasing in order $\mathcal{O}(t)$, by $\tau_m = 2^m$, we have

$$\begin{aligned} & = \tau_{m_0-1} + 52\sigma\sqrt{Kd} \sum_{m=m_0}^{m_1} \mathcal{O}(\frac{1}{\sqrt{2^{m-2}}}) 2^{m-1} + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & = \tau_{m_0-1} + 52\sigma\sqrt{Kd} \sum_{m=m_0}^{m_1} \mathcal{O}(\sqrt{2^m}) + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & \leq \tau_{m_0-1} + 52\sigma\sqrt{Kd} \int_{m_0}^{\log_2(T)} 2^{\frac{x}{2}} dx + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & \leq \tau_{m_0-1} + \frac{104}{\ln 2} \sigma\sqrt{KdT} + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \\ & < \tau_{m_0-1} + 151\sigma\sqrt{KdT} + \sqrt{8(T - \tau_{m_0-1}) \log(\frac{2}{\delta})} \end{aligned} \quad (\text{D.30})$$

E ADDITIONAL EXPERIMENTS RESULTS

E.1 ADDITIONAL SIMULATION SETTINGS AND RESULTS ANALYSIS

Setting 3 (ϵ effects): We consider the RCB algorithm's effect over different budget parameters with $\epsilon = [0.01, 0.03, 0.05]$ and prior variances $\Sigma_{i,0} = 1/\lambda \mathbf{I}_d = [1/3, 1/5, 1/10] \mathbf{I}_d$. For rest parameters, $T = 5 \times 10^4$, $K = 5$, $d = 5$, $\sigma = 0.05$, and $\beta_{i,0} = \mathbf{0}_d, \forall i \in [K]$.

Setting 4 (Prior Decay and Prior-Posterior Gap Assumption Mis-specification Effects): We also test the robustness of RCB algorithm when the Assumption 4 is mis-specified. Here we assume $\Sigma_{i,0} = [0.02, 0.04, 0.1] \mathbf{I}$ and the prior decay rate are *linear decay*, *square root decay*, and *log decay*. We set the environment parameters to be $T = 5^4$, $K = 5$, $d = 5$. We set $\epsilon = 0.05$ and the prior mean $\beta_{1,0} = [1, 1, 1, 1, 1]^T$ and $\beta_{i,0} = [0, 0, 0, 0, 0]^T$.

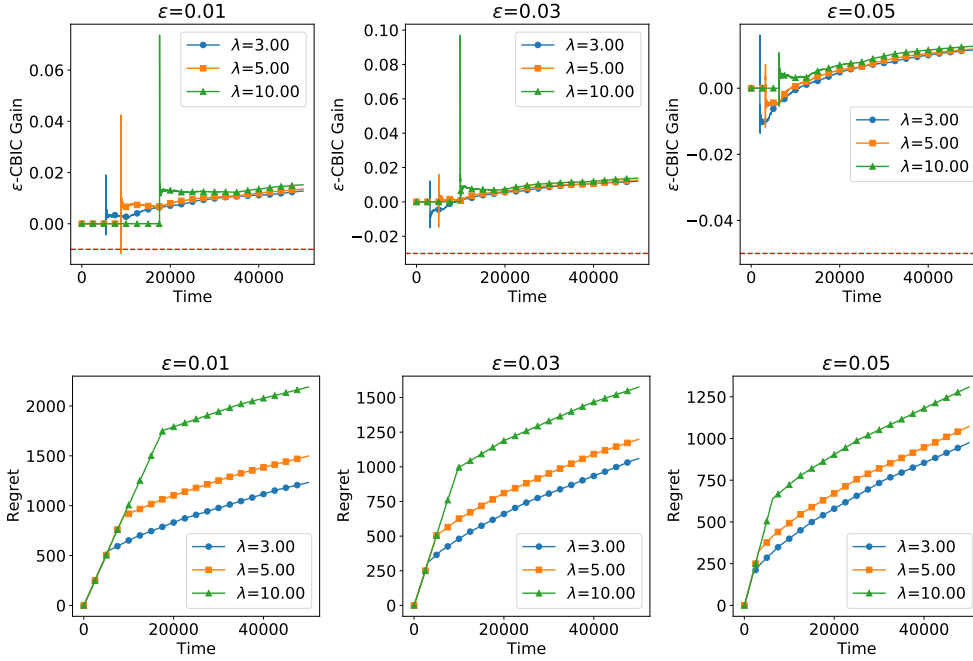


Figure 3: Gain (top) and Regret (bottom) of Setting 2.

Setting 3 - ϵ Effects Analysis: In Figure 3, three columns represent different ϵ 's effects over the ϵ -CBIC gain and regret.

For the top of the figure, we found that RCB can satisfy the ϵ -CBIC property under different ϵ and λ 's scenario. What's more, all the instantaneous gains have the uplift trend (increasing gain), which shows similar pattern to the setting 1.

The bottom shows the relationship between the regret, ϵ , and the prior variance $\Sigma_{i,0} = 1/\lambda \mathbf{I}_d$. We found that the regret of $\Sigma_{i,0} = 1/10 \mathbf{I}_d$ is much larger than the regret of $\Sigma_{i,0} = 1/3 \mathbf{I}_d$ and $\Sigma_{i,0} = 1/5 \mathbf{I}_d$. The reason is that in order to satisfy ϵ -CBIC property, the length of the cold start stage is linearly inverse proportion to the order of minimum eigenvalue ϕ_0 , which is demonstrated in Theorem 1. In other words, when the prior variance is small, it means that the customers have strong opinions over arms and the platform needs a long length of the cold start stage to make the RCB algorithm to satisfy the ϵ -CBIC property. In addition, the regret will decreases when ϵ increases. That is, when the platform wants to avoid long length of the cold start stage, it can sacrifice the ϵ to avoid a large regret, which is a trade-off between the guarantee of ϵ -CBIC property and the regret.

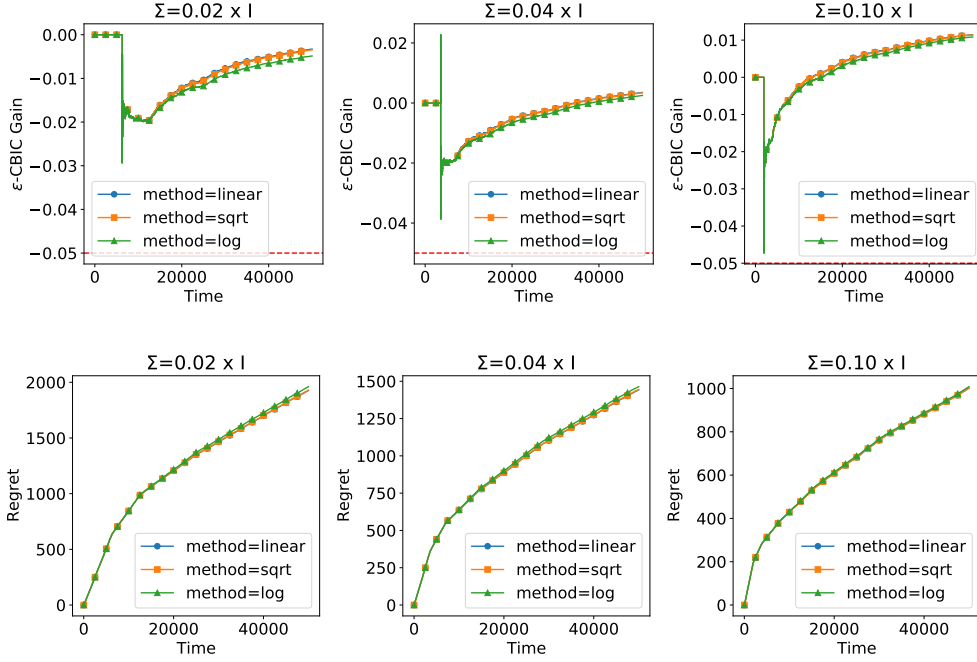


Figure 4: Gain (top) and Regret (bottom) of Setting 2.

Setting 4 - Misspecified Effects Analysis: In Figure 4, the three columns represent different prior margin $\tau_{\mathcal{P}_0}$'s effects over the regret and decay rate mis-specified over the ϵ -CBIC gain. For top figure, we found RCB can still protect the ϵ -CBIC under different Σ scenario. Besides, we found that all the instantaneous ϵ -CBIC gains still have the uplift trend, which shows similar pattern to the setting 1 and setting 2. And the linear decay rate has the largest ϵ -CBIC gain and as $\Sigma_{i,0}$ increases, the platform gains more.

The second row shows the relationship between the regret and margin, and the decay rate misspecified. We found that in any decay rate that the RCB algorithm employs, the regret of are really similar. The reason is that for any element of $\beta_{i,0}$ is small within $[0, 1]$ and the prior variance is moderate, three decay rates has similar effect. And we found that when variance increases, regret decrease. It indicates that when prior variance is large, the regret difference among three different decay rates is shrinkage. In other words, when costumers do not have strong opinions over arms (variance is large), different decay rates have similar regret effects.

E.2 ADDITIONAL REAL DATA ANALYSIS

Data Description: This data contains the true patient-specific optimal warfarin doses (which are initially unknown but are eventually found through the physician-guided dose adjustment process over the course of a few weeks) for 5528 patients with more than 70 features. It also includes patient-level covariates such as clinical factors, demographic variables, and genetic information that have been found to be predictive of the optimal warfarin dosage (Consortium, 2009). We follow the similar data construction method in (Bastani & Bayati, 2020). These covariates include:

- Demographics: gender, race, ethnicity, age, height (cm), weight (kg).
- Diagnosis: reason for treatment (e.g. deep vein thrombosis, pulmonary embolism, etc.).
- Pre-existing diagnoses: indicators for diabetes, congestive heart failure or cardiomyopathy, valve replacement, smoker status.
- Medications: indicators for potentially interacting drugs (aspirin, Tylenol, and Zocor).
- Genetics: presence of genotype variants of CYP2C9 and VKORC1.

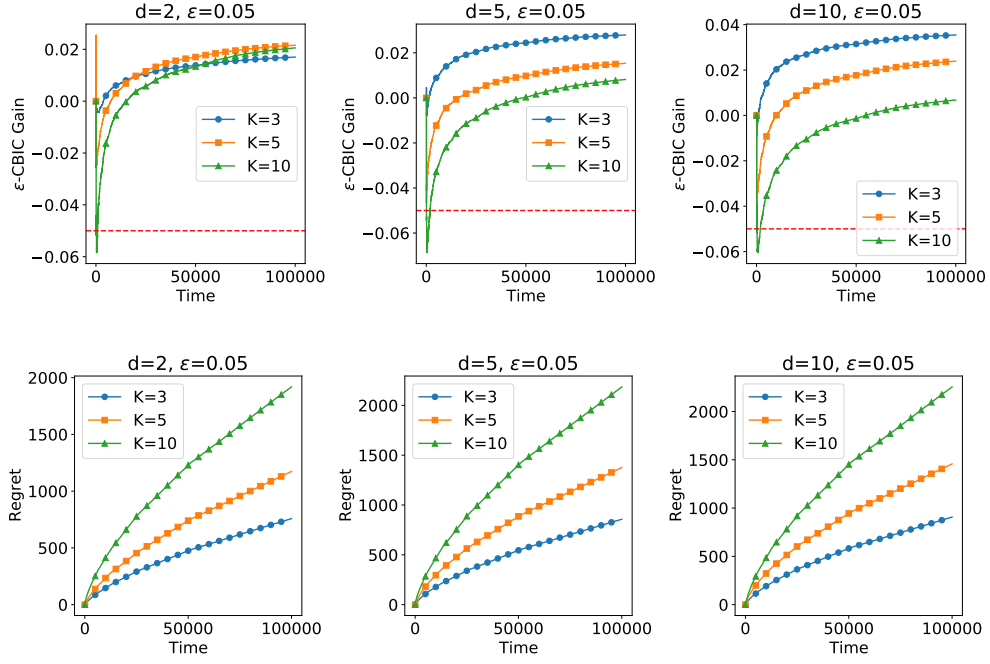


Figure 5: Gain (top) and Regret (bottom) of Setting 2 with $N = 10^2$.

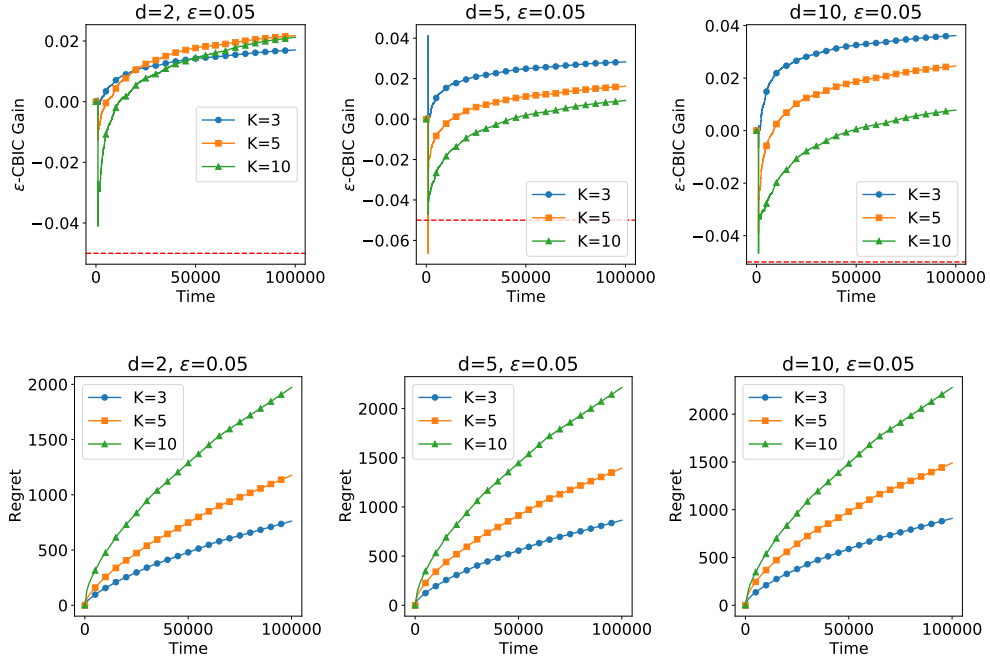


Figure 6: Gain (top) and Regret (bottom) of Setting 2 with $N = 10^3$.

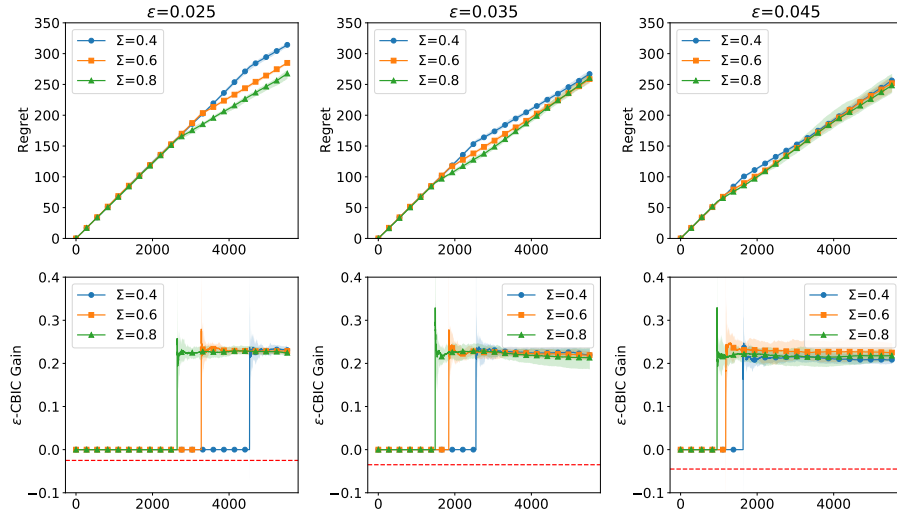


Figure 7: Regret and incentive compatibility of warfarin dosing.

The details can be found in Appendix 1 of [Consortium \(2009\)](#). All these covariates were hand-selected by professionals as being relevant to the task of warfarin dosing based on medical literature; there are no extraneously added variables. Since the detailed feature construction is not available in [\(Bastani & Bayati, 2020\)](#), we construct features follow the description in [\(Bastani & Bayati, 2020\)](#). For *diagnosis variables*, we categorize the reason for treatment with 0/1 (1 represents patients have reason for treatment, 0 represents patients have no reason or unknown reason for treatment). For *medications variables*, we only include three medications: aspirin, Tylenol, Zocor, and all other medications are set to be 0. For *genetics variables*, we considered genotype variants of CYP2C9 and VKORC1 and the rest are set to be 0. The previous feature construction aims to avoid to high dimensional feature space. All categorical variables are transformed into dummy variables and all missing values are set to 0. After the data construction, we have 70 features and 5528 patients. In [\(Bastani & Bayati, 2020\)](#), they have 93 features, which is similar to our constructions.

Model Hyperparameter Setup: The prior mean’s setup follow the fixed-dose strategy and detailed explanation is provided in the following. We assume the prior variance increases linearly over time after the *cold start*. This allows physicians decrease the confidence of their prior dose strategy and trust the RCB algorithm over time. In addition, the length of the *cold start* is determined by Theorem 1.

Addition Result Analysis.

Regret: In the first row, we show the regret of RCB with different confidence strengths (prior variance). When Σ is small that means physicians have stronger opinion over the medium dosage, and the reverse is that the physicians have weaker opinion over the medium dosage. With different prior, we found that when $\Sigma = 0.4\mathbf{I}$, it has the largest regret since we need more samples in the cold start stage to let physicians trust RCB, which means that we need a large N . Interestingly, we found that when ϵ increases (left to right), the regret difference between different prior variance shrinks because when we can tolerate with a higher ratio of non- ϵ -CBIC compatible patients, the prior’s effect decreases and the overall regret decreases because of a shorter cold start stage.

ϵ - CBIC Gain: In the second row, we show ϵ -CBIC gain of the RCB with different confidence strengths. Different prior variance has similar effect on the CBIC gain and all variants’ gain are above $-\epsilon$, which satisfies the property since the gain after the cold start stage is only determined by the posterior difference within the arm RCB selected.

F NONLINEAR REWARD DISCUSSION

If the true model has a non-linear structure, we can approximate the nonlinear functions of the covariates by using basis expansion methods in from statistical learning [\(Hastie et al., 2009\)](#).