Batched Nonparametric Contextual Bandits

Rong Jiang[1] and Cong Ma[2]

[1]Committee on Computational and Applied Mathematics, University of Chicago
[2]Department of Statistics, University of Chicago

February 2024; Revised June 2024

Abstract

We study nonparametric contextual bandits under batch constraints, where the expected reward for each action is modeled as a smooth function of covariates, and the policy updates are made at the end of each batch of observations. We establish a minimax regret lower bound for this setting and propose a novel batch learning algorithm that achieves the optimal regret (up to logarithmic factors). In essence, our procedure dynamically splits the covariate space into smaller bins, carefully aligning their widths with the batch size. Our theoretical results suggest that for nonparametric contextual bandits, a nearly constant number of policy updates can attain optimal regret in the fully online setting.

1 Introduction

Recent years have witnessed substantial progress in the field of sequential decision making under uncertainty. Especially noteworthy are the advancements in personalized decision making, where the decision maker uses side-information to make customized decision for a user. The contextual bandit framework has been widely adopted to model such problems because of its applicability and elegance [35, 53, 6]. In this framework, one interacts with an environment for a number of rounds: at each round, one is given a context, picks an action, and receives a reward. One can update the action-assignment policy based on previous observations and the goal is to maximize the expected cumulative rewards. For example, in online news recommendation, a recommendation algorithm selects an article for each newly arrived user based on the user's contextual information, and observes whether the user clicks the article or not. The goal is to try to maximize the number of clicks received. Apart from news recommendation, contextual bandits have found numerous applications in other fields such as clinical trials, personalized medicine, and online advertising [30, 62, 13].

At the core of designing a contextual bandit algorithm is deciding how to update the policy based on prior observations. A standard metric of performance for bandit algorithms is regret, which is the expected difference between the cumulative rewards obtained by an oracle who knows the optimal action for every context and that obtained by the actual algorithm under consideration. Many existing regret optimal bandit algorithms require a policy update per observation (unit) [4, 1, 39, 34]. At a first glance, such frequent policy updates are needed so that the algorithm can quickly learn the optimal action under each context and reduce regret. However, this kind of algorithm ignores an important concern in the practice of sequential decision making—the batch constraint.

In many real world scenarios, the data often arrive in batches: the statistician can only observe the outcomes of the policy at the end of a batch, and then decides what to do for the next batch. For example, this batch constraint is ubiquitous in clinical trials: statisticians need to divide the participants into batches, determine a treatment allocation policy before the batch starts, and then observe all the outcomes at the end of the batch [49]. Policy updates are made per batch instead of per unit. In fact, it is infeasible to apply unit-wise policy update in this case because observing the effect of a treatment takes time and if one waits for the result before deciding how to treat the next patient, the entire experiment will take too long to complete when the number of participants is huge. The batch constraint also appears in areas such as online marketing, crowdsourcing, and simulations [8, 50, 31, 15]. Clearly, the batch constraint presents additional

challenges to online learning. Indeed, from an information perspective, the statistician's information set is largely restricted since she can only observe all the responses at the end of a batch. The following questions naturally arise:

Given a batch budget M and a total number of T rounds, how should the statistician determine the size of each batch, and how should she update the policy after each batch? Can the statistician design batch learning algorithms that achieve regret performances on par with the fully online setting using as few policy updates as possible?

1.1 Main contributions

In this work, we address the aforementioned questions under a classical framework for personalized decision making—nonparametric contextual bandits [48, 39]. In this framework, the expected reward associated with each treatment (or arm in the language of bandits) is modeled as a nonparametric smooth function of the covariates [59]. In the fully online setup, seminal works [48, 39] establish the minimax optimal regret bounds for the nonparametric contextual bandits. Nevertheless, under the more challenging setting with the batch constraint, the fundamental limits for nonparametric bandits remain unknown. Our paper aims to bridge this gap. More concretely, we make the following contributions:

- First, we establish a minimax regret lower bound for the nonparametric bandits with the batch constraint. Our lower bound holds even when the batch size is adaptively chosen (based on the data observed in prior batches). The proof relies on a simple but useful insight that the worst-case regret over the entire horizon is greater than the worst-case regret over the first i batches for any $1 \leq i \leq M$. To exploit this insight, for each different batch, we construct different families of hard instances to target it, leading to a maximal regret over this batch.

- In addition, we demonstrate that the aforementioned lower bound is tight by providing a matching upper bound (up to log factors). Specifically, we design a novel algorithm—Batched Successive Elimination with Dynamic Binning (BaSEDB)—for the nonparametric bandits with batch constraints. BaSEDB progressively splits the covariate space into smaller bins whose widths are carefully selected to align well with the corresponding batch size. The delicate interplay between the batch size and the bin width is crucial for obtaining the optimal regret in the batch setting.

- On the other hand, we show the suboptimality of static binning under the batch constraint by proving an algorithm-specific lower bound. Unlike the fully online setting where policies that use a fixed number of bins can attain the optimal regret [39], our lower bound indicates that batched successive elimination with static binning is strictly suboptimal.[1] This highlights the necessity of dynamic binning in some sense under the batch setting, which is uncommon in classical nonparametric estimation.

- Last but not least, we demonstrate the challenge of adapting to the margin parameter in tha batch setting. Specifically, we show that when M is small, the price of not knowing the true margin parameter for an algorithm is at least a polynomial increase in terms of the regret.

It is also worth mentioning that an immediate consequence of our results is that $M \asymp \log\log T$ number of batches sufices to achieve the optimal regret in the fully online setting. In other words, we can use a nearly constant number of policy updates in practice to achieve the optimal regret obtained by policies that require one update per round.

1.2 Related work

Nonparametric contextual bandits. [58] introduced the mathematical framework of contextual bandit. The theory of contextual bandits in the fully online setting has been continuously developed in the past few decades. On one hand, [4, 1, 23, 6, 7, 41] obtained learning guarantees for linear contextual bandits in both low and high dimensional settings. On the other hand, [59] introduced the nonparametric approach to model the mean reward function. [48] proved a minimax lower bound on the regret of nonparametric

[1] In a certain regime the BSE policy from [39] which uses a fixed number bins could loose by log factors compared to the optimal fully online regret. However, we will show the price of fixed binning is polynomial under the batch setting.

bandit and developed an upper-confidence-bound (UCB) based policy to achieve a near-optimal rate. [39] improved this result and proposed the Adaptively Binned Successive Elimination (ABSE) policy that can also adapt to the unknown margin parameter. Further insights in this nonparametric setting were developed in subsequent works [42, 43, 45, 24, 27, 52, 25, 10, 51, 9]. The smoothness assumption is also adopted in another line of work [37, 36, 33, 11] on the continuum-armed bandit problems. However in contrast to what we study, the reward is assumed to be a Lipschitz function of the action, and the covariates are not taken into considerations.

Batch learning. The batch constraint has received increasing attention in recent years. [40, 21] considered the multi-armed bandit problem under the batch setting and showed that $O(\log \log T)$ batches are adequate in achieving the rate-optimal regret, compared to the fully online setting. [26, 47] extended batch learning to the (generalized) linear contextual bandits and [46, 56, 17] further studied the setting with high-dimensional covariates. [29, 28] established batch learning guarantees for the Thompson sampling algorithm. [18] considered Lipschitz continuum-armed bandit problem with the batch constraint. Inference for batched bandits was considered in [60]. A concept related to batch learning in literature is called delayed feedback [14, 13, 55, 19]. These works consider the setting where rewards are observed with delay and analyze effects of delay on the regret. [32, 2] studied delayed feedback in nonparametric bandits and the key difference to batch learning is that the batch size is given, whereas in our case, it is a design choice by the statistician. Batch learning's focus is different to that of delayed feedback in the sense that the former gives the decision maker discretion to choose the batch size which makes it possible to approximate the optimal standard online regret with a small number of batches. Finally, the notion switching cost is intimately related to the batch constraint. [12] studied online learning with low switching cost and obtained minimax optimal regret with $O(\log \log T)$ batches. [5, 61, 20, 57, 44] developed regret guarantees with low switching cost for reinforcement learning. Low switching cost can be interpreted as infrequent policy updates, but it does not require the learner to divide the samples into batches with feedback only becoming available at the end of a batch.

2 Problem setup

We begin by introducing the problem setup for nonparametric bandits with the batch constraint.

A two-arm nonparametric bandit with horizon $T \geq 1$ is specified by a sequence of independent and identically distributed random vectors

$$(X_t, Y_t^{(1)}, Y_t^{(-1)}), \qquad \text{for } t = 1, 2, \ldots, T, \tag{1}$$

where X_t is sampled from a distribution P_X. Throughout the paper, we assume that $X_t \in \mathcal{X} = [0,1]^d$, and P_X has a density (w.r.t. the Lebesgue measure) that is bounded below and above by some constants $\underline{c}, \bar{c} > 0$, respectively. For $k \in \{1, -1\}$ and $t \geq 1$, we assume that $Y_t^{(k)} \in [0,1]$ and that

$$E[Y_t^{(k)} \mid X_t] = f^{(k)}(X_t).$$

Here $f^{(k)}$ is the unknown mean reward function for the arm k.

Without the batch constraint, the game of nonparametric bandits plays sequentially. At each step t, the statistician observes the context X_t, and pulls an action $A_t \in \{1, -1\}$ according to a rule $\pi_t : \mathcal{X} \to \{1, -1\}$. Then she receives the corresponding reward $Y_t^{(A_t)}$. In this case, the rule π_t for selecting the action at time t is allowed to depend on all the observations strictly anterior to t.

In an M-batch game, the statistician needs to design an M-batch policy (Γ, π), where $\Gamma = \{t_0, t_1, \ldots, t_M\}$ is a partition of the entire time horizon T that satisfies $0 = t_0 < t_1 < \ldots < t_{M-1} < t_M = T$, and $\pi = \{\pi_t\}_{t=1}^T$ is a sequence of random functions $\pi_t : \mathcal{X} \to \{1, -1\}$. The grid Γ can be chosen adaptively, meaning that the statistician can use all information up to t_{i-1} to determine t_i. More specifically, prior to the start of the game, she will specify the first batch t_1, and at the end of t_1, she will use all observations she have to decide the next batch t_2, and this process repeats in batches. In contrast to the case without the batch constraint, only the rewards associated with timesteps prior to the current batch are observed and available for making decisions for the current batch. Specifically, let $\Gamma(t)$ be the batch index for the time t, i.e., $\Gamma(t)$ is the unique integer such that $t_{\Gamma(t)-1} < t \leq t_{\Gamma(t)}$. Then at time t, the available information for π_t is only

$\{X_l\}_{l=1}^{t} \cup \{Y_l^{(A_l)}\}_{l=1}^{\Gamma(t)-1}$, which we denote by $\mathcal{F}_t$. The statistician's policy π_t at time t is allowed to depend on $\mathcal{F}_t$.

The goal of the statistician is to design an M-batch policy (Γ, π) that can compete with an oracle that has perfect knowledge (i.e., the law of $(X_t, Y_t^{(1)}, Y_t^{(-1)})$) of the environment. Formally, we define the cumulative regret as

$$R_T(\pi) := \mathbb{E}\left[\sum_{t=1}^{T} f^*(X_t) - f^{(\pi_t(X_t))}(X_t)\right], \tag{2}$$

where $f^*(x) := \max_{k\in\{1,-1\}} f^{(k)}(x)$ is the maximum mean reward one could obtain on the context x. Note here we omit the dependence on Γ for simplicity.

2.1 Assumptions

We adopt two standard assumptions in the nonparametric bandits literature [48, 39]. The first assumption is on the smoothness of the mean reward functions.

Assumption 1 (Smoothness). We assume that the reward function for each arm is (β, L)-smooth, that is, there exist $\beta \in (0, 1]$ and $L > 0$ such that for $k \in \{1, -1\}$,

$$|f^{(k)}(x) - f^{(k)}(x')| \le L\|x - x'\|_2^{\beta}$$

holds for all $x, x' \in \mathcal{X}$.

The second assumption is about the separation between the two reward functions.

Assumption 2 (Margin). We assume that the reward functions satisfy the margin condition with parameter $\alpha > 0$, that is there exist $\delta_0 \in (0, 1)$ and $D_0 > 0$ such that

$$P_X\left(0 < f^{(1)}(X) - f^{(-1)}(X) \le \delta\right) \le D_0\delta^{\alpha}$$

holds for all $\delta \in [0, \delta_0]$.

Assumption 2 is related to the margin condition in classification [38, 54, 3] and is introduced to bandits in [22, 48, 39]. The margin parameter affects the complexity of the problem. Intuitively, a small α, say $\alpha \approx 0$, means the two mean functions are entangled with each other in many regions and hence it is challenging to distinguish them; a large α, on the other hand, means the two reward functions are mostly well-separated.

From now on, we use $\mathcal{F}(\alpha, \beta)$ to denote the class of nonparametric bandit instances (i.e., distributions over (1)) that satisfy Assumptions 1-2.

Remark 1. Throughout the paper, we assume that $\alpha\beta \le 1$. By proposition 2.1 from [48], when $\alpha\beta > 1$, one of the arms will dominate the other one for the entire covariate space. The instance is reduced to a multi-armed bandit without covariates which is not the interest of the current paper. Therefore, we focus on the case $\alpha\beta \le 1$ hereafter.

3 Fundamental limits of batched nonparametric bandits

In this section, we establish minimax lower bounds for the regret achievable by any M-batch policy (Γ, π); see Theorem 2. To begin with, we state a minimax lower bound, together with its proof, when the grid Γ is prespecified, that is, the statistician divides the horizon $[1 : T]$ into M disjoint batches $[1 : t_1]$, $[t_1 + 1 : t_2]$, $\ldots$, $[t_{M-1} + 1, T]$ before the game begins; see Theorem 1. As we will soon see, the proof of the lower bound with fixed grid is not only useful for establishing the lower bound for any general M-batch policy (Γ, π), but also instrumental in our development of an optimal policy to be detailed in Section 4.

Recall that $\mathcal{F}(\alpha, \beta)$ denotes the class of nonparametric bandit instances (i.e., distributions over (1)) that obey Assumptions 1-2. We have the following minimax lower bound for any M-batch policy with a fixed grid, in which we define

$$\gamma := \frac{\beta(1 + \alpha)}{2\beta + d} \in (0, 1).$$

Theorem 1. Suppose that $\alpha\beta \le 1$, and assume that P_X is the uniform distribution on $\mathcal{X} = [0,1]^d$. For any M-batch policy (Γ, π) where Γ is prespecified, there exists a nonparametric bandit instance in $\mathcal{F}(\alpha, \beta)$ such that the regret of (Γ, π) on this instance is lower bounded by

$$\mathbb{E}[R_T(\pi)] \ge \tilde{D} T^{\frac{1-\gamma}{1-\gamma^M}},$$

where $\tilde{D} > 0$ is a constant independent of T and M.

See Section 3.1 for the proof of this lower bound.

As a sanity check, one sees that as M increases, the lower bound decreases. This is intuitive, as the policy is more powerful as M increases. As a result, the problem of batched nonparametric bandits becomes easier.

3.1 Proof of Theorem 1

Let (Γ, π) be the M-batch policy under consideration, with

$$\Gamma = \{t_0 = 0, t_1, t_2, \ldots, t_M = T\}.$$

Throughout this proof, we consider Bernoulli reward distributions, that is $Y_t^{(1)}, Y_t^{(-1)}$ are Bernoulli random variables with mean $f^{(1)}(X_t)$, and $f^{(-1)}(X_t)$, respectively. In addition, we fix $f^{(-1)}(x) = \frac{1}{2}$. Let f be the mean reward function of the first arm. To make the dependence on the reward instance clear, we write the cumulative regret up to time n as $R_n(\pi; f)$.

Our proof relies on a simple observation: the worst-case regret over $[T]$ is larger than the worst-case regret over the first i batches. Formally, we have

$$\sup_{(f, \frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} R_T(\pi; f) \ge \max_{1 \le i \le M} \sup_{(f, \frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} R_{t_i}(\pi; f). \tag{3}$$

Though simple, this observation lends us freedom on choosing different families of instances in $\mathcal{F}(\alpha, \beta)$ targeting different batch indices i.

Our proof consists of four steps. In Step 1, we reduce bounding the regret of a policy to lower bounding its inferior sampling rate to be defined. In Step 2, we detail the choice of different families of instances for each different batch index i. Then in Step 3, we apply an Assouad-type of argument to lower bound the average inferior sampling rate of the family of hard instances. Lastly in Step 4, we combine the arguments to complete the proof.

Step 1: Relating regret to inferior sampling rate. Given an M-batch policy, we define its inferior sampling rate at time n on an instance $(f, \frac{1}{2})$ to be

$$S_n(\pi; f) := \mathbb{E}\left[\sum_{t=1}^{n} 1\{\pi_t(X_t) = \pi^{\boxtimes}(X_t), f(X_t) = \frac{1}{2}\}\right].$$

In words, $S_n(\pi; f)$ counts the number of times π selects the strictly suboptimal arm up to time n. Thanks to the following lemma, we can reduce lower bounding the regret to the inferior sampling rate.

Lemma 1 (Lemma 3.1 in [48]). Suppose that $(f, \frac{1}{2}) \in \mathcal{F}(\alpha, \beta)$. Then for any $1 \le n \le T$, we have

$$S_n(\pi; f) \le D n^{\frac{1}{1+\alpha}} R_n(\pi; f)^{\frac{\alpha}{1+\alpha}},$$

for some constant $D > 0$.

As an immediate consequence of the above lemma, we obtain

$$\begin{aligned}\sup_{(f, \frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} R_T(\pi; f) &\ge \max_{1 \le i \le M} \sup_{(f, \frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} \left(\frac{1}{D}\right)^{\frac{1+\alpha}{\alpha}} t_i^{-\frac{1}{\alpha}} (S_{t_i}(\pi; f))^{\frac{1+\alpha}{\alpha}} \\ &= \left(\frac{1}{D}\right)^{\frac{1+\alpha}{\alpha}} \max_{1 \le i \le M} t_i^{-\frac{1}{\alpha}} \left[\sup_{(f, \frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} S_{t_i}(\pi; f)\right]^{\frac{1+\alpha}{\alpha}}.\end{aligned}$$

From now on, we focus on lower bounding $\sup_{(f, \frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} S_{t_i}(\pi; f)$.

Step 2: Introducing the family of reward instances for t_i. Our construction of the family of hard instances is adapted from [48]. Define $z_1 = 1$, and $z_i = \lceil t_{i-1}^{1/(2\beta+d)} \rceil$ for $i = 2, 3, \ldots, M$. Henceforth, we will fix some i and write z_i as z. We partition $[0,1]^d$ into z^d bins with equal width. Denote the bins by C_j for $j = 1, \ldots, z^d$, and let q_j be the center of C_j.

Define a set of binary sequences $\Omega_s := \{\pm 1\}^s$, with $s := \lceil z^{d-\alpha\beta} \rceil$. For each $\omega \in \Omega_s$ we define a function $f_\omega : [0,1]^d \to \mathbb{R}$:

$$f_\omega(x) = \frac{1}{2} + \sum_{j=1}^{s} \omega_j \phi_j(x),$$

where $\phi_j(x) = D_\varphi z^{-\beta} \varphi(2z(x - q_j)) 1\{x \in C_j\}$ with $\varphi(x) = (1 - \|x\|_\infty)^\beta 1\{\|x\|_\infty \le 1\}$, and $D_\varphi = \min(2^{-\beta} L, 1/4)$. In all, we consider the family of reward instances

$$\mathcal{C}_z := \left\{ f^{(1)}(x) = f_\omega(x), f^{(-1)}(x) = \frac{1}{2} \;\middle|\; \omega \in \Omega_s \right\}. \tag{4}$$

With slight abuse of notation, we also use $\mathcal{C}_z$ to denote $\{f_\omega : \omega \in \Omega_s\}$. It is straightforward to check that $\mathcal{C}_z \subseteq \mathcal{F}(\alpha, \beta)$.

Step 3: Lower bounding the inferior sampling rate. Fix some $i \in [M]$, and consider $z = z_i$. Since $\mathcal{C}_z \subseteq \mathcal{F}(\alpha, \beta)$, we have

$$\sup_{(f, \frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} S_{t_i}(\pi; f) \ge \sup_{f \in \mathcal{C}_z} S_{t_i}(\pi; f).$$

Using the definitions of $\mathcal{C}_z$ and $S_{t_i}(\pi; f)$, we have

$$\begin{aligned}
\sup_{f \in \mathcal{C}_z} S_{t_i}(\pi; f) &= \sup_{\omega \in \Omega_s} E_{\pi, f_\omega} \left[\sum_{t=1}^{t_i} 1\{\pi_t(X_t) = \mathrm{sign}(f_\omega(X_t) - \tfrac{1}{2}), f_\omega(X_t) = \tfrac{1}{2}\} \right] \\
&\ge \frac{1}{2^s} \sum_{\omega \in \Omega_s} E_{\pi, f_\omega} \left[\sum_{t=1}^{t_i} 1\{\pi_t(X_t) = \mathrm{sign}(f_\omega(X_t) - \tfrac{1}{2}), f_\omega(X_t) = \tfrac{1}{2}\} \right].
\end{aligned}$$

Since $f_\omega(x) = \frac{1}{2}$ for $x \notin \bigcup_{j=1,\ldots,s} C_j$, we further obtain

$$\sup_{f \in \mathcal{C}_z} S_{t_i}(\pi; f) \ge \frac{1}{2^s} \sum_{\omega \in \Omega_s} \sum_{t=1}^{t_i} \sum_{j=1}^{s} E^t_{\pi, f_\omega} [1\{\pi_t(X_t) = \omega_j, X_t \in C_j\}]. \tag{5}$$

Here we use P^t_{π, f_ω} to denote the joint distribution of $\{X_l\}_{l=1}^t \cup \{Y_l^{\pi_l(X_l)}\}_{l=1}^{\Gamma(t)-1}$, where $\Gamma(t)$ is the batch index for t, i.e., the unique integer such that $t_{\Gamma(t)-1} < t \le t_{\Gamma(t)}$. We use E^t_{π, f_ω} to denote the corresponding expectation. Expand the right hand side of (5) to see that

$$\sup_{f \in \mathcal{C}_z} S_{t_i}(\pi; f) \ge \frac{1}{2^s} \sum_{j=1}^{s} \sum_{t=1}^{t_i} \sum_{\omega_{[-j]} \in \Omega_{-1}} \sum_{h \in \{\pm 1\}} E^t_{\pi, f_{\omega^h_{[-j]}}} [1\{\pi_t(X_t) = h, X_t \in C_j\}], \tag{6}$$

$$\text{where } W_{j,t,\omega_{[-j]}} \text{ denotes } \sum_{h \in \{\pm 1\}} E^t_{\pi, f_{\omega^h_{[-j]}}} [1\{\pi_t(X_t) = h, X_t \in C_j\}]$$

where $\omega^h_{[-j]}$ is the same as ω except for the j-th entry being h. Note that here we use the fact that for $f_{\omega^h_{[-j]}}$, the optimal arm in the bin C_j is h. We then relate $W_{j,t,\omega_{[-j]}}$ to a binary testing error,

$$\begin{aligned}
W_{j,t,\omega_{[-j]}} &= \frac{1}{z^d} \sum_{h \in \{\pm 1\}} P^t_{\pi, f_{\omega^h_{[-j]}}} (\pi_t(X_t) = h \mid X_t \in C_j) \\
&\ge \frac{1}{4z^d} \exp\left(-KL(P^t_{\pi, f_{\omega^{-1}_{[-j]}}}, P^t_{\pi, f_{\omega^1_{[-j]}}}) \right), \qquad (7)
\end{aligned}$$

where the second step invokes Le Cam's method. Under the batch setting, at time t, the available information is only up to $t_{\Gamma(t)-1}$. Consequently, we can apply Lemma 5 to obtain

$$KL(P^{t}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{t}_{\pi,f_{\omega^{1}_{[-j]}}}) = KL(P^{t_{\Gamma(t)-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{t_{\Gamma(t)-1}}_{\pi,f_{\omega^{1}_{[-j]}}}) \le 2z^{-(2\beta+d)}t_{\Gamma(t)-1}. \tag{8}$$

Combining (6), (7), and (8), we arrive at

$$\begin{aligned}
\sup_{f\in C_z} S_{t_i}(\pi; f) &\ge \frac{1}{8}\sum_{j=1}^{s}\sum_{t=1}^{t_i}\frac{1}{z^d}\exp\left(-2z^{-(2\beta+d)}t_{\Gamma(t)-1}\right) \\
&\ge \frac{1}{8}\sum_{l=1}^{z^{d-\alpha\beta}}\sum_{t=t_{l-1}+1}^{t_i}\frac{1}{z^d}\exp\left(-2z^{-(2\beta+d)}t_{l-1}\right)\mathbb{1}_{j=1} \\
&\ge \frac{1}{8}\sum_{l=1}^{z^{d-\alpha\beta}}\sum_{t=t_{l-1}+1}^{t_i}\frac{1}{z^d}\exp\left(-2z^{-(2\beta+d)}t_{i-1}\right)\mathbb{1}_{j=1}
\end{aligned}$$

where the second line uses the fact that $s = \lfloor z^{d-\alpha\beta}\rfloor$, and the last inequality holds since $t_{l-1} \le t_{i-1}$ for all $1 \le l \le i$. Now recall that $z = z_i = \lceil (t_{i-1})^{1/(2\beta+d)}\rceil$ for $i \ge 1$, and $z = 1$ for $i = 1$. We can continue the lower bound to see that

$$\begin{aligned}
\sup_{f\in C_{z_i}} S_{t_i}(\pi; f) &\ge \frac{1}{8}\sum_{j=1}^{z^{d-\alpha\beta}}\sum_{l=1}^{i}\frac{t_l - t_{l-1}}{z^d}\exp\left(-2z^{-(2\beta+d)}t_{i-1}\right) \\
&\ge c^{\natural}\sum_{l=1}^{z^{d-\alpha\beta}}\sum_{t=t_{l-1}+1}^{t_i}\frac{1}{z^d}\mathbb{1}_{j=1} \\
c\cdot\frac{t_i}{z^{\alpha\beta}} &= \begin{cases} c^{\natural}\cdot\frac{t_i}{t_{i-1}^{\frac{\alpha\beta}{2\beta+d}}}, & i > 1 \\ c^{\natural}t_1, & i = 1 \end{cases},
\end{aligned}$$

for some $c^{\natural} > 0$.

Step 4: Combining bounds together. Combining the previous arguments together leads to the conclusion that

$$\begin{aligned}
\sup_{(f,\frac{1}{2})\in F(\alpha,\beta)} R_T(\pi; f) &\ge \max_{1\le i\le M}\sup_{f\in C_{z_i}} R_{t_i}(\pi; f) \\
&\ge \left(\frac{1}{D}\right)^{\frac{1+\alpha}{\alpha}}\max_{1\le i\le M} t_i^{-\frac{1}{\alpha}}\sup_{f\in C_{z_i}} S_{t_i}(\pi; f)\Bigg)^{\frac{1+\alpha}{\alpha}} \\
&\gtrsim \max\left(t_1, \frac{t_2}{t_1^{\gamma}}, \ldots, \frac{T}{t_{M-1}^{\gamma}}\right) \\
&\ge \tilde{D}T^{\frac{1-\gamma}{1-\gamma^M}}.
\end{aligned} \tag{9}$$

This finishes the proof.

3.2 Lower bound for general M-batch policy

Now we are ready to state the minimax lower bound for any general M-batch policy (Γ, π), i.e., when the grid Γ is allowed to be adaptively chosen.

Theorem 2. Suppose that $\alpha\beta \le 1$, and assume that P_X is the uniform distribution on $\mathcal{X} = [0,1]^d$. For any M-batch policy (Γ, π), there exists a nonparametric bandit instance in $\mathcal{F}(\alpha, \beta)$ such that the regret of π on this instance is lower bounded by

$$\mathbb{E}[R_T(\pi)] \ge \tilde{D}_1 \left(\frac{1}{M}\right)^{\tilde{D}_2} T^{\frac{1-\gamma}{1-\gamma^M}},$$

where $\tilde{D}_1, \tilde{D}_2 > 0$ are constants independent of T and M.

See Appendix A for the proof.

Since our focus is on $M > \log\log T$ (when $M \asymp \log\log T$, by Corollary 1 there exists an algorithm whose regret attains the optimal fully online regret), we can see Theorem 1 and Theorem 2 differ at most by poly-log factors in T.

Unlike the fixed grid case where we choose a specific family of hard instances to target the regret in a certain batch, we cannot directly do so when the grid is adaptively selected because the adversary does not know $\{t_i\}_{i=1}^M$ in advance. Inspired by [21], we overcome this dificulty by using an appropriately defined bad event that happens with suficient probability to reduce the adaptive case to the fixed grid case. However, the proof under the nonparametric setting is much more challenging because the presence of contexts makes the instances for different batches less indistinguishable with each other. A key ingredient of our proof is to establish tight control of the total variation distance between two mixture distributions. The full proof can be found in Appendix A.

3.3 Implications on design of optimal M-batch policy

As we have mentioned, the proof of the lower bound with fixed grid, i.e., Theorem 1 facilitates the design of optimal M-batch policy.

Grid selection. First, the lower bound of the whole horizon is reduced to the worst-case regret over a specific batch; see (3). Consequently, we need to design the grid $\Gamma = (t_0, t_1, t_2, \ldots, t_{M-1}, t_M)$ such that the total regret is evenly distributed across batches. More concretely, in view of the lower bound (9), one needs to set $t_i \asymp_{1-\gamma} \left(\frac{T}{t_{i-1}}\right)^{1-\gamma^{-\frac{1}{M}}}$ for $2 \le i \le M$.

Dynamic binning. In addition, in the proof of the lower bound, for each different batch i, we use different families of hard reward instances, parameterized by the number of bins $z_i = \lceil t_{i-1}^{1/(2\beta+d)} \rceil$. In other words, from the lower bound perspective, the granularity (i.e., the bin width $1/z_i$) at which we investigate the mean reward function depends crucially on the grid points $\{t_i\}$: the larger the grid point t_i, the finer the granularity. This key observation motivates us to consider the batched successive elimination with dynamic binning algorithm to be introduced below.

4 Batched successive elimination with dynamic binning

In this section, we present the batched successive elimination with dynamic binning policy (BaSEDB) that nearly attains the minimax lower bound, up to log factors; see Algorithm 1. On a high level, Algorithm 1 gradually partitions the covariate space $\mathcal{X}$ into smaller hypercubes (i.e., bins) throughout the batches based on a list of carefully chosen cube widths, and reduces the nonparametric bandit in each cube to a bandit problem without covariates.

A tree-based interpretation. The process is best illustrated with the notion of a tree $\mathcal{T}$ of depth M; see Figure 1. Each layer of of the tree $\mathcal{T}$ is a set of bins that form a regular partition of $\mathcal{X}$ using hypercubes with equal widths. And the common width of the bins B_i in layer i is dictated by a list $\{g_i\}_{i=0}^{M-1}$ of split factors. More precisely, we let

$$w_i := \left(\prod_{l=0}^{i-1} g_l\right)^{-1} \tag{10}$$

be the width of the cubes in the i-th layer B_i for $i \ge 1$, and $w_0 = 1$. In other words, B_i contains all the cubes

Algorithm 1 Batched successive elimination with dynamic binning (BaSEDB)

Input: Batch size M, grid $\Gamma = \{t_i\}_{i=0}^{M}$, split factors $\{g_i\}_{i=0}^{M-1}$.
$L \leftarrow B_1$
for $C \in L$ **do**
    $I_C = I$
for $i = 1, \ldots, M-1$ **do**
    for $t = t_{i-1}+1, \ldots, t_i$ **do**
        $C \leftarrow L(X_t)$
        Pull an arm from I_C in a round-robin way.
        if $t = t_i$ **then**
            Update L and $\{I_C\}_{C \in L}$ by Algorithm 2 $(L, \{I_C\}_{C \in L}, i, g_i)$.
for $t = t_{M-1}+1, \ldots, T$ **do**
    $C \leftarrow L(X_t)$
    Pull any arm from I_C.

$$C_{i,v} = \{x \in X : (v_j - 1)w_i \le x_j < v_j w_i, 1 \le j \le d\},$$

where $v = (v_1, v_2, \ldots, v_d) \in [\frac{1}{w_i}]^d$. As a result, there are in total $(\frac{1}{w_i})^d$ bins in B_i.

Algorithm 1 proceeds in batches and maintains two key objects: (1) a list L of active bins, and (2) the corresponding active arms I_C for each $C \in L$; see Figure 1 for an example. Specifically, prior to the game (i.e., prior to the first batch), L is set to be B_1, all bins in layer 1, and $I_C = \{1, -1\}$ for all $C \in L$. Within this batch, the statistician tries the arms in I_C equally likely for all bins in L. Then at the end of the batch, given the revealed rewards in this batch, we update the active arms I_C for each $C \in L$ via successive elimination. If no arm were eliminated from I_C, this suggests that the current bin is not fine enough for the statistician to tell the difference between the two arms. As a result, she splits the bin $C \in L$ into its children child(C) in T. All the child nodes will be included in L, while the parent C stops being active (i.e., C is removed from L). The whole process repeats in a batch fashion. [2]

Grid Γ and split factors $\{g_i\}_{i=0}^{M-1}$. As one can see, the split factor g_i controls how many children a node at layer i can have and its appropriate choice is crucial for obtaining small regret. Intuitively, g_i should be selected in a way such that a node C_{i+1} with width w_i can fully leverage the number of samples allocated to it during the $(i+1)$-th batch. With these goals in mind, we design the grid $\Gamma = \{t_i\}$ and split factors $\{g_i\}$ as follows. Recall that $\gamma = \frac{\beta(1+\alpha)}{2\beta+d}$. We set

$$b = \Theta T^{\frac{1-\gamma}{1-\gamma^M}}.$$

The split factors are chosen according to

$$g_0 = \lceil b^{\frac{1}{2\beta+d}} \rceil, \qquad \text{and} \qquad g_i = \lceil g_{i-1}^{\gamma} \rceil, \; i = 1, \ldots, M-2. \tag{11}$$

In addition, the grid is chosen such that

$$t_i - t_{i-1} = \lceil l_i w_i^{-(2\beta+d)} \log(T w_i^d) \rceil, 1 \le i \le M-1, \tag{12}$$

where $l_i > 0$ is a constant to be specified later. It is easy to check that with these choices, we have

$$t_1 \in T^{\frac{1-\gamma}{1-\gamma^M}}, \qquad \text{and} \qquad t_i = \lceil b(t_{i-1})^{\gamma} \rceil, \quad \text{for } i = 2, \ldots, M.$$

[2] For the final batch M, the split factor $g_{M-1} = 1$ by default because there is no need to further partition the nodes for estimation.

Algorithm 2 Tree growing subroutine

Input: Active nodes L, active arm sets $\{I_C\}_{C \in L}$, batch number i, split factor g_i.
$L' \leftarrow \{\}$
for $C \in L$ do
  if $|I_C| = 1$ then
    $L' \leftarrow L' \cup \{C\}$
    Proceed to next C in the iteration.
  $\bar{Y}_{C,i}^{\max} \leftarrow \max_{k \in I_C} \bar{Y}_{C,i}^{(k)}$
  for $k \in I_C$ do
    if $\bar{Y}_{C,i}^{\max} - \bar{Y}_{C,i}^{(k)} > U(m_{C,i}, T, C)$ then $I_C \leftarrow I_C - \{k\}$
  if $|I_C| > 1$ then
    $I_{C'} \leftarrow I_C$ for $C' \in \text{child}(C, g_i)$
    $L' \leftarrow L' \cup \text{child}(C, g_i)$
  else
    $L' \leftarrow L' \cup \{C\}$
Return L'

In particular, we set b properly to make $t_M = T$. Indeed, these choices taken together meet the expectation laid out in Section 3.3: we need to choose the grid and the split factors appropriately so that (1) the total regret spreads out across different batches, and (2) the granularity becomes finer as we move further to later batches.

When to eliminate arms? Now we zoom in on the elimination process described in Algorithm 2. The basic idea follows from successive elimination in the bandit literature [16, 39, 21]: the statistician eliminates an arm from I_C if she expects the arm to be suboptimal in the bin C given the rewards collected in C. Specifically, for any node $C \in \mathcal{T}$, define

$$U(\tau, T, C) := 4 \sqrt{\frac{\log(2T|C|^d)}{\tau}},$$

where $|C|$ denotes the width of the bin. Let $m_{C,i} := \sum_{t=t_{i-1}+1}^{t_i} 1\{X_t \in C\}$ be the number of times we observe contexts from C in batch i. We then define for $k \in \{1, \dots, K\}$ that

$$\bar{Y}_{C,i}^{(k)} := \frac{\sum_{t=t_{i-1}+1}^{t_i} Y_t \cdot 1\{X_t \in C, A_t = k\}}{\sum_{t=t_{i-1}+1}^{t_i} 1\{X_t \in C, A_t = k\}},$$

which is the empirical mean reward of arm k in node C during the i-th batch. It is easy to check that $\bar{Y}_{C,i}^{(k)}$ has expectation $\bar{f}_C^{(k)}$ given by

$$\bar{f}_C^{(k)} := \mathbb{E}[f^{(k)}(X) \mid X \in C] = \frac{1}{P_X(C)} \int_C f^{(k)}(x) dP_X(x).$$

Similarly, we define the average optimal reward in bin C to be

$$\bar{f}_C^{*} := \frac{1}{P_X(C)} \int_C f^{*}(x) dP_X(x).$$

The elimination threshold $U(m_{C,i}, T, C)$ is chosen such that an arm k with $f^* - \bar{f}_C^{(k)} \gtrsim |C|^\beta$ is eliminated with high probability at the end of batch i. Therefore, when $|I_C| > 1$, the remaining arms are statistically

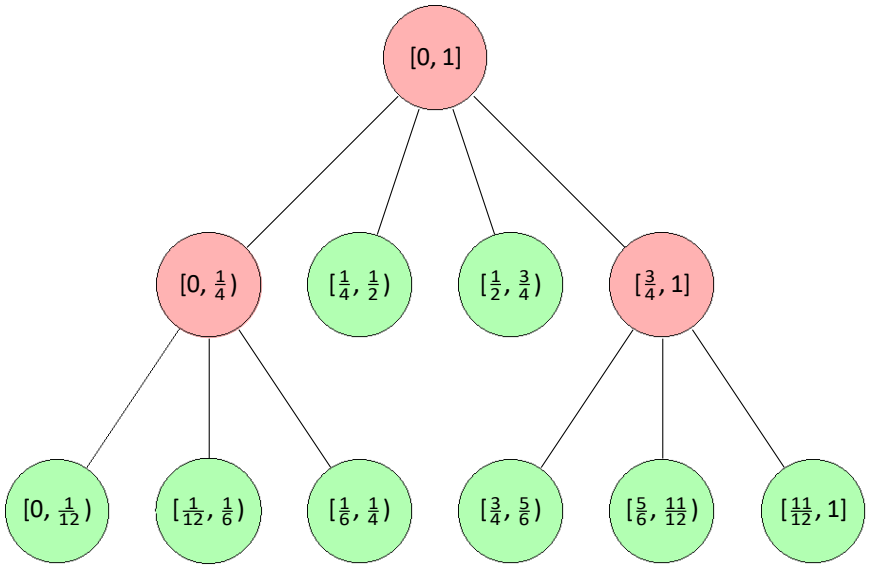

Figure 1: An example of the tree growing process for $d = 1, M = 3, G = \{4, 3, 1\}$. The root node is at depth 0. For the first batch, the 4 nodes located at depth 1 of the tree were used. Both $[\frac{1}{4}, \frac{1}{2})$ and $[\frac{1}{2}, \frac{3}{4})$ only had one active arm remaining so they were not further split and remained in the set of active nodes (green). Meanwhile, $|I_{[0,\frac{1}{4})}| = |I_{[\frac{3}{4},1]}| = 2$ so each of them was split into 3 smaller nodes, and both nodes were marked as inactive (red). For the second batch, all the green nodes were actively used but arm elimination was performed at the end of batch 2 only for nodes located at depth 2 (the green nodes at depth 1 already have 1 active arm remaining so there is no need to eliminate again).

indistinguishable from each other, so C is split into smaller nodes to estimate those arms more accurately using samples from future batches. On the other hand, when $|I_C| = 1$, the remaining arm is optimal in C with high probability—a consequence of the smoothness condition, and it will be exploited in the later batches.

Connections and differences with ABSE in [39]. In appearance, BaSEDB (Algorithm 1) looks quite similar to the Adaptively Binned Successive Elimination (ABSE) proposed in [39]. However, we would like to emphasize several fundamental differences. First, the motivations for the algorithms are completely different. [39] designs ABSE to adapt to the unknown margin condition α, while our focus is to tackle the batch constraint. In fact, without the batch constraints, if α is known, adaptive binning is not needed to achieve the optimal regret [39]. This is certainly not the case in the batched setting. Fixing the number of bins used across different batches is suboptimal because one can construct instances that cause the regret incurred during a certain batch to explode. We will expand on this phenomenon in Section 4.3. Secondly, the algorithm in [39] partitions a bin into a fixed number 2^d of smaller ones once the original bin is unable to distinguish the remaining arms. In this way, the algorithm can adapt to the difference in the local dificulty of the problem. In comparison, one of our main contributions is to carefully design the list of varying split factors that allows the new cubes to maximally utilize the number of samples allocated to it during the next batch.

4.1 Regret guarantees

Now we are ready to present the regret performance of BaSEDB (Algorithm 1).

Theorem 3. Suppose that $\alpha\beta \leq 1$. Fix any constant $D_1 > 0$ and suppose that $M \leq D_1 \log T$. Equipped with the grid and split factors list that satisfy (12) and (11), the policy $\hat{\pi}$ given by Algorithm 1 obeys

$$E[R_T(\hat{\pi})] \leq \tilde{C}(\log T)^2 \cdot T^{\frac{1-\gamma}{1-\gamma^M}},$$

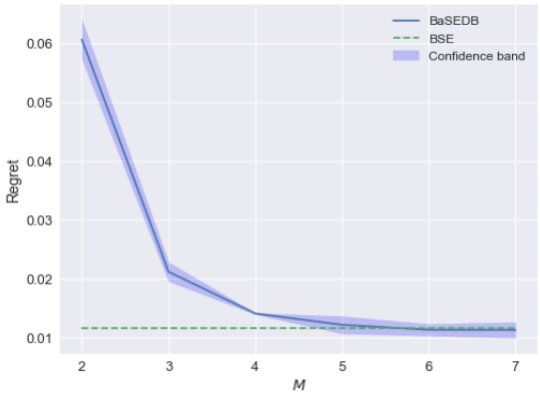

Figure 2: Regret vs. batch budget M.

where $\tilde{C} > 0$ is a constant independent of T and M.

See Appendix B for the proof.

While Theorem 3 requires $M \gtrsim \log T$, we see from the corollary below that it is in fact suficient to show the optimality of Algorithm 1.

Corollary 1. As long as $M \geq D_2 \log\log(T)$, where D_2 depends on $\gamma = \frac{\beta(1+\alpha)}{2\beta+d}$, Algorithm 1 achieves

$$E[R_T(\hat{\pi})] \leq \tilde{C}(\log T)^2 \cdot T^{1-\gamma},$$

where $\tilde{C} > 0$ is a constant independent of T and M.

Theorem 3, together with Corollary 1 and Theorem 2 establish the fundamental limits of batch learning for the nonparametric bandits with covariates, as well as the optimality of BaSEDB, up to logarithmic factors. To see this, when $M \gtrsim \log\log(T)$, the upper bound in Theorem 3 matches the lower bounds in Theorem 1 and Theorem 2, apart from log factors. On the other end, when $M \gtrsim \log\log(T)$, Algorithm 1, while splitting the horizon into M batches, achieves the optimal regret (up to log factors) for the setting without the batch constraint [39]. It is evident that Algorithm 1 is optimal in this case.

4.2 Numerical experiments

In this section, we provide some experiments on the empirical performance of Algorithm 1. We set $T = 50000, d = \beta = 1, \alpha = 0.2$. We let P_X be the uniform distribution on $[0,1]$. Denote $q_j = (j - 1/2)/4$ and $C_j = [q_j - 1/8, q_j + 1/8]$ for $1 \leq j \leq 4$. For the mean reward functions, we choose $f^{(1)}, f^{(-1)} : [0,1] \to R$ such that

$$f^{(1)}(x) = \frac{1}{2} + \sum_{j=1}^{4} \omega_j \phi_j(x), \qquad f^{(-1)}(x) = \frac{1}{2},$$

where $\omega_j's$ are sampled i.i.d. from $Rad(\frac{1}{2})$, $\phi_j(x) = \frac{1}{4}\phi(8(x - q_j))1\{x \in C_j\}$ and $\varphi(x) = (1 - |x|)1\{|x| \leq 1\}$. We let $Y^{(k)} \sim Bernoulli(f^{(k)}(x))$. To illustrate the performance of Algorithm 1, we compare it with the Binned Successive Elimination (BSE) policy from [39], which is shown to be minimax optimal in the fully online case. Figure 2 shows the regret of Algorithm 1 under different batch budegts. One can see that it is suficient to have $M = 5$ batches to achieve the fully online eficiency.

4.3 Failure of static binning

We have seen the power of dynamic binning in solving batched nonparametric bandits by establishing its rate-optimality in minimizing regret. Now we turn to a complimentary but intriguing question: is it necessary to use dynamic binning to achieve optimal regret under the batch constraint? To formally address this

question, we investigate the performance of successive elimination with static binning, i.e., Algorithm 1 with $g_0 = g$, and $g_1 = g_2 = \cdots g_{M-2} = 1$. Although static binning works when M is large (e.g., a single choice of g attains the optimal regret [48, 39] in the fully online setting), we show that it must fail when M is small.

To bring the failure mode of static binning into focus, we consider the simplest scenario when $M = 3$, and $\alpha = \beta = d = 1$. Note that the successive elimination with static binning algorithm is parameterized by the grid choice $\Gamma = \{t_0 = 0, t_1, t_2, t_3 = T\}$ and the fixed number g of bins. The following theorem formalizes the failure of static binning in achieving optimal regret when $M = 3$.

Theorem 4. Consider $M = 3$, and $\alpha = \beta = d = 1$. For any choice of $1 \le t_1 < t_2 \le T - 1$, and any choice of g, there exists a nonparametric bandit instance in $F(1, 1)$ such that the resulting successive elimination with static binning algorithm $\hat{\pi}_{\text{static}}$ satisfies

$$E[R_T(\hat{\pi}_{\text{static}})] \ge \tilde{C}_1 T^{\frac{9}{19}+\kappa},$$

for some $\kappa, \tilde{C}_1 > 0$ that are independent of T. Here $T^{\frac{9}{19}}$ is the optimal regret achieved by BaSEDB—an successive elimination algorithm with dynamic binning.

While the formal proof is deferred to Appendix C, we would like to immediately point out the intuition underlying the failure of static binning.

Necessary choice of grid Γ. It is evident from the proof of the minimax lower bound (Theorem 1) that one needs to set $t_1 \asymp T^{9/19}$, and $t_2 \asymp T^{15/19}$. Otherwise, the inequality (9) guarantees the worst-case regret of $\hat{\pi}_{\text{static}}$ exceeds the optimal one $T^{\frac{9}{19}}$. Consequently, we can focus on the algorithm with $t_1 \asymp T^{9/19}$, $t_2 \asymp T^{15/19}$, and only consider the design choice g.

Why fixed g fails. As a baseline for comparison, recall that in the optimal algorithm with dynamic binning, we set $g_0 \asymp T^{3/19}$, and $g_0 g_1 \asymp T^{5/19}$ so that the worst case regret in three batches are all on the order of $T^{\frac{9}{19}}$. In view of this, we split the choice of g into three cases.

- Suppose that $g \asymp T^{3/19}$. In this case, we can construct an instance such that the reward difference only appears on an interval with length $1/z \asymp 1/g$; see Figure 3. In other words, the static binning is finer than that in the reward instance. As a result, the number of pulls in the smaller bin (used by the algorithm) in the first batch is not suficient to tell the two arms apart, that is with constant probability, arm elimination will not happen after the first batch. This necessarily yields the blowup of the regret in the second batch.
- Suppose that $g \asymp T^{3/19}$. In this case, we can construct an instance such that the reward difference only appears on an interval with length $1/z \asymp 1/g$; see Figure 4. In other words, the static binning is coarser than that in the reward instance. Since the aggregated reward difference on the larger bin is so small, the number of pulls in the larger bin (used by the algorithm) in the first batch is still not suficient to result in successful arm elimination. Again, the regret on the second batch blows up.
- Suppose that $g \asymp T^{3/19}$. Since this choices matches g_0 used in the optimal dynamic binning algorithm, there is no reward instance that can blow up the regret in the first two batches. Nevertheless, since $g \asymp g_0 g_1 \asymp T^{5/19}$, one can construct the instance similar to the previous case (i.e., Figure 4) such that the regret on the third batch blows up.

5 Adaptivity to margin parameter α

In this section, we provide some discussions on the possibility of adapting to the margin parameter α if it is unknown. Recall in Section 4, the grid choice of Algorithm 1 requires knowledge of α. One may ask if such knowledge is essential in obtaining small regret. Unfortunately, the following theorem demonstrates that the price of not knowing α is at least a polynomial increase in regret.

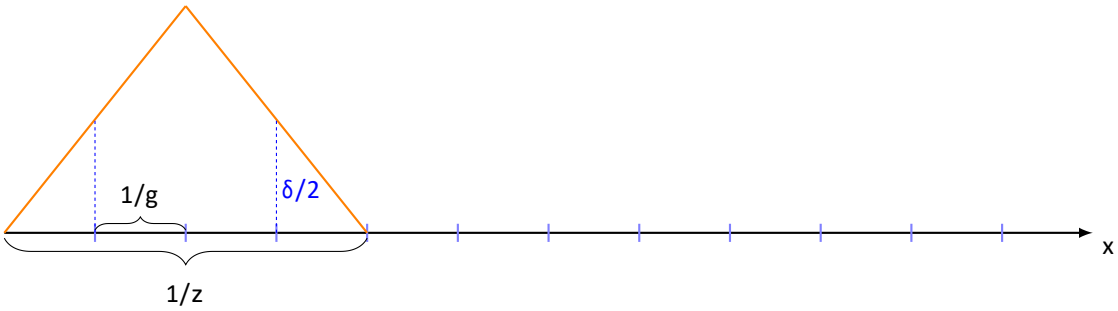

Figure 3: Instance with $g > z$. Each bin B produced by $\hat{\pi}_{\text{static}}$ has width $1/g$.

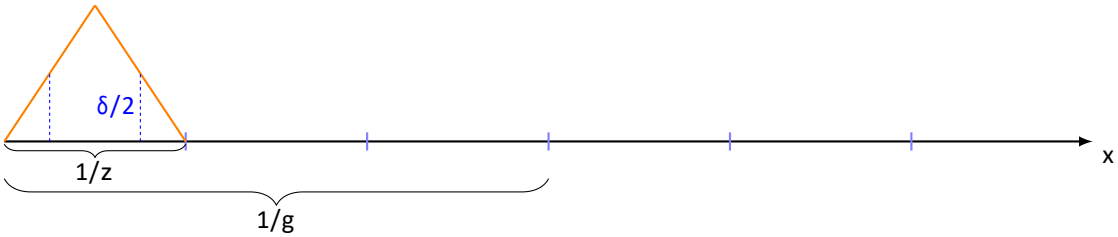

Figure 4: Instance with $g < z$. Each bin B produced by $\hat{\pi}_{\text{static}}$ has width $1/g$.

Theorem 5. Consider $M = 2$ (or 3) and $\beta = d = 1$. For any algorithm that does not know the true margin parameter α, there exists a choice of α such that

$$\sup_{+\kappa_1,\, F(\alpha,1)} E[R_T(\pi)] \geq \tilde{D}_3 T^{\frac{1-\frac{\alpha+1}{\alpha+2}}{1-(\frac{\alpha+1}{3})^M}}$$

for some $\tilde{D}_3 > 0, \kappa_1$ that are independent of T.

See Appendix D for the proof.

Theorem 5 says for any algorithm that does not have knowledge of α, its regret is at least a polynomial factor larger than the optimal regret attained by Algorithm 1. This result shows batch learning for nonparametric bandits is much harder than the fully online case to some extent, where adaptivity to α could be achieved for free [39].

The intuition behind the proof of Theorem 5 is that since the algorithm does not know α, it has little hope to pick the first batch size t_1 optimally. If t_1 is too large, then the adversary can choose a big α, which corresponds to the family of reward functions with larger gaps, so that the algorithm's regret during the first batch explodes. If t_1 is too small, then the adversary can choose a little α, which corresponds to the family of reward functions with smaller gaps, so that the algorithm's knowledge gathered during the first batch is not enough to distinguish the arms and its regret will explode in later batches.

6 Conclusions

In this paper, we characterize the fundamental limits of batch learning in nonparametric contextual bandits. In particular, our optimal batch learning algorithm (i.e., Algorithm 1) is able to match the optimal regret in the fully online setting with only $O(\log \log T)$ policy updates. Our work open a few interesting avenues to explore in the future.

Extensions to multiple arms. With slight modification, our algorithm works for nonparametric contextual bandits with more than two arms. However, it remains unclear what the fundamental limits of batch learning are in this multi-armed case (i.e., when K is large).

Improving the log factor. Comparing the upper and lower bounds, it is evident that Algorithm 1 is near-optimal up to log factors. It is certainly interesting to improve this log factor, either by strengthening the lower bound, or making the upper bound more eficient.

Adapting to margin parameters. While we have shown that adaptivity to the margin parameter is not possible when the batch constraint is stringent, i.e., when $M = 2$ (or 3), it leaves open the question of designing optimal adaptive algorithm when M is large, say $M \asymp \log\log T$. In fact, when $M = T$, i.e., in the fully online setting, [39] provides an adaptively binned successive elimination algorithm that is capable of adapting to the margin parameter optimally.

Acknowledgements

CM is partially supported by the National Science Foundation via grant DMS-2311127.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 24, 2011.

[2] Sakshi Arya and Yuhong Yang. Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards. Statistics & Probability Letters, 164:108818, 2020.

[3] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. The Annals of Statistics, 35(2):608–633, 2007.

[4] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3(Nov):397–422, 2002.

[5] Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably eficient q-learning with low switching cost. Advances in Neural Information Processing Systems, 32, 2019.

[6] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. Operations Research, 68(1):276–294, 2020.

[7] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. Management Science, 67(3):1329–1349, 2021.

[8] Dimitris Bertsimas and Adam J Mersereau. A learning approach for interactive marketing to a customer segment. Operations Research, 55(6):1120–1135, 2007.

[9] Moise Blanchard, Steve Hanneke, and Patrick Jaillet. Non-stationary contextual bandits and universal learning. arXiv preprint arXiv:2302.07186, 2023.

[10] Changxiao Cai, T Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. arXiv preprint arXiv:2211.12612, 2022.

[11] T Tony Cai and Hongming Pu. Stochastic continuum-armed bandits with additive models: Minimax regrets and adaptive algorithm. The Annals of Statistics, 50(4):2179–2204, 2022.

[12] Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. Advances in Neural Information Processing Systems, 26, 2013.

[13] Olivier Chapelle. Modeling delayed feedback in display advertising. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1097–1105, 2014.

[14] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. Advances in neural information processing systems, 24, 2011.

[15] Stephen E Chick and Noah Gans. Economic analysis of simulation selection problems. Management Science, 55(3):421–437, 2009.

[16] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. Journal of machine learning research, 7(6):1079–1105, 2006.

[17] Jianqing Fan, Zhaoran Wang, Zhuoran Yang, and Chenlu Ye. Provably eficient high-dimensional bandit learning with batched feedbacks. arXiv preprint arXiv:2311.13180, 2023.

[18] Yasong Feng, Zengfeng Huang, and Tianyu Wang. Lipschitz bandits with batched feedback. Advances in Neural Information Processing Systems, 35:19836–19848, 2022.

[19] Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In International Conference on Machine Learning, pages 3348–3356. PMLR, 2020.

[20] Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably eficient algorithm for linear markov decision process with low switching cost. arXiv preprint arXiv:2101.00494, 2021.

[21] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. Advances in Neural Information Processing Systems, 32, 2019.

[22] Alexander Goldenshluger and Assaf Zeevi. Woodroofe's one-armed bandit problem revisited. The Annals of Applied Probability, 19(4):1603–1633, 2009.

[23] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. Stochastic Systems, 3(1):230–261, 2013.

[24] Melody Guan and Heinrich Jiang. Nonparametric stochastic contextual bandits. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[25] Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. Smoothness-adaptive contextual bandits. Operations Research, 70(6):3198–3216, 2022.

[26] Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. arXiv preprint arXiv:2004.06321, 2020.

[27] Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and nondifferentiable regret regimes. Operations Research, 70(6):3261–3281, 2022.

[28] Cem Kalkanli and Ayfer Ozgur. Batched thompson sampling. Advances in Neural Information Processing Systems, 34:29984–29994, 2021.

[29] Amin Karbasi, Vahab Mirrokni, and Mohammad Shadravan. Parallelizing thompson sampling. Advances in Neural Information Processing Systems, 34:10535–10548, 2021.

[30] Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein Jr, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus Jr, Sanjay Gupta, et al. The battle trial: personalizing therapy for lung cancer. Cancer discovery, 1(1):44–53, 2011.

[31] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In Proceedings of the SIGCHI conference on human factors in computing systems, pages 453–456, 2008.

[32] Anders Bredahl Kock and Martin Thyrsgaard. Optimal sequential treatment allocation. arXiv preprint arXiv:1705.09952, 2017.

[33] Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. The Journal of Machine Learning Research, 21(1):5402–5446, 2020.

[34] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

[35] 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, pages 661–670, 2010.

[36] Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In Conference on Learning Theory, pages 1463–1492. PMLR, 2018.

[37] Tyler Lu, Dávid Pál, and Martin Pál. Showing relevant ads via context multi-armed bandits. In Proceedings of AISTATS, 2009.

[38] Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. The Annals of Statistics, 27(6):1808–1829, 1999.

[39] Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. Ann. Statist., 41(2):693–721, 2013.

[40] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. Ann. Statist., 44(2):660–681, 2016.

[41] Wei Qian, Ching-Kang Ing, and Ji Liu. Adaptive algorithm for multi-armed bandit problem with high-dimensional covariates. Journal of the American Statistical Association, pages 1–13, 2023.

[42] Wei Qian and Yuhong Yang. Kernel estimation and model combination in a bandit problem with covariates. Journal of Machine Learning Research, 17(149), 2016.

[43] Wei Qian and Yuhong Yang. Randomized allocation with arm elimination in a bandit problem with covariates. Electronic Journal of Statistics, 10(1):242–270, 2016.

[44] Dan Qiao, Ming Yin, Ming Min, and Yu-Xiang Wang. Sample-eficient reinforcement learning with loglog (t) switching cost. In International Conference on Machine Learning, pages 18031–18061. PMLR, 2022.

[45] Henry Reeve, Joe Mellor, and Gavin Brown. The k-nearest neighbour ucb algorithm for multi-armed bandits with covariates. In Algorithmic Learning Theory, pages 725–752. PMLR, 2018.

[46] Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. Management Science, 2023.

[47] Zhimei Ren, Zhengyuan Zhou, and Jayant R Kalagnanam. Batched learning in generalized linear contextual bandits with general decision sets. IEEE Control Systems Letters, 6:37–42, 2020.

[48] Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. arXiv preprint arXiv:1003.1630, 2010.

[49] Herbert E. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58:527–535, 1952.

[50] Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. Marketing Science, 36(4):500–522, 2017.

[51] Joe Suk and Samory Kpotufe. Tracking most significant shifts in nonparametric contextual bandits. arXiv preprint arXiv:2307.05341, 2023.

[52] Joseph Suk and Samory Kpotufe. Self-tuning bandits over unknown covariate-shifts. In Algorithmic Learning Theory, pages 1114–1156. PMLR, 2021.

[53] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In Mobile Health, pages 495–517. Springer, 2017.

[54] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. The Annals of Statistics, 32(1):135–166, 2004.

[55] Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. arXiv preprint arXiv:1706.09186, 2017.

[56] Chi-Hua Wang and Guang Cheng. Online batch decision-making with high-dimensional covariates. In International Conference on Artificial Intelligence and Statistics, pages 3848–3857. PMLR, 2020.

[57] Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably eficient reinforcement learning with linear function approximation under adaptivity constraints. Advances in Neural Information Processing Systems, 34:13524–13536, 2021.

[58] Michael Woodroofe. A one-armed bandit problem with a concomitant variable. Journal of the American Statistical Association, 74(368):799–806, 1979.

[59] Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. Ann. Statist., 30(1):100–121, 2002.

[60] Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. Advances in neural information processing systems, 33:9818–9829, 2020.

[61] Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. Advances in Neural Information Processing Systems, 33:15198–15207, 2020.

[62] Zhijin Zhou, Yingfei Wang, Hamed Mamani, and David G Coffey. How do tumor cytogenetics in-form cancer treatments? dynamic risk stratification and precision medicine using multi-armed bandits. Dynamic Risk Stratification and Precision Medicine Using Multi-armed Bandits (June 17, 2019).

A Proof of Theorem 2

It is worth emphasizing that Theorem 2 aims to establish the hardness of batched nonparametric bandits even when the grid Γ is allowed to be adaptively chosen. Nevertheless, the proof of Theorem 1 in the fixed-grid case is still useful.

Define $b = T^{(1-\gamma)/(1-\gamma^M)}$. For each $1 \le i \le M$, we set $T_i = \lfloor b^{(1-\gamma^i)/(1-\gamma)} \rfloor$, $z_i = \lceil (36 T_{i-1} M^2)^{1/(2\beta+d)} \rceil$, and $s_i := \lfloor z_i^{d-\alpha\beta} \rfloor$. We reuse the family of hard instances $\mathcal{C}_{z_i}$ as defined in (4), and define the mixture distribution

$$Q_i(\cdot) = \frac{1}{s_i 2^{s_i-1}} \sum_{j=1}^{s_i} \sum_{\omega_{[-j]} \in \Omega_{s_i-1}} P_{\pi,i,f_{\omega_{[-j]}^{-1}}}(\cdot) = \sum_{\omega \in \Omega_{s_i}} q_i(\omega) P_{\pi,i,f_\omega}(\cdot), \tag{13}$$

where $q_i : \Omega_{s_i} \to [0,1]$ is selected so that the above equality holds, P_{π,i,f_ω} is the distribution of the observations when $f_\omega \in \mathcal{C}_{z_i}$. It is easy to see $\sum_{\omega \in \Omega_{s_i}} q_i(\omega) = 1$.

We pause here to state a useful claim regarding the family $\{Q_i\}_{i=1}^M$ of mixture distributions, namely, they are all close to each other under the total variation distance.

Lemma 2. For any $1 \le i \le M$, one has $\mathrm{TV}(Q_M^{T_{i-1}}, Q_i^{T_{i-1}}) \le \frac{1}{2}\sqrt{T_{i-1} z_i^{-(2\beta+d)}}$.

Define the event

$$A_i = \{t_{i-1} < T_{i-1} < T_i \le t_i\}.$$

Intuitively, A_i models the event when the algorithm's selected grid points t_{i-1} and t_i are suboptimal. When A_i happens, the goal is to design a problem instance such that using observations up t_{i-1} cannot distinguish the optimal arm and therefore the policy must incur a large regret between t_{i-1} and t_i.

The following lemma ensures that at least one of the bad events A_i's happens with suficiently large probability under the mixture distribution.

Lemma 3. There exists $i^\star \in [M]$ such that $Q_{i^\star}(A_{i^\star}) \ge 1/(2M)$.

The next lemma indeed shows that when A_i happens, the regret must be large.

Lemma 4. If $Q_i(A_i) \ge 1/(2M)$, then

$$\sup_{f \in \mathcal{C}_{z_i}} R_{T_i}(\pi; f) \gtrsim T_i z_i^{-\beta(1+\alpha)} M^{-\frac{1+\alpha}{\alpha}}.$$

Now we are ready to establish the desired claim in the theorem. It is straightforward to see that

$$\sup_{(f,\frac{1}{2}) \in \mathcal{F}(\alpha,\beta)} R_T(\pi; f) \ge \sup_{f \in \mathcal{C}_{z_{i^\star}}} R_{T_{i^\star}}(\pi; f) \gtrsim T_{i^\star} z_{i^\star}^{-\beta(1+\alpha)} M^{-\frac{1+\alpha}{\alpha}} = \tilde{D}_1 \Big(\frac{1}{M}\Big)^{\tilde{D}_2} T^{\frac{1-\gamma}{1-\gamma^M}},$$

where the second inequality uses Lemma 4, and the last one arises from the definitions of T_i and z_i.

A.1 Proof of Lemma 2

It sufices to bound their KL divergence. By the standard decomposition of KL divergence and Bernoulli reward structure,

$$\mathrm{KL}(Q_M^{T_{i-1}}, Q_i^{T_{i-1}}) \le 8 \sum_{t=1}^{T_{i-1}} \mathbb{E}_{Q_M}\Bigg[\Bigg(\underbrace{\sum_{\omega \in \Omega_M} q_M(\omega) f_{M,\omega}(X_t) - \sum_{\omega \in \Omega} q_i(\omega) f_{i,\omega}(X_t)}_{\Delta_t}\Bigg)^2 \mathbb{1}\{\pi_t(X_t) = 1\}\Bigg], \tag{14}$$

where $f_{i,\omega}$ denotes an instance from $\mathcal{C}_{z_i}$. To control Δ_t, we further decompose it as

$$\Delta_t = \sum_{j=1}^{s_i} \Bigg(\sum_{\omega \in \Omega_M} q_M(\omega) f_{M,\omega}(X_t) - \sum_{\omega \in \Omega} q_i(\omega) f_{i,\omega}(X_t)\Bigg) \mathbb{1}\{X_t \in C_{i,j}\},$$

where $C_{i,j}$ is the j^{th} bin corresponding to the instance family C_{z_i}. Here, the difference between the two sums can be restricted to $\bigcup_{j=1}^{s_i} C_{i,j}$ because $\bigcup_{j=1}^{s_M} C_{M,j} \supseteq \bigcup_{j=1}^{s_i} C_{i,j}$. Indeed, the area of effective bins for an instance in C_{z_i} is $s_i z_i^{\ d} = z^{-\alpha\beta}$, which decreases as i increases. Notice for $X_t \in C_{i,j}$,

$$\left| \sum_{\omega \in \Omega} q_i(\omega) f_{i,\omega}(X_t) - \frac{1}{2} \right| \le \frac{2^{s_i - 1}}{s_i \cdot 2^{s_i - 1}} \cdot \frac{z_i^{-\beta}}{4} = \frac{z_i^{-\beta}}{4 s_i},$$

because all the $\omega_{[-j]}^{-1}$'s have a negative sign in the j^{th} bin and there are $2^{s_i - 1}$ of them, while for each $k = j$, the positive and negative spikes within the j^{th} bin cancel out each other when summing over $\omega_{[-k]}^{-1}$'s due to symmetry. Therefore,

$$|\Delta_t| \le \frac{z_i^{-\beta}}{4 s_i} \sum_{j=1}^{s_i} 1\{X_t \in C_{i,j}\} = \frac{1}{4} z_i^{-\beta - d + \alpha\beta} \sum_{j=1}^{s_i} 1\{X_t \in C_{i,j}\}.$$

Plugging the above back to (14) we obtain

$$\begin{aligned} KL(Q_M^{T_{i-1}}, Q_i^{T_{i-1}}) &\le \frac{1}{2} \sum_{t=1}^{T_{i-1}} E_{Q_M}\left[z_i^{-2(\beta + d - \alpha\beta)} \sum_{j=1}^{s_i} 1\{X_t \in C_{i,j}\} 1\{\pi_t(X_t) = 1\} \right] \\ &= \frac{1}{2} z_i^{-2(\beta + d - \alpha\beta)} \sum_{t=1}^{T_{i-1}} \sum_{j=1}^{s_i} \left[P_{Q_M}(X_t \in C_{i,j}, \pi_t(X_t) = 1) \right] \\ &\le \frac{1}{2} z_i^{-2(\beta + d - \alpha\beta)} \sum_{t=1}^{T_{i-1}} \sum_{j=1}^{s_i} z_i^{-d} = \frac{1}{2} T_{i-1} z_i^{-(2\beta + d + (d - \alpha\beta))}, \end{aligned}$$

where the last inequality is because $P_{Q_M}(X_t \in C_{i,j}, \pi_t(X_t) = 1) = z_i^{-d} P_{Q_M}(\pi_t(X_t) = 1 \mid X_t \in C_{i,j}) \le z_i^{-d}$. Since $\alpha\beta \le 1$,

$$KL(Q_M^{T_{i-1}}, Q_i^{T_{i-1}}) \le \frac{1}{2} T_{i-1} z_i^{-(2\beta + d + (d - \alpha\beta))} \le \frac{1}{2} T_{i-1} z_i^{-(2\beta + d)}.$$

By Pinsker's inequality, we can conclude

$$TV(Q_M^{T_{i-1}}, Q_i^{T_{i-1}}) \le \sqrt{\frac{1}{2} KL(Q_M^{T_{i-1}}, Q_i^{T_{i-1}})} \le \sqrt{\frac{1}{2} T_{i-1} z_i^{-(2\beta + d)}}.$$

A.2 Proof of Lemma 3

For any $1 \le i \le M$, we have

$$|Q_M(A_i) - Q_i(A_i)| \overset{(i)}{=} |Q_M^{T_{i-1}}(A_i) - Q_i^{T_{i-1}}(A_i)| \overset{(ii)}{\le} TV(Q_M^{T_{i-1}}, Q_i^{T_{i-1}}) \overset{(iii)}{\le} \frac{1}{2M}, \tag{15}$$

where step (i) is because A_i can be determined by observations up to T_{i-1}, step (ii) uses the definition of TV, and step (iii) applies Lemma 2 and the definition of z_i. Consequently,

$$\begin{aligned} \sum_{i=1}^{M} Q_i(A_i) &= Q_M(A_M) + \sum_{i=1}^{M-1} Q_i(A_i) \\ &= Q_M(A_M) + \sum_{i=1}^{M-1} (Q_i(A_i) - Q_M(A_i) + Q_M(A_i)) \\ &\overset{(iv)}{\ge} Q_M(A_M) + \sum_{i=1}^{M-1} (Q_M(A_i) - \frac{1}{2M}) \ge \sum_{i=1}^{M} Q_M(A_i) - \frac{1}{2} \overset{(v)}{=} \frac{1}{2}, \end{aligned}$$

where step (iv) uses inequality (15), and step (v) uses the fact that $\sum_{i=1}^{M} Q_M(A_i) = 1$.

A.3 Proof of Lemma 4

We try to lower-bound the number of mistakes we make up to T_i. By inequality (5),

$$\begin{aligned}
\sup_{f \in C_z} S_{T_i}(\pi, f, \frac{1}{2}) &\geq \frac{1}{2^s} \sum_{j=1}^{s} \sum_{t=1}^{T_i} \sum_{\omega_{[-j]} \in \Omega_{-1}} \frac{1}{z^d} \sum_{h \in \{\pm 1\}} P_{\pi, f_{\omega^h_{[-j]}}}(\pi_t(X_t) = h \mid X_t \in C_j) \\
&\geq \frac{1}{2^s} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} \frac{1}{z^d} \sum_{t=1}^{T_i} \int \min\{dP^t_{\pi, f_{\omega^{-1}_{[-j]}}}, dP^t_{\pi, f_{\omega^1_{[-j]}}}\} \\
&\geq \frac{1}{2^s} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} \frac{T_i}{z^d} \min\{dP^{T_i}_{\pi, f_{\omega^{-1}_{[-j]}}}, dP^{T_i}_{\pi, f_{\omega^1_{[-j]}}}\},
\end{aligned}$$

where the second inequality invokes Le Cam's method, and the last inequality holds since $\int \min\{dP, dQ\} = 1 - TV(P, Q)$, and $TV(P^t_{\pi, f_{\omega^{-1}_{[-j]}}}, P^t_{\pi, f_{\omega^1_{[-j]}}}) \leq TV(P^{T_i}_{\pi, f_{\omega^{-1}_{[-j]}}}, P^{T_i}_{\pi, f_{\omega^1_{[-j]}}})$ for $t \leq T_i$. We continue the lower-bound to see that

$$\begin{aligned}
\sup_{f \in C_z} S_{T_i}(\pi, f, \frac{1}{2}) &\geq \frac{1}{2^s} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} \frac{T_i}{z^d} \int \min\{dP^{T_i}_{\pi, f_{\omega^{-1}_{[-j]}}}, dP^{T_i}_{\pi, f_{\omega^1_{[-j]}}}\} \\
&\geq \frac{1}{2^s} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} \frac{T_i}{z^d} \int_{A_i} \min\{dP^{T_i}_{\pi, f_{\omega^{-1}_{[-j]}}}, dP^{T_i}_{\pi, f_{\omega^1_{[-j]}}}\} \\
&= \frac{1}{2^s} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} \frac{T_i}{z^d} \int_{A_i} \min\{dP^{T_{i-1}}_{\pi, f_{\omega^{-1}_{[-j]}}}, dP^{T_{i-1}}_{\pi, f_{\omega^1_{[-j]}}}\},
\end{aligned}$$

where the last step uses the fact that under A_i, the available observations for π at T_i are the same as those at T_{i-1}. Using properties of TV, we reach

$$\begin{aligned}
\sup_{f \in C_z} S_{T_i}(\pi, f, \frac{1}{2}) &\geq \frac{1}{2^s} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} \frac{T_i}{z^d} \int_{A_i} \min\{dP^{T_{i-1}}_{\pi, f_{\omega^{-1}_{[-j]}}}, dP^{T_{i-1}}_{\pi, f_{\omega^1_{[-j]}}}\} \\
&\geq \frac{1}{2^s} \cdot \frac{T_i}{z^d} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} P_{\pi, f_{\omega^{-1}_{[-j]}}}(A_i) - \frac{3}{2} TV(P^{T_{i-1}}_{\pi, f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi, f_{\omega^1_{[-j]}}}) \\
&\geq \frac{1}{2^s} \cdot \frac{T_i}{z^d} \sum_{j=1}^{s} \sum_{\omega_{[-j]} \in \Omega_{-1}} P_{\pi, f_{\omega^{-1}_{[-j]}}}(A_i) - \frac{3}{2} \sqrt{\frac{1}{2} KL(P^{T_{i-1}}_{\pi, f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi, f_{\omega^1_{[-j]}}})},
\end{aligned}$$

where the second inequality applies Lemma 6, and the third inequality is due to Pinsker's inequality. Take $z = z_i$ and use Lemma 5, we have

$$\begin{aligned}
\sup_{f \in C_{z_i}} S_{T_i}(\pi, f, \frac{1}{2}) &\geq \frac{1}{2^{s_i}} \cdot \frac{T_i}{z_i^d} \sum_{j=1}^{s_i} \sum_{\omega^{[-j]} \in \Omega_{s_i - 1}} P_{\pi, i, f_{\omega^{-1}_{[-j]}}}(A_i) - \frac{3}{2} \sqrt{z_i^{-(2\beta+d)} T_{i-1}} \\
&= \frac{1}{2^{s_i}} \cdot \frac{T_i}{z_i^d} \left(s_i 2^{s_i - 1} Q_i(A_i) - \frac{3}{2} s_i 2^{s_i - 1} \sqrt{\frac{1}{36M^2}} \right) \\
&\geq \frac{1}{8} T_i z_i^{-\alpha\beta} \frac{1}{M},
\end{aligned}$$

where the last inequality uses the assumption $Q_i(A_i) \geq 1/(2M)$. Therefore, we arrive at

$$\sup_{f \in C_{z_i}} R_{T_i}(\pi; f) \gtrsim T_i^{-\frac{1}{\alpha}} \left(\sup_{f \in C_{z_i}} S_{t_i}(\pi; f) \right)^{\frac{1+\alpha}{\alpha}} \gtrsim T_i z_i^{-\beta(1+\alpha)} M^{-\frac{1+\alpha}{\alpha}}.$$

Lemma 5. Fix $z > 0$ and suppose $f_\omega \in C_z$. For any $n \in [T]$ and any policy π, one has

$$KL(P^n_{\pi,f_{\omega^{-1}_{[-j]}}}, P^n_{\pi,f_{\omega^1_{[-j]}}}) \le 2nz^{-(2\beta+d)}.$$

Proof. We can compute

$$\begin{aligned}
KL(P^n_{\pi,f_{\omega^{-1}_{[-j]}}}, P^n_{\pi,f_{\omega^1_{[-j]}}}) &\overset{(i)}{\le} 8E_{\pi,f_{\omega^{-1}_{[-j]}}}\Big[\sum_{t=1}^{n}(f_{\omega^{-1}_{[-j]}}(X_t) - f_{\omega^1_{[-j]}}(X_t))^2 1\{\pi_t(X_t) = 1\}\Big] \\
&\overset{(ii)}{\le} 32D_\varphi^2 z^{-2\beta} E_{\pi,f_{\omega^{-1}_{[-j]}}}\Big[\sum_{t=1}^{n} 1\{\pi_t(X_t) = 1, X_t \in C_j\}\Big] \\
&\overset{(iii)}{=} 32D_\varphi^2 z^{-(2\beta+d)} \sum_{t=1}^{n} P^t_{\pi,f_{\omega^{-1}_{[-j]}}}(\pi_t(X_t) = 1 \,|\, X_t \in C_j) \\
&\overset{(iv)}{\le} 32D_\varphi^2 z^{-(2\beta+d)} n \le 2nz^{-(2\beta+d)}.
\end{aligned}$$

Here, step (i) uses the standard decomposition of KL divergence and Bernoulli reward structure; step (ii) is due to the definition of f_ω; step (iii) uses $P(X_t \in C_j) = 1/z^d$, and step (iv) arises from $P^t_{\pi,f_{\omega^{-1}_{[-j]}}}(\pi_t(X_t) = 1 \,|\, X_t \in C_j) \le 1$ for any $1 \le t \le n$. □

Lemma 6. For any $i \in [M]$, one has

$$\int_{A_i} \min\{dP^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, dP^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}\} \ge P_{\pi,f_{\omega^{-1}_{[-j]}}}(A_i) - \frac{3}{2}TV(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}).$$

Proof. We can compute

$$\begin{aligned}
\int_{A_i} \min\{dP^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, dP^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}\} &= \int_{A_i} \frac{dP^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}} + dP^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}} - |dP^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}} - dP^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}|}{2} \\
&\ge \frac{1}{2}(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}(A_i) + P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}(A_i)) - TV(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}) \\
&= \frac{1}{2}(2P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}(A_i) + P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}(A_i) - P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}(A_i)) - TV(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}) \\
&\overset{(i)}{\ge} P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}(A_i) - \frac{1}{2}TV(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}) - TV(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}) \\
&\overset{(ii)}{=} P_{\pi,f_{\omega^{-1}_{[-j]}}}(A_i) - \frac{3}{2}TV(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}),
\end{aligned}$$

where step (i) is due to $|P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}}(A_i) - P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}(A_i)| \le TV(P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}, P^{T_{i-1}}_{\pi,f_{\omega^1_{[-j]}}})$, and step (ii) uses the fact that $P_{\pi,f_{\omega^{-1}_{[-j]}}}$ and $P^{T_{i-1}}_{\pi,f_{\omega^{-1}_{[-j]}}}$ are equivalent on A_i. □

B Proof of Theorem 3

Our proof of Theorem 3 is inspired by the framework developed in [39]. Our setting presents additional technical dificulty due to the batch constraint.

We begin with introducing some useful notations. Recall the tree growing process described in section 4, where we have defined a tree T of depth M. The root (depth 0) of the tree is the whole space X. In depth 1, X has g_0^d children, each of which is a bin of width $1/g_0$. For each bin in depth 1, it has g_1^d children, each of which is a bin of width $1/(g_0 g_1)$. These children form the depth 2 nodes of the tree T. We form the tree recursively until depth M.

For a bin $C \in \mathcal{T}$, we define its parent by $p(C) = \{C' \in \mathcal{T} : C \in \text{child}(C')\}$. Moreover, we let $p^1(C) = p(C)$ and define $p^k(C) = p(p^{k-1}(C))$ for $k \geq 2$ recursively. In all, we denote by $P(C) = \{C' \in \mathcal{T} : C' = p^k(C) \text{ for some } k \geq 1\}$ all the ancestors of the bin C.

We also define L_t to be the set of active bins at time t, with the dummy case $L_0 = \{X\}$. Clearly, for $1 \leq t \leq t_1$, one has $L_1 = B_1$, where B_1 are all the bins in the first layer.

B.1 Two clean events

The regret analysis relies on two clean events. First, fix a batch $i \geq 1$, and recall $L_{t_{i-1}+1}$ is the set of active bins at time $t_{i-1} + 1$. We denote the random number of pulls for a bin $C \in L_{t_{i-1}+1}$ within batch i to be

$$m_{C,i} := \sum_{t=t_{i-1}+1}^{t_i} 1\{X_t \in C\}.$$

Clearly, it has expectation

$$m^*_{C,i} = E[m_{C,i}] = (t_i - t_{i-1})P_X(X \in C).$$

The first clean event claims that $m_{C,i}$ concentrates well around its expectation $m^*_{C,i}$ uniformly over all $C \in \mathcal{T}$. We denote this event by E.

Lemma 7. Suppose that $M \leq D_1 \log(T)$ for some constant $D_1 > 0$. With probability at least $1 - 1/T$, for all $1 \leq i \leq M$, and $C \in L_{t_{i-1}+1}$, we have

$$\frac{1}{2} m^*_{C,i} \leq m_{C,i} \leq \frac{3}{2} m^*_{C,i}.$$

See Section B.5.1 for the proof.

Since $M \leq D_1 \log(T)$ by assumption, we can apply Lemma 7 to obtain

$$E[R_T(\hat{\pi})1(E^c)] \leq T P(E^c) = 1.$$

Therefore, in the remaining proof, we condition on E and focus on bounding $E[R_T(\hat{\pi})1(E)]$.

The second clean event is on the elimination process. Since we use successive elimination in each bin, it is natural to expect that the optimal arm in each bin is not eliminated during the process. To mathematically specify this event, we need a few notations.

For each bin $C \in L_i$, let I'_C be the set of remaining arms at the end of batch i, i.e., after Algorithm 2 is invoked. Define

$$\bar{I}_C = \left\{k \in \{1, -1\} : \sup_{x \in C} f^*(x) - f^{(k)}(x) \leq c_1 |C|^\beta\right\},$$

$$\underline{I}_C = \left\{k \in \{1, -1\} : \sup_{x \in C} f^*(x) - f^{(k)}(x) \leq c_0 |C|^\beta\right\},$$

where $c_0 = 2Ld^{\beta/2} + 1$ and $c_1 = 8c_0$. Clearly, we have

$$\underline{I}_C \subseteq \bar{I}_C.$$

Define a good event $A_C = \{\underline{I}_C \subseteq I'_C \subseteq \bar{I}_C\}$, which is the event that the remaining arms in C have gaps of correct order. In addition, define $G_C = \cap_{C' \in P(C)} A_{C'}$. Recall B_i is the set of bins C with $|C| = (\prod_{l=1}^{i} g_l)^{-1} = w_i$ for $i \geq 1$.

Lemma 8. For any $1 \leq i \leq M - 1$ and $C \in B_i$, we have

$$P(E \cap G_C \cap A_C^c) \leq \frac{4m^*_{C,i}}{T|C|^d}.$$

In words, Lemma 8 guarantees that A_C happens with high probability if E holds and $A_{C'}$ holds for all the ancestors C' of C. See Section B.5.2 for the proof.

B.2 Regret decomposition

In this section, we decompose the regret into three terms. First, for a bin C, we define

$$r_T^{\text{live}}(C) := \sum_{t=1}^{T} f^{(1)}(X_t) - f^{(\pi_t(X_t))}(X_t) \quad 1(X_t \in C)1(C \in L_t).$$

In addition, define $J_t := \cup_{s \le t} L_s$ to be the set of bins that have been live up until time t. Correspondingly we define

$$r_T^{\text{born}}(C) := \sum_{t=1}^{T} f^{(1)}(X_t) - f^{(\pi_t(X_t))}(X_t) \quad 1(X_t \in C)1(C \in J_t).$$

It is clear from the definition that for any $C \in \mathcal{T}$, one has

$$\begin{aligned} r_T^{\text{born}}(C) &= r_T^{\text{live}}(C) + \sum_{C' \in \text{child}(C)} r_T^{\text{born}}(C') \\ &= r_T^{\text{born}}(C)1(A_C^c) + r_T^{\text{live}}(C)1(A_C) + \sum_{C' \in \text{child}(C)} r_T^{\text{born}}(C')1(A_C). \end{aligned}$$

Applying this relation recursively leads to the following regret decomposition:

$$\begin{aligned} R_T(\pi) &= r_T^{\text{born}}(X) \\ &= \underbrace{r_T^{\text{live}}(X)}_{=0} + \sum_{C' \in \text{child}(X)} r_T^{\text{born}}(C') \\ &= \sum_{1 \le i < M} \Bigg[\sum_{C \in B_i} r_T^{\text{born}}(C)1(G_C \cap A_C^c) + \sum_{C \in B_i} r_T^{\text{live}}(C)1(G_C \cap A_C) \Bigg] \\ &\qquad + \sum_{C \in B_M} r_T^{\text{live}}(C)1(G_C), \end{aligned}$$

where the second equality arises from the fact that $r_T^{\text{live}}(X) = 0$. Indeed, $X \notin L_t$ for any $1 \le t \le T$.

B.3 Controlling three terms

In what follows, we control V_i, U_i and the last batch separately.

B.3.1 Controlling V_i

Fix some $1 \le i \le M-1$, and some bin $C \in B_i$. On the event G_C we have $I_{p(C)}' \subset I_{p(C)}^-$, that is, for any $k \in I_{p(C)}'$,

$$\sup_{x \in p(C)} f^{(1)}(x) - f^{(k)}(x) \le c_1 |p(C)|^{\beta}.$$

This implies that for any $x \in C$, and $k \in I_{p(C)}'$,

$$f^{(1)}(x) - f^{(k)}(x) \quad 1\{G_C\} \le c_1 |p(C)|^{\beta} 1(0 < f^{(1)}(x) - f^{(-1)}(x) \le c_1 |p(C)|^{\beta}). \tag{16}$$

As a result, we obtain

$$E[r_T^{\text{live}}(C)1(G_C \cap A_C)] = E\left[\sum_{t=1}^{T} f^{(1)}(X_t) - f^{(\pi_t(X_t))}(X_t) \quad 1(X_t \in C)1(C \in L_t)1(G_C \cap A_C) \right]$$

$$\overset{(i)}{\leq} E\left[\sum_{t=1}^{T} c_1|p(C)|^\beta 1(0 < f^{(1)}(X_t) - f^{(-1)}(X_t) \leq c_1|p(C)|^\beta) 1(X_t \in C, C \in L_t) 1(G_C \cap A_C)\right]$$

$$\overset{(ii)}{\leq} c_1|p(C)|^\beta E\left[\sum_{t=t_{i-1}+1}^{t_i} 1(0 < f^{(1)}(X_t) - f^{(-1)}(X_t) \leq c_1|p(C)|^\beta, X_t \in C) 1(G_C \cap A_C)\right]$$

$$\overset{(iii)}{\leq} c_1|p(C)|^\beta \sum_{t=t_{i-1}+1}^{t_i} P(0 < f^{(1)}(X_t) - f^{(-1)}(X_t) \leq c_1|p(C)|^\beta, X_t \in C)$$

$$= c_1|p(C)|^\beta (t_i - t_{i-1}) P(0 < f^{(1)}(X) - f^{(-1)}(X) \leq c_1|p(C)|^\beta, X \in C).$$

Here, step (i) uses relation (16), and the fact that $\pi_t(X_t) \in I'_{p(C)}$ when $X_t \in C$. For step (ii), if C is split, then it is no longer live, so the live regret incurred on the remaining batches is zero. On the other hand, if C is not split, then $|I'_C| = 1$. Without loss of generality, assume that arm -1 is eliminated. Conditioned on A_C, this means $-1 \notin I_C$ and there exists $x_0 \in C$ such that $f^{(1)}(x_0) - f^{(-1)}(x_0) > c_0|C|^\beta$. By the smoothness condition, having a gap at least $c_0|C|^\beta$ on a single point in C implies $f^{(1)}(x) - f^{(-1)}(x) > |C|^\beta$ for all $x \in C$. Therefore, arm 1 which is the remaining one is the optimal arm for all $x \in C$ and would not incur any regret further. The third inequality holds since $1(G_C \cap A_C) \leq 1$.

Taking the sum over all bins in B_i and using the fact that $|p(C)| = w_{i-1}$, we obtain

$$\sum_{C \in B_i} E[r_T^{live}(C) 1(G_C \cap A_C)] \leq \sum_{C \in B_i} c_1 w_{i-1}^\beta (t_i - t_{i-1}) P(0 < f^{(1)}(X) - f^{(-1)}(X) \leq c_1|p(C)|^\beta, X \in C)$$

$$= c_1 w_{i-1}^\beta (t_i - t_{i-1}) \sum_{C \in B_i} P(0 < f^{(1)}(X) - f^{(-1)}(X) \leq c_1 w_{i-1}^\beta, X \in C). \quad (17)$$

Note that

$$\sum_{C \in B_i} P(0 < f^{(1)}(X) - f^{(-1)}(X) \leq c_1 w_{i-1}^\beta, X \in C) = P(0 < f^{(1)}(X) - f^{(-1)}(X) \leq c_1 w_{i-1})^\beta$$

$$\leq D_0 \cdot \left[c_1 w_{i-1}^\beta\right]^\alpha, \quad (18)$$

where the last inequality follows from the margin condition. Combining relations (18) and (17), we reach

$$\sum_{C \in B_i} E[r_T^{live}(C) 1(G_C \cap A_C)] \leq (t_i - t_{i-1}) \cdot [c_1 w_{i-1}^\beta]^{1+\alpha} \cdot D_0.$$

B.3.2 Controlling U_i

Fix some $1 \leq i \leq M - 1$, and some bin $C \in B_i$. Again, using the definition of G_C, we obtain

$$E[r_T^{born}(C) 1(G_C \cap A_C^c)] = E\left[\sum_{t=1}^{T} f^*(X_t) - f^{(\pi_t(X_t))}(X_t) \quad 1(X_t \in C) 1(C \in J_t) 1(G_C \cap A^c{}_C)\right]$$

$$\leq E\left[\sum_{t=1}^{T} c_1|p(C)|^\beta 1(0 < f^{(1)}(X_t) - f^{(-1)}(X_t) \leq c_1|p(C)|^\beta) 1(X_t \in C, C \in J_t) 1(G_C \cap A^c{}_C)\right]$$

$$\leq c_1|p(C)|^\beta T P(0 < f^{(1)}(X) - f^{(-1)}(X) \leq c_1|p(C)|^\beta, X \in C) P(G_C \cap A^c{}_C)$$

Apply Lemma 8 to see that

$$E[r_T^{born}(C) 1(G_C \cap A^c{}_C)] \leq c_1|p(C)|^\beta T P(0 < f^{(1)}(X) - f^{(-1)}(X) \leq c_1|p(C)|^\beta, X \in C) \frac{4m}{T|C|^d}$$

$$= c_1 w_{i-1}^{\beta} P(0 < f^{(1)}(X) - f^{(-1)}(X) \le c_1 w_{i-1}^{\beta}, X \in C) \frac{4(t_i - t_{i-1}) P_X(X \in C)}{|C|^d}$$
$$\le 4\bar{c} c_1 w_{i-1}^{\beta} P(0 < f^{(1)}(X) - f^{(-1)}(X) \le c_1 w_{i-1}^{\beta}, X \in C)(t_i - t_{i-1}),$$

where we use the fact that $P_X(X \in C) \le \bar{c}|C|^d$ in the second inequality. Summing over all bins in B_i, we obtain

$$\sum_{C \in B_i} E[r_T^{born}(C) 1(G_C \cap A_C^c)] \le 4\bar{c} c_1 w_{i-1}^{\beta}(t_i - t_{i-1}) \sum_{C \in B_i} P(0 < f^{(1)}(X) - f^{(-1)}(X) \le c_1 w_{i-1}^{\beta}, X \in C)$$
$$\le 4\bar{c} c_1 w_{i-1}^{\beta}(t_i - t_{i-1}) D_0 \cdot \left[c_1 w_{i-1}^{\beta}\right]^{\alpha}$$
$$= 4 D_0 \bar{c}(t_i - t_{i-1})[c_1 w_{i-1}^{\beta}]^{1+\alpha},$$

where the second inequality reuses the bound in (18).

B.3.3 Last Batch

For $C \in B_M$, one can similarly obtain

$$E[r_T^{live}(C) 1(G_C)] \le c_1 |p(C)|^{\beta}(T - t_{M-1}) P(0 < f^{(1)}(X) - f^{(-1)}(X) \le c_1 |p(C)|^{\beta}, X \in C).$$

Consequently, summing over $C \in B_M$ yields

$$\sum_{C \in B_M} E[r_T^{live}(C) 1(G_C)] \le \sum_{C \in B_M} c_1 |p(C)|^{\beta}(T - t_{M-1}) P(0 < f^{(1)}(X) - f^{(-1)}(X) \le c_1 |p(C)|^{\beta}, X \in C)$$
$$\le c_1 w_{M-1}^{\beta}(T - t_{M-1}) D_0 \cdot \left[c_1 w_{M-1}^{\beta}\right]^{\alpha}$$
$$= D_0 (T - t_{M-1})[c_1 w_{M-1}^{\beta}]^{1+\alpha}.$$

B.4 Putting things together

In sum, the total regret is bounded by

$$E[R_T(\pi)] \le c \left(t_1 + \sum_{i=2}^{M-1} (t_i - t_{i-1}) \cdot w_{i-1}^{\beta+\alpha\beta} + (T - t_{M-1}) w_{M-1}^{\beta+\alpha\beta} \right),$$

where c is a constant that depends on (α, β, D, L). Recall that $w_i = (\prod_{l=0}^{Q_{i-1}} g_l)^{-1}$, and the choices for the batch size and the split factors (12)-(11). We then obtain

$$t_1 \asymp T^{\frac{1-\gamma}{1-\gamma^M}} \log T,$$
$$(t_i - t_{i-1}) \cdot w_{i-1}^{\beta+\alpha\beta} \asymp T^{\frac{1-\gamma}{1-\gamma^M}} \log T, \qquad \text{for } 2 \le i \le M-1,$$
$$(T - t_{M-1}) w_{M-1}^{\beta+\alpha\beta} \le T w_{M-1}^{\beta+\alpha\beta} \asymp T^{\frac{1-\gamma}{1-\gamma^M}} \log T.$$

The proof is finished by combining the above three bounds.

B.5 Proofs for the clean events

We are left with proving that the two clean events happen with high probability.

B.5.1 Proof of Lemma 7

Fix the batch index i, and a node C in layer-i of the tree $\mathcal{T}$. By relation (12), we have

$$\begin{aligned} m_{C,i} &= (t_i - t_{i-1}) P_X(X \in C) \\ &\geq |C|^{-(2\beta+d)} \log(T|C|^d) P_X(X \in C) \\ &\geq |C|^{-2\beta} \geq g_0^{2\beta} \cdot (T^{\frac{-1+\gamma}{1-\gamma+M\cdot\gamma} \cdot \frac{2\beta}{2\beta+d}}), \end{aligned}$$

where the last step uses the fact that $P_X(X \in C) \geq \underline{c}|C|^d$. Therefore, $m_{C,i} \geq \frac{3}{4}\log(2T^2)$ for all i and C, as long as T is suficiently large. This allows us to invoke Chernoff's bound to obtain that with probability at most $1/T^2$

$$\sum_{t=t_{i-1}+1}^{t_i} 1\{X_t \in C\} - m_{C,i} \geq \sqrt{3\log(2T^2) m_{C,i}}.$$

Denote $E^c = \{\exists 1 \leq i \leq M, C \in \mathcal{L}_{t_{i-1}+1} \text{ such that } |\sum_{t=t_{i-1}+1}^{t_i} 1\{X_t \in C\} - m_{C,i}| \geq \sqrt{3\log(2T^2) m_{C,i}}\}$. Applying union bound to reach

$$P(E^c) \leq \sum_{C \in \mathcal{T}} \frac{1}{T^2} \overset{\text{(i)}}{\leq} \frac{1}{T^2} \sum_{i=1}^{M} \Big(\prod_{l=0}^{i-1} g_l\Big)^d \overset{\text{(ii)}}{\leq} \frac{1}{T^2} \cdot M \cdot \Big(\prod_{l=0}^{M-1} g_l\Big)^d,$$

where step (i) sums over all possible nodes of $\mathcal{T}$ across batches, and step (ii) is due to $(\prod_{l=0}^{i-1} g_l)^d \leq (\prod_{l=0}^{M-1} g_l)^d$ for any $1 \leq i \leq M$. Since $g_{M-1} = 1$, we further obtain

$$P(E^c) \leq \frac{1}{T^2} \cdot M \cdot \Big(\prod_{l=0}^{M-2} g_l\Big)^d \overset{\text{(iii)}}{\leq} \frac{1}{T^2} \cdot M \cdot t_{M-1}^{\frac{d}{2\beta+d}} \overset{\text{(iv)}}{\leq} D_1 \frac{1}{T^2} \cdot \log T \cdot T^{\frac{d}{2\beta+d}} \leq \frac{1}{T},$$

where step (iii) invokes relation (12), and step (iv) uses the assumption $M \leq D_1 \log T$. This completes the proof.

B.5.2 Proof of Lemma 8

To simplify notation, for any event F, we define $P^{G_C}(F) = P(E \cap G_C \cap F)$.

Let D_C^1 be the event that an arm $k \in \underline{I}_C$ is eliminated at the end of batch i, and D_C^2 be the event that an arm $k \notin \bar{I}_C$ is not eliminated at the end of batch i. Consequently, we have

$$P^{G_C}(A_C^c) = P^{G_C}(D_C^1) + P^{G_C}((D_C^1)^c \cap D_C^2).$$

Recall $U(\tau, T, C) = 4\sqrt{\frac{\log(2T|C|^d)}{\tau}}$. By relation (12), we can write

$$\begin{aligned} m_{C,i} &= (t_i - t_{i-1}) P_X(X \in C) \\ &= l_i |C|^{-(2\beta+d)} \log(T|C|^d) P_X(X \in C), \end{aligned}$$

where $l_i > 0$ is a constant chosen such that $U(2m_{C,i}, T, C) = 2c_0|C|^\beta$. Under E, we have $U(m_{C,i}, T, C) \leq 4c_0|C|^\beta$ because $m_{C,i} \geq \frac{1}{2} m_{C,i}$.

1. Upper bounding $P^{G_C}(D_C^1)$: when D_C^1 occurs, an arm $k \in \underline{I}_C$ is eliminated by some $k' \in I_{p(C)}$ at the end of batch i. This means $\bar{Y}_{C,i}^{(k')} - \bar{Y}_{C,i}^{(k)} > U(m_{C,i}, T, C)$. Meanwhile,

$$\bar{f}_C^{(k')} - \bar{f}_C^{(k)} \leq \bar{f}_C - \bar{f}_C^{(k)} \overset{\text{(i)}}{\leq} c_0|C|^\beta \leq \frac{1}{2} U(2m_{C,i}, T, C),$$

where step (i) uses the definition of $\underline{I}_C$. Consequently, $|\bar{Y}_{C,i}^{(k')} - \bar{f}_C^{(k')}| \leq U(m_{C,i}, T, C)/4$ and $|\bar{Y}_{C,i}^{(k)} - \bar{f}_C^{(k)}| \leq U(m_{C,i}, T, C)/4$ cannot hold simultaneously. Otherwise, this would contradict with $\bar{Y}_{C,i}^{(k')} - \bar{Y}_{C,i}^{(k)} > U(m_{C,i}, T, C)$ because $m_{C,i} \leq 2m_{C,i}$ under E. Therefore,

$$P^{G_C}(D_C^1) \leq P\Big(\exists k \in I_{p(C)}, m_{C,i} \leq 2m_{C,i} : |\bar{Y}_{C,i}^{(k)} - \bar{f}_C^{(k)}| \geq \frac{1}{4} U(m_{C,i}, T, C)\Big).$$

2. Upper bounding $P^{G_C}((D_C^1)^c \cap D_C^2)$: when $(D_C^1)^c \cap D_C^2$ happens, no arm in $\underline{I}_C$ is eliminated while some $k \notin \bar{I}_C$ remains in the active arm set. By definition, there exists $x^{(k)}$ such that $f^*(x^{(k)}) - f^{(k)}(x^{(k)}) > 8c_0|C|^\beta$. Let $\eta(k)$ be any arm that satisfies $f^*(x^{(k)}) = f^{(\eta(k))}(x^{(k)})$, and one can easily verify $\eta(k) \in \underline{I}_C$. Since k is not eliminated, we have $\bar{Y}_{C,i}^{(\eta(k))} - \bar{Y}_{C,i}^{(k)} \le U(m_{C,i}, T, C)$. On the other hand,

$$\begin{aligned} \bar{f}_C^{(\eta(k))} &\overset{\text{(iii)}}{\ge} f^{(\eta(k))}(x^{(k)}) - c_0|C|^\beta \\ &\ge f^{(k)}(x^{(k)}) + 8c_0|C|^\beta - c_0|C|^\beta \\ &= f^{(k)}(x^{(k)}) + 7c_0|C|^\beta \\ &\overset{\text{(iv)}}{\ge} \bar{f}_C^{(k)} + 6c_0|C|^\beta \ge \bar{f}_C^{(k)} + \frac{3}{2}U(m_{C,i}, T, C), \end{aligned} \tag{19}$$

where steps (iii) and (iv) use Lemma 10. Inequality (19) together with the fact that $\bar{Y}_{C,i}^{(\eta(k))} - \bar{Y}_{C,i}^{(k)} \le U(m_{C,i}, T, C)$ imply $|\bar{Y}_{C,i}^{(k_0)} - \bar{f}_C^{(k_0)}| \ge U(m_{C,i}, T, C)/4$ for either $k_0 = k$ or $k_0 = \eta(k)$. Consequently,

$$P^{G_C}((D_C^1)^c \cap D_C^2) \le P\left(\exists k \in I'_{p(C)}, m_{C,i} \le 2m_{C,i}^* : |Y_{C,\bar{I}(k)}^{(k)} - f_C^{(k)}| \ge \frac{1}{4}U(m_{C,i}, T, C)\right).$$

Combining the two parts we obtain

$$\begin{aligned} P^{G_C}(A_C^c) &= P^{G_C}(D_C^1) + P^{G_C}((D_C^1)^c \cap D_C^2) \\ &\le 2 \cdot P\left(\exists k \in I_{p(C)}, m_{C,i} \le 2m_{C,i}^* : |\bar{Y}_{C,i}^{(k)} - f_C^{(k)}| \ge \frac{1}{4}U(m_{C,i}, T, C)\right) \\ &\le \frac{4m_{C,i}^*}{T|C|^d}, \end{aligned}$$

where the last inequality applies Lemma 9.

B.6 Auxiliary lemmas

Lemma 9. For any $1 \le i \le M-1$ and $C \in \mathcal{B}_i$, one has

$$P\left(\exists k \in I'_{p(C)}, m_{C,i} \le 2m_{C,i}^* : |\bar{Y}_{C,i}^{(k)} - f_C^{(k)}| \ge \frac{1}{4}U(m_{C,i}, T, C)\right) \le \frac{2m_{C,i}^*}{T|C|^d}.$$

Proof. Recall in Algorithm 1 we pull each arm in a round-robin fashion within a bin during batch i. Fix $\tau > 0$. Let $\bar{Y}^{(k)}_\tau = \sum_{j=1}^{\tau} Y_j^{(k)}/\tau$ where $Y_j^{(k)}$'s are i.i.d. random variables with $Y_j^{(k)} \in [0,1]$ and $E[Y_j^{(k)}] = f_C^{(k)}$. By Hoeffding's inequality, with probability $1/(T|C|^d)$, we have

$$|\bar{Y}_\tau^{(k)} - \bar{f}_C^{(k)}| \ge \sqrt{\frac{\log(2T|C|^d)}{2\tau}}.$$

Applying union bound to get

$$P\left(\exists k \in I_{p(C)}, 0 \le \tau \le m_{C,i}^* : |\bar{Y}_\tau^{(k)} - f_C^{(k)}| \ge \sqrt{\frac{\log(2T|C|^d)}{2\tau}}\right) \le \frac{2m_{C,i}^*}{T|C|^d},$$

which completes the proof. □

Lemma 10. Fix $k \in \{1, -1\}$ and $C \in \mathcal{T}$, for any $x \in C$, one has

$$|\bar{f}_C^{(k)} - f^{(k)}(x)| \le c_0|C|^\beta,$$

where $c_0 = 2Ld^{\beta/2} + 1$.

Proof. For notation simplicity, we write f for $f^{(k)}$ in the following proof. By definition,

$$\begin{aligned}|\bar{f}_C - f(x)| &= |\frac{1}{P(C)}\int_C (f(y) - f(x))dP(y)| \\ &\le \frac{1}{P(C)}\int_C |f(y) - f(x)|dP(y) \\ &\le \frac{1}{P(C)}\int_C L\|x - y\|_2^\beta dP(y),\end{aligned}$$

where the first inequality uses the triangle inequality, and the second inequality is due to the smoothness condition. Since $x \in C$, we further have

$$\begin{aligned}|\bar{f}_C - f(x)| &\le \frac{1}{P(C)}\int_C L\|x - y\|_2^\beta dP(y) \\ &\le \frac{1}{P(C)}\int_C Ld^{\beta/2}|C|^\beta dP(y) \\ &\le c_0|C|^\beta.\end{aligned}$$

This completes the proof. □

C Proof of Theorem 4

As we argued after the statement of Theorem 4, one needs to set $t_1 \asymp T^{9/19}$, and $t_2 \asymp T^{15/19}$. Therefore, throughout the proof, we assume this is true and only focus on the number g of bins.

To construct a hard instance, we partition $[0,1]$ into z bins with equal width. Denote the bins by C_j for $j = 1, ..., z$, and let q_j be the center of C_j. Define a function $\varphi : [0,1] \to R$ as $\varphi(x) = (1 - |x|)1\{|x| \le 1\}$. Correspondingly define a function $\phi_j : [0,1] \to R$ as $\phi_j(x) = D_\varphi z^{-1}\varphi(2z(x - q_j))1\{x \in C_j\}$, where $D_\varphi = \min(2^{-1}L, 1/4)$. Define a function $f : [0,1] \to R$:

$$f(x) = \frac{1}{2} + \phi_1(x).$$

The problem instance of interest is $v = (f^{(1)}(x) = f(x), f^{(-1)}(x) = \frac{1}{2})$. It is easy to verify $v \in F(1,1)$. Throughout the proof, we condition on the event E specified by Lemma 7, which says the number of samples allocated to a bin concentrates well around its expectation. We will show even under this good event, there exists a choice of z that makes successive elimination fail to remove the suboptimal arms at the end of a batch with constant probability.

C.1 A helper lemma

We begin with presenting a helper lemma that will be used extensively in the later part of the proof. The claim is intuitive: if the sample size is small, it is not suficient to tell apart two Bernoulli distributions with similar means. Then, in our context, arm elimination will not occur.

Lemma 11. Assume $m_{B,i} \le 2\bar{m}_{B,i}$. For any $B \in [0,1]$ and $i \in \{1,2\}$. If $f_B^{(1)} - f_B^{(-1)} \le \delta \le 1/\sqrt{\bar{m}_{B,i}}$ for some $\delta > 0$, then

$$P\left(\bar{Y}_{B,i}^{(1)} - \bar{Y}_{B,i}^{(-1)} > U(m_{B,i}, T, B)\right) \le \frac{t_i}{T}.$$

Proof. Fix $0 < \tau \le \bar{m}_{B,i}$. Let $\bar{Y}_\tau^{(k)} = \sum_{l=1}^{\tau} Y_l^{(k)}/\tau$ where $Y_l^{(k)}$'s are i.i.d. random variables with $Y_l^{(k)} \in [0,1]$ and $E[Y_l^{(k)}] = f_B^{(k)}$ for $k \in \{1, -1\}$. Recall $U(\tau, T, B) = 4\sqrt{\frac{\log(2T|B|)}{\tau}}$[3]. Then,

$$P\left(Y_\tau^{(1)} - Y_\tau^{(-1)} > U(2\tau, T, B)\right) \le P^{(i)}\left(Y_\tau^{(1)} - Y_\tau^{(-1)} > \delta + \sqrt{\frac{\log(2T/g)}{2\tau}}\right)$$

[3] We remark the constant 4 is not essential for the proof to work. For any $c > 0$, $c\log(2T|B|) = \log((2T|B|)^c)$ so the final success probability is still tiny as long as T is suficiently large.

$$\overset{\text{(ii)}}{\leq} P\left(\bar{Y}_\tau^{(1)} - \bar{Y}_\tau^{(-1)} > f_B^{(1)} - f_B^{(-1)} + \sqrt{\frac{\log(2T/g)}{2\tau}}\right)$$

$$\overset{\text{(iii)}}{\leq} \frac{g}{T},$$

where step (i) is because $\delta \leq 1/\sqrt{m_{B,i}} \leq 1/\sqrt{\tau}$, step (ii) is due to $f_B^{(1)} - f_B^{(-1)} \leq \delta$, and step (iii) uses Hoeffding's inequality. Applying union bound to get

$$P\left(0 < \tau \leq m_B : \bar{Y}_\tau^{(1)} - \bar{Y}_\tau^{(-1)} > U(2\tau, T, B)\right) \leq \frac{m_{B,i}g}{T} \leq \frac{t_i}{T}.$$

This finishes the proof. □

C.2 Three failure cases for g

Fix some small constant $\epsilon > 0$ to be specified later. From now on, we use $\hat{\pi}$ to denote $\hat{\pi}_{\text{static}}$ for simplicity. We split the proof into three cases: (1) $g \geq T^{3/19+\epsilon}$; (2) $g \leq T^{3/19-\epsilon}$; (3) and $g \in (T^{3/19-\epsilon}, T^{3/19+\epsilon})$.

Case 1: $g \geq T^{3/19+\epsilon}$. Set $z = T^{3/19-\epsilon/2}$. Assume without loss of generality that $g = H \cdot z$ for some $H \geq 4$; see Figure 3 for an illustration of the instance. Suppose $C_1 = \cup_{l=1}^{H} B_l$, where B_l's are the bins produced by $\hat{\pi}$ that lie in C_1. It is clear that

$$\begin{aligned}
E[R_T(\hat{\pi})] &\overset{\text{(i)}}{\geq} E\left[\sum_{t=t_1+1}^{t_2} f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\right] \\
&\overset{\text{(ii)}}{=} E\left[\sum_{t=t_1+1}^{t_2} f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\ \mathbb{1}\{X_t \in C_1\}\right] \\
&\overset{\text{(iii)}}{\geq} \sum_{t=t_1+1}^{t_2} \sum_{l=H/4}^{3H/4} E\left[f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\ \mathbb{1}\{X_t \in B_l\}\right], \qquad (20)
\end{aligned}$$

where step (i) is because the total regret is greater than the regret incurred during the second batch, step (ii) uses the fact that under the instance ν, the mean rewards of the two arms differ only in C_1, and step (iii) arises since $C_1 = \cup_{l=1}^{H} B_l$. Now we turn to lower bounding $E\left[f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\ \mathbb{1}\{X_t \in B_l\}\right]$ for each $H/4 \leq l \leq 3H/4$.

Consider any such B_l. We drop the subscripts and write B instead for simplicity. By the design of ν, we have $f_B^{(1)} - f_B^{(-1)} \leq D_\varphi z^{-1} = \delta$, which obeys $D_\varphi z^{-1} \leq 1/\sqrt{m_{B,1}}$—a consequence of the choice of z. Additionally, we have $m_{B,1} \leq 2m_{B,1}$ under E. Therefore, we can invoke Lemma 11 to obtain

$$P\left(\bar{Y}_{B,1}^{(1)} - \bar{Y}_{B,1}^{(-1)} > U(m_{B,1}, T, B)\right) \leq \frac{t_1}{T} \leq \frac{1}{2}.$$

In words, with probability exceeding 1/2, no elimination will happen for the bin B. As a result, we obtain

$$\begin{aligned}
E[R_T(\hat{\pi})] &\geq \sum_{t=t_1+1}^{t_2} \sum_{l=H/4}^{3H/4} E\left[f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\ \mathbb{1}\{X_t \in B_l\}\right] \\
&\geq H \cdot \frac{t_2}{g} \cdot z^{-1} \geq \frac{t}{z^2} = T^{\frac{9}{19}+\epsilon},
\end{aligned}$$

where we have used the choice of z. So Theorem 4 holds with $\kappa = \epsilon$.

Case 2: $g \le T^{3/19-\epsilon}$. Set $z = T^{3/19-\epsilon/8}$. We have $g < z$ and there exists $H > 1$ such that $z = H \cdot g$; see Figure 4 for an illustration of the instance. Let B be the bin produced by $\hat{\pi}$ such that $C_1 \in B$. By the design of v, we have

$$\bar{f}_B^{(1)} - \bar{f}_B^{(-1)} \le \frac{1}{H}(1/2 + D_\varphi z^{-1}) + (1 - \frac{1}{H})\frac{1}{2} - \frac{1}{2} = \frac{D_\varphi z^{-1}}{H}.$$

Let $\delta = \frac{D_\varphi z^{-1}}{H}$, we have $\delta \le 1/\sqrt{m_{B,1}}$ due to our choice of z. Additionally, we have $m_{B,1} \le 2m_{B,1}$ under E. Therefore, we can invoke Lemma 11 to obtain

$$P\left(Y_{B,1}^{(1)} - Y_{B,1}^{(-1)} > U(m_{B,1}, T, B)\right) \le T^{-2}.$$

Thus, with probability exceeding 1/2, the suboptimal arm is not eliminated in B. Similar to the previous case, we obtain

$$\begin{aligned} E[R_T(\hat{\pi})] &\ge E\left[\sum_{t=t_2}^{t_2} f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\right] \\ &= E\left[\sum_{t=t_1+1}^{t_2} f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\ \mathbb{1}\{X_t \in C_1\}\right] \\ &\gtrsim \frac{t_2}{z^2} = T^{\frac{9}{19}+\frac{\epsilon}{4}}. \end{aligned}$$

So Theorem 4 holds with $\kappa = \epsilon/4$.

Case 3: $g \in (T^{3/19-\epsilon}, T^{3/19+\epsilon})$. Set $z \in T^{1/4}$. We then have $g < z$, as long as $\epsilon \le 1/19$. And there exists $H > 1$ such that $z = H \cdot g$; see Figure 4 for an illustration of the instance. Let B be the bin produced by $\hat{\pi}$ such that $C_1 \in B$. By the design of v, we have

$$\bar{f}_B^{(1)} - \bar{f}_B^{(-1)} \le \frac{1}{H}(1/2 + D_\varphi z^{-1}) + (1 - \frac{1}{H})\frac{1}{2} - \frac{1}{2} = \frac{D_\varphi z^{-1}}{H}.$$

Let $\delta = \frac{D_\varphi z^{-1}}{H}$, we have $\delta \le 1/\sqrt{m_{B,1}}$ due to our choice of z. Additionally, we have $m_{B,1} \le 2m_{B,1}$ under E. Therefore, we can invoke Lemma 11 to obtain

$$P\left(Y_{B,1}^{(1)} - Y_{B,1}^{(-1)} > U(m_{B,1}, T, B)\right) \le T^{-4}.$$

This means with probability at least 3/4, arm elimination does not occur in B after the first batch. Moreover, since $\delta \le 1/\sqrt{m_{B,2}}$ by the choice of z, and $m_{B,2} \le 2m_{B,2}$ under E, we can apply Lemma 11 again to get

$$P\left(Y_{B,2}^{(1)} - Y_{B,2}^{(-1)} > U(m_{B,2}, T, B)\right) \le T^{-4}.$$

In all, with probability at least 1/2, arm elimination does not occur in B after the second batch. Similar to before, we reach the conclusion that

$$\begin{aligned} E[R_T(\hat{\pi})] &\ge E\left[\sum_{t=t_2+1}^{T} f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t)\right] \\ &= E\left[\sum_{t=t_2+1}^{T} (f^*(X_t) - f^{\hat{\pi}_t(X_t)}(X_t))\mathbb{1}\{X_t \in C_1\}\right] \\ &\gtrsim \frac{T}{z^2} = T^{\frac{1}{2}}. \end{aligned}$$

We see that Theorem 4 holds with $\kappa = 1/38$.

D Proof of Theorem 5

When $\beta = d = 1$, we denote $\gamma(\alpha) = (\alpha + 1)/3$. Fix some small constant $\epsilon > 0$ to be specified later. We deal with $M = 2$ and $M = 3$ separately. In both cases, we use the fact that the algorithm needs to provide its first batch size t_1 prior to the game, and design α such that any choice of t_1 would fail.

D.1 When M = 2

Under $M = 2$, the theoretical optimal rate is $T^{(1-\gamma(\alpha))/(1-\gamma(\alpha)^2)} = T^{3/(\alpha+4)}$.

Case of $t_1 = T^{3/5+\epsilon}$. Take $\alpha = 1$. Since fixed grid and adaptive grid are the same when $M = 2$, by relation (9), we have

$$\sup_{f \in F(\alpha,\beta)} E[R_T(\pi)] \geq \max\left\{t_1, \frac{T}{t_1^{\frac{\alpha+1}{3}}}\right\} \geq t_1 = T^{\frac{3}{5}+\kappa_1},$$

where $\kappa_1 = \epsilon$.

Case of $t_1 = T^{3/4-\epsilon}$. Take $\alpha = o(1)$. By relation (9), we have

$$\sup_{f \in F(\alpha,\beta)} E[R_T(\pi)] \geq \max\left\{t_1, \frac{T}{t_1^{\frac{\alpha+1}{3}}}\right\} \geq \frac{T}{t_1^{\frac{\alpha+1}{3}}} = T^{1-(\frac{3}{4}-\epsilon)\frac{\alpha+1}{3}} = T^{\frac{3}{4}+\kappa_1},$$

where $\kappa_1 = (\alpha + 1)\epsilon/3 - \alpha/4 > 0$.

D.2 When M = 3

Under $M = 3$, the theoretical optimal rate is $T^{(1-\gamma(\alpha))/(1-\gamma(\alpha)^3)} = T^{9/((\alpha)^2+5\alpha+13)}$.

Case of $t_1 = T^{9/19+\epsilon}$. Take $\alpha = 1$. During the first batch, the learner can do no better than pull an arm uniformly at random. We can use the instance $(f_1(x) = 1, f_2(x) = 0)$ so that

$$\sup_{f \in F(\alpha,\beta)} E[R_T(\pi)] \geq t_1 = T^{\frac{9}{19}+\kappa_1},$$

where $\kappa_1 = \epsilon$.

Case of $t_1 = T^{9/13-\epsilon}$. Take $\alpha = o(1)$. Denote $T_2 = T^{1/(\gamma(\alpha)+1)} t_1^{\gamma(\alpha)/(\gamma(\alpha)+1)}$. Define the events $E_2 = \{T_2 < t_2\}$ and $E_3 = \{t_2 \leq T_2\}$. Recall $Q_i(\cdot) = \sum_{\omega \in \Omega_i} q_i(\omega) P_{\pi,i,\omega}(\cdot)$ from (13) given $z_i > 0$ and $s_i = z_i^{d-\alpha\beta}$. Take $z_2 = (36 t_1 2^2)^{1/(2\beta+d)}$ and $z_3 = (36 T_2 2^2)^{1/(2\beta+d)}$. Since E_2 can be determined by observations up to t_1, we have

$$|Q_2(E_2) - Q_3(E_2)| = |Q_2^{t_1}(E_2) - Q_3^{t_1}(E_2)| \leq TV(Q_2^{t_1}, Q_3^{t_1}) \overset{(i)}{\leq} \frac{1}{2}\sqrt{t_1 z_2^{-(2\beta+d)}} \overset{(ii)}{\leq} \frac{1}{4}, \tag{21}$$

where step (i) applies Lemma 2, and step (ii) is due to the definition of z_2. Consequently,

$$\begin{aligned} Q_2(E_2) + Q_3(E_3) &= Q_2(E_2) - Q_3(E_2) + Q_3(E_2) + Q_3(E_3) \\ &\geq -\frac{1}{4} + Q_3(E_2) + Q_3(E_3) = \frac{3}{4}, \end{aligned}$$

where the second step uses inequality (21) and the last equality is because E_2 and E_3 form a whole partition of the probability space. Then we would have at least one of $Q_2(E_2) \geq 1/4$ or $Q_3(E_3) \geq 1/4$. If $Q_2(E_2) \geq 1/4$, by Lemma 4 we obtain

$$\sup_{f \in F(\alpha,\beta)} E[R_T(\pi)] \geq \sup_{f \in C_{z_2}} R_{T_2}(\pi; f) \gtrsim T_2 z_2^{-\beta(1+\alpha)} \gtrsim T^{\frac{9}{13}+\kappa_1},$$

for some $\kappa_1 > 0$. If $Q_3(E_3) \geq 1/4$, we similarly have

$$\sup_{f \in F(\alpha, \beta)} E[R_T(\pi)] \geq \sup_{f \in C_{z_3}} R_{T_3}(\pi; f) \asymp T z_3^{-\beta(1+\alpha)} \asymp T^{-\frac{9}{13}+\kappa_1},$$

for some $\kappa_1 > 0$.