Self-Tuning Bandits over Unknown Covariate-Shifts

Joseph Suk   JS5338@COLUMBIA.EDU
Columbia University

Samory Kpotufe   SKK2175@COLUMBIA.EDU
Columbia University

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

Bandits with covariates, a.k.a. *contextual bandits*, address situations where optimal actions (or arms) at a given time t, depend on a *context* x_t, e.g., a new patient's medical history, a consumer's past purchases. While it is understood that the distribution of contexts might change over time, e.g., due to seasonalities, or deployment to new environments, the bulk of studies concern the most adversarial such changes, resulting in regret bounds that are often worst-case in nature.

Covariate-shift on the other hand has been considered in classification as a middle-ground formalism that can capture mild to relatively severe changes in distributions. We consider non-parametric bandits under such middle-ground scenarios, and derive new regret bounds that tightly capture a continuum of changes in context distribution. Furthermore, we show that these rates can be *adaptively* attained without knowledge of the time of shift (change point) nor the amount of shift.

1. Introduction

Bandits with covariates, or *contextual* bandits, concern situations where the reward of an action depends on a current context x, e.g., a patient's medical record (actions are treatments), or a user's profile and past history (actions are new products to propose). The problem is to maximize the total rewards of actions over time as similar contexts appear and rewards are observed over past actions. We consider the *stochastic* setting where covariates and rewards are jointly distributed at any time.

In the nonparametric version, little is assumed about the distribution of rewards over contexts, beyond *Lipschitz* conditions that capture the idea that rewards should be somewhat close for nearby contexts. Now suppose contexts are drawn from a *fixed* distribution; this ensures typicality, i.e., we can expect similar contexts to have appeared previously, so there is much potential to gradually learn the right actions for each context. Most recent advances have been made in nonparametric settings with a fixed distribution, with early consistency results in Yang et al. (2002), minimax rates in Rigollet and Zeevi (2010); Perchet and Rigollet (2013), followed by various refinements on the earlier algorithmic approaches (Reeve et al., 2018; Guan and Jiang, 2018).

However, it is understood that the distribution of contexts can change over time, e.g., seasonal changes in consumers' profile and purchases, or the extension of clinical trials to new populations. This reality has received little attention so far[1] in the nonparametric literature on the problem, although it has long been recognized in the more established parametric setting on contextual bandits, under various formalisms of often adversarial nature (Besbes et al., 2014; Hariri et al., 2015; Karnin and Anava, 2016; Luo et al., 2018; Liu et al., 2018; Wu et al., 2018; Chen et al., 2019; Chi Cheung et al., 2019).

---

1. We actually are not aware of any such work to date for the nonparametric setting.

As an initial take on a nonparametric setting with changes in distributions, we focus attention to the less adversarial case of *covariate-shift*, a formalism often adopted in works on *domain adaptation* in classification, starting with Sugiyama et al. (2008); Cortes et al. (2008); Gretton et al. (2009); Ben-David and Urner (2012). For example, consider a situation where clinical trials are to be extended to a new population, or where the makeup of the underlying population changes unbeknownst to the experimenter. While the distribution on patients' profiles X changes, the predictors in X (e.g., biometrics, medical history) are expected to remain predictive of treatment outcomes; in other words, the conditional distribution of rewards given X remains unchanged. Formally, in *covariate-shift*, $Q_{Y|X} = P_{Y|X}$ but $Q_X \neq P_X$, where P and Q denote previous and new joint-distributions on context-reward pairs (X, Y).

Algorithmic and Statistical Goals. We are interested in achievable regret in the time period corresponding to a fixed distribution Q_X over contexts, right after one or multiple unknown shifts in context distribution. Let P_X be a previous such distribution. Intuitively, performance under Q_X depends on *how far* P_X is from Q_X, since typical contexts under Q_X might not have been observed under P_X. In particular, prior distributions P_X can bias an algorithm towards choices that are suboptimal to performance under Q_X (see Remark 2 for biased choices under typical nonparametric approaches). We therefore aim for a procedure which *not only* can automatically recover from such bias, but can draw as much benefit as possible from past covariate distributions P_X, whenever they are not too far from Q_X. *As such, our analysis is more optimistic, and not only recovers the vanilla rates achievable without distributional shifts, but also integrates the information gained before the shift, resulting in faster rates.* In particular, our regret bounds tightly characterize the *effective amount of past experience* contributed to Q by previous runs on earlier distributions P's, in terms of both the length of previous runs and the discrepancies $P_X \rightarrow Q_X$.

We remark that the covariate-shift problem in this setting is harder than in classification where established approaches would compare observed $X \sim Q_X$ to prior observations $X \sim P_X$ to evaluate and adjust to the change. Here, however, we might not know the change point, and therefore cannot readily identify which data is which: for example, in an ongoing clinical study, or online recommender system, seasonal shifts in population makeup are unlikely to be known a priori. Interestingly, while one might then try and detect such change points, we show that this is not necessary. Our proposed procedure automatically *adapts* to unknown change points, along with unknown levels of change in covariate distributions, while achieving regrets of near optimal order for the setting.

Capturing distributional changes. In order to admit a more natural range of covariate distributions across changes, we had to significantly relax the usual distributional conditions in prior work on nonparametric bandits. Namely, a so-called *strong-density* condition was always assumed, which roughly states that the density of covariates is nearly uniform on the support and greatly simplifies the analysis with respect to margin conditions on rewards (see Remark 1). We instead only assume that distributions are supported inside $[0, 1]^D$, yet are able to properly capture the benefits of margins in rewards through careful integration arguments. *Such new arguments are likely of independent interest even in the vanilla bandit settings with no distributional shift.*

Finally the relation between distributions across runs is captured through a so-called *transfer-exponent* adapted from earlier work on covariate-shift in classification (Kpotufe and Martinet, 2018).

Other Related Work. There is by now an expansive literature on *parametric*[2] contextual bandits, ranging from fully adversarial to stochastic settings (see e.g., Hazan and Megiddo, 2007; Langford and Zhang, 2008; Bubeck and Cesa-Bianchi, 2012; Auer and Chiang, 2016; Rakhlin and Sridharan, 2016).

In the stochastic parametric setting, the earlier cited results (Besbes et al., 2014; Hariri et al., 2015; Karnin and Anava, 2016; Luo et al., 2018; Liu et al., 2018; Wu et al., 2018; Chen et al., 2019; Chi Cheung et al., 2019) are closest in spirit to the present work. However, they consider settings of a more adversarial nature, as their aim is to achieve regrets – over stationary periods of length Δ_t – of similar order $O(\sqrt{\Delta_t})$ as would have been achieved without distribution shifts; in other words, they contend that past experiences could adversarially affect regret over stationary periods, and the aim is to mitigate such adversity. In contrast, as we will see, past experience is actually useful under covariate-shift (to a variable extent depending on shift characteristics), as long as the bandits procedure is reasonably conservative in least-observed regions of context space. In a similar vein, Azar et al. (2013) considers a non-contextual setting where it is possible to benefit from previous experience. Moreover, the work assumes knowledge of the time of shift, whereas we do not.

In recent works on *non-stationary* bandits (Besbes et al., 2014; Luo et al., 2018; Chen et al., 2019; Chi Cheung et al., 2019), regret is expressed either in terms of the number of shifts or in terms of a budget on the total variation between subsequent distributions. The latter is somewhat more natural, as it captures a total shift in distribution rather than the number of shifts. In a similar natural way, our regret bounds are expressed over a total shift, as captured by an *aggregate transfer-exponent*, and thus do not worsen with the unknown number of shifts in distribution (see Theorem 2 of Appendix C, while for ease of presentation, Theorem 1 of the main text covers the case of a single unknown shift).

Slivkins (2014) considers nonparametric settings with Lipschitz rewards, however, with possibly adversarial non-stochastic contexts. Given adversarial assumptions on past contexts, the work requires more conservative procedures than ours (see e.g., experimental comparison in Appendix A). In particular, their regret bounds do not account for the benefits of *margins in rewards between arms*, which can significantly reduce achievable regrets. Nonetheless, their bounds are expressed in terms of notions of metric dimensions which can be tighter than the notion of *box-dimension* employed in the present work; however, their resulting bounds are only clearly tighter than ours in the regime without margin ($\alpha = 0$ in our notation) since otherwise (e.g., for sufficient margin $\alpha \geq d$) our rates' exponent changes to $1/2$, i.e. no longer depend exponentially on dimension, which they cannot avoid without considering noise margin.

Finally, in the setting of active online regression with multiple domains, Chen et al. (2020) establishes adaptive regret guarantees in terms of the domain dimensions and durations.

We start with a formal setup in Section 2, followed by an overview of results in Section 3. Algorithms and proof ideas are discussed in Section 4.

2. Setup

2.1. Bandits With Covariate-Shift

We consider a finite set of actions (or arms) $[K] \doteq \{1, 2 \ldots, K\}$, and let $Y \in [0, 1]^K$ denote the rewards of each action $i \in [K]$. We assume that the covariate X, lying in $\mathcal{X} \doteq [0, 1]^D$, is jointly

---

2. The term characterizes settings displaying $O(\sqrt{n})$ regrets, due to either parametric constraints on rewards or on *hindsight* baseline policies.

distributed with Y, and we therefore assume a random independent sequence[3] of covariate-reward pairs $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$, *identically distributed* over *different* and possibly unknown periods of time[4].

Single-shift vs Multiple shifts. To simplify presentation, in the main part of the paper we focus on the case of a single shift in distribution, and analyze the case of multiple shifts in the appendix. The discussion loses little generality as the multiple shift case follows easily as shown in Appendix C.

In the simplest case of a single shift, we assume that for some $n_P \geq 0$, possibly a priori unknown, the sequence $\{(X_t, Y_t)\}_{t\in[n_p]}$ is i.i.d. according to a distribution P, while $\{(X_t, Y_t)\}_{t>n_p}$ is i.i.d. according to a *new* distribution Q with different marginals. We will be interested in performance under Q, i.e., after such a shift.

Assumption 1 (Covariate shift) *While the distribution of covariates X_t might change overtime, the conditional distribution of $Y_t \mid X_t$ remains fixed (i.e., in our context $Q_X \neq P_X$, but $P_{Y|X} = Q_{Y|X}$). In particular, the aim is to* maximize *expected rewards conditioned on X_t; this is captured through the* **fixed regression function** $f : \mathcal{X} \to [0,1]^K$ as $f^i(x) \doteq \mathbb{E}(Y^i | X = x)$, $i \in [K]$.

In the bandits setting, a so-called *policy*[5] (or bandit procedure) chooses actions at each round t, based on observed covariates (up to round t) and passed rewards, whereby at each round t only the rewards Y_t^i of chosen actions i are revealed. We say an arm i is pulled if action i is chosen by the policy. We adopt the following formalism.

Definition 1 (Policy) *A policy $\pi \doteq \{\pi_t\}_{t\in\mathbb{N}}$ is a random sequence of functions $\pi_t : \mathcal{X}^t \times [K]^{t-1} \times [0,1]^{t-1} \to [K]$. In an abuse of notation, in the context of a sequence of observations till round t, we will let $\pi_t \in [K]$ also denote the action chosen at round t. In the case of a **randomized** policy, i.e., where π_t in fact maps to distributions on $[K]$, we will still let $\pi_t \in [K]$ denote the (random) action chosen at round t.*

We let $\mathbf{X}_t \doteq \{X_s\}_{s\leq t}, \mathbf{Y}_t \doteq \{Y_s\}_{s\leq t}$ denote the observed covariates and (observed and unobserved) rewards from rounds 1 to t. The performance of a policy is evaluated as follows (visualized in Figure 1).

Definition 2 (Cumulative regret) *Define the regret between rounds $n_P < n$ of a policy π, as*

$$\boldsymbol{R}_{n_P,n}(\pi) \doteq \sum_{t=n_P+1}^{n} \max_{i\in[K]} \left(f^i(X_t) - f^{\pi_t}(X_t) \right).$$

In our context of a shift to Q, we often will use the short notation $\boldsymbol{R}_n^Q(\pi)$ to denote $\boldsymbol{R}_{n_P,n}(\pi)$.

The oracle policy π^* refers to the strategy that maximizes the expected reward at any round t, and is given by $\pi_t^*(X_t) \in \mathrm{argmax}_{i\in[K]} f^i(X_t)$. The regret of a policy π is therefore the *excess* expected reward of π^* relative to π over $\mathbf{X}_n$. We seek a policy π that minimizes $\mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n} \boldsymbol{R}_n^Q(\pi)$.

We emphasize that, while we will be interested in regret over particular periods $n_P + 1 : n$ (corresponding to a fixed Q), it is understood by definition that π runs starting at $t = 1$, and $\boldsymbol{R}_n^Q(\pi) \doteq \mathbf{R}_{n_P,n}(\pi)$ therefore depends on prior decisions up till time n_P. Finally, usual bounds for stationary distributions are recovered simply by letting $n_P = 0$.

---

3. The indexing set $\mathbb{N}$ denotes the natural numbers excluding 0.

4. We use the terms *time t* or *round t* interchangeably, the latter in the context of a procedure.

5. We remark that the term *policy* is often used to denote a mapping from state (or covariate) to action; here we simply equate it with any decision procedure taking action based on current and past observations.
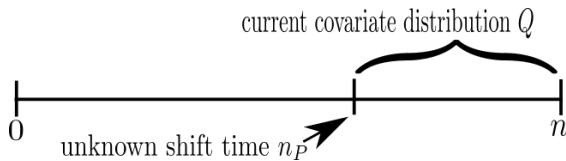
Fig. 1: We are interested in the regret $\mathbf{R}_n^Q(\pi)$, w.r.t. a static distribution phase Q, restricted to the corresponding rounds.

2.2. Nonparametric Setting

Our main assumptions below are stated under the ℓ_∞ norm on $[0,1]^D$ for convenience, as we build our procedures π over regular grids of $[0,1]^D$. It should be clear however that the relevant conditions hold under any norm (e.g., any ℓ_p, $p \geq 1$) when they hold under ℓ_∞, by the equivalence of $\mathbb{R}^d$ norms.

● **Standard Assumptions and Conditions.** We assume, as in prior work on nonparametric contextual bandits (Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013; Slivkins, 2014; Reeve et al., 2018; Guan and Jiang, 2018), that the regression function is Lipschitz, with some known upper-bound λ on the Lipschitz constant (often simply assumed to be 1).

Assumption 2 (Lipschitz f) *There exists $\lambda > 0$ such that for all $i \in [K]$ and $x, x' \in \mathcal{X}$,*

$$|f^i(x) - f^i(x')| \leq \lambda \|x - x'\|_\infty. \tag{1}$$

Furthermore, the difficulty of detecting the optimal arm $\pi^*(x)$ at any x is parametrized through the following *margin* condition of f w.r.t. Q, originally due to Tsybakov et al. (2004) (for nonparametric classification).

Definition 3 (Margin Condition) *Let $f^{(1)}(x)$, $f^{(2)}(x)$ denote the highest and second highest values of $f^i(x), i \in [K]$, if they are not all equal; otherwise let $f^{(1)}(x) = f^{(2)}(x)$ be that value. There exists $\delta_0 > 0, C_\alpha > 0$ so that $\forall \delta \in [0, \delta_0]$,*

$$Q_X(0 < |f^{(1)}(X) - f^{(2)}(X)| \leq \delta) \leq C_\alpha \delta^\alpha. \tag{2}$$

In particular, the above is always satisfied with at least $\alpha = 0$. Intuitively, the larger the margin $f^{(1)}(x) - f^{(2)}(x)$ at x, the easier it is to detect the best arm, in the sense that a rough approximation to f is sufficient. The above condition, common in prior work on nonparametric bandits, encodes the margin distribution under Q_X. Interestingly, we need no assumption on the margin distribution under P_X, although our setting assumes that the procedure π is first ran on covariates $X_t \sim P_X, t \leq n_P$; in fact, we will see that we only need to ensure that π maintains good choices of arms for every potential $x \in \mathcal{X}$, along with sufficient arm pulls, up till the distribution shifts at round $n_P + 1$.

● **A Relaxed Distributional Condition.** In this work we only assume that marginals P_X, Q_X are supported in $[0,1]^D$. The following definition then serves to capture the *complexity* of a support. The quantity d therein, called a *box dimension*, can be viewed as capturing the intrinsic dimension of the support $\mathcal{X}_Q$: low-dimensional supports intersect fewer cells of a partition of $[0,1]^D$. We need no direct condition on past distributions P_X, as we simply need to capture their discrepancy from Q_X.
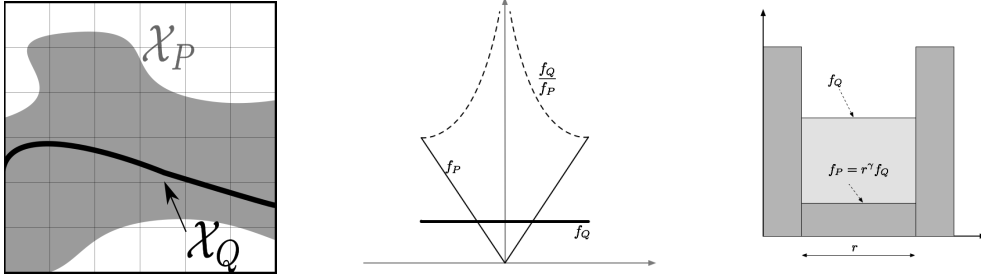
5

Fig. 2: Some settings with $0 < \gamma < \infty$. Left: Q_X is of smaller dimension d, while P_X is of higher dimension d_P (with gray support); γ can then be shown to be $(d_P - d)$ under mild regularity. Center: the density $f_P \propto |x|^\gamma$ goes fast to 0, while f_Q is uniform; f_Q/f_P then diverges (so *density ratios*, and *f-divergences* are ill defined). Right: P_X moves mass away from regions of large Q_X mass, with relative densities captured by γ and the size of the region (r).

Definition 4 (Support complexity) *For $r \in \left\{ 2^{-i} : i \in \mathbb{N} \right\}$, let $\mathcal{P}_r$ denote the regular partition of $[0,1]^D$ into hypercubes of side length r. Denote by $\mathcal{G}(\mathcal{X}_Q, r)$ the number of cells of $\mathcal{P}_r$ which intersect $\mathcal{X}_Q$. We say that $\mathcal{X}_Q$ has (C_d, d) **box dimension**, for $d \geq 1$, $C_d \geq 1$, if $\forall r \in (0,1]$, $\mathcal{G}(\mathcal{X}_Q, r) \leq C_d \cdot r^{-d}$.*

The condition clearly holds for at least $d = D$, and would imply faster rates when $d \ll D$. It has been employed as a measure of the complexity of a data space, and we refer the reader, e.g., to Scott and Nowak (2006); Clarkson (2006) for detailed expositions.

   We emphasize that we do not assume knowledge of such support complexity d.

Remark 1 (Strong Density Condition) *Prior cited work on nonparametric contextual bandits invariably assumed that Q_X is **nearly** uniform on its support, namely that for any ball B of radius r, $Q_X(B) \gtrsim r^d$ for some d ($d = D$ is usual). This in particular ensures that, **uniformly over the data space**, any small region receives a sufficient amount of covariates X's for good estimation of the expected reward f at candidate arms; this simplifies both algorithmic design and analysis. This is not the case here, and we therefore require refined algorithmic choices and integration arguments.*

• **Quantifying the Shift from P to Q.** Next we aim to quantify how much the earlier covariate distribution P_X differs from the shift Q_X. Intuitively, P_X has information on Q_X if it yields data useful to Q_X, in other words, if it has sufficient mass in regions of large Q_X mass. The next condition, adapted from recent work (Kpotufe and Martinet, 2018) on classification, parametrizes such intuition.

Definition 5 *We call $\gamma \geq 0$ a **transfer exponent** between P_X and Q_X, if $\exists\, C_\gamma \geq 0$ such that, for all ℓ_∞ balls $B \subset [0,1]^D$ of diameter $r \in (0,1]$, we have $P_X(B) \geq C_\gamma \cdot r^\gamma \cdot Q_X(B)$.*

   Note that the above condition always holds with at least $\gamma = \infty$ (this occurs, e.g., when Q_X has mass outside P_X's support). The larger the shift, the larger γ, with $\gamma = 0$ capturing the mildest such shifts in covariate distribution. Some examples, borrowed from Kpotufe and Martinet (2018), are given in Figure 2. As we will see, the transfer exponent γ manages to tightly capture a continuum of easy to hard shifts in covariate distributions as evident in achievable regret rates $\mathbf{R}_n^Q$ over the period corresponding to the current distribution Q_X.

3. Results Overview

A common algorithmic approach in nonparametric contextual bandits, starting from earlier work (Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013), is to maintain *tree*-based (regression) estimates $\hat{f}_t$ of the expected reward function f, so that at any time t, upon observing X_t, only those arms i with $f^i(X_t)$ close to $f^{(1)}(X_t)$ might be played. This assumes a good estimate of f at any time t, which in the context of tree-based estimates boils down to choosing an optimal *level* in the tree – where each level r corresponds to a piecewise-constant regression estimate $\hat{f}_t$ over bins of side-length r in $\mathcal{X}$.

Remark 2 (Crucial algorithmic choices can be biased) *Assuming a **strong density** condition on Q_X (where balls of radius r have mass of order r^d for some d), in the usual setting with a stationary distribution Q, an optimal level $r = r_t$ is chosen as $O(t^{-1/(2+d)})$ yielding optimal regression that would result in the best provable regret rates (Rigollet and Zeevi, 2010). However, suppose in our context with distribution drift that covariates were at first distributed as P_X, forcing a different choice of level $O(t^{-1/(2+d_P)})$, i.e., under a different covariate dimension d_P. This would be suboptimal under Q_X, which we wish to quickly detect.*

Furthermore, under our relaxed conditions on Q_X, it turns out that the optimal oracle choice of level r_t (even for a fixed distribution) is of the form $t^{-1/(2+\alpha+d)}$ – i.e., higher levels in terms of *unknown* margin parameter α – so as to ensure sufficient samples near decision boundaries.

The right choice of level is further complicated in context with distributional shift: if we want to optimally benefit from past observations under P, after the shift to Q, the optimal choice of level has to also account for the discrepancy between P and Q (as captured here by γ). As it turns out, a first analysis (not shown, but implicit in our arguments) reveals that the optimal choice of level is then of the form $r_t(\gamma, n_P) \doteq O(\min\{n_P^{-1/(2+\alpha+d+\gamma)}, (t - n_P)^{-1/(2+\alpha+d)}\})$, i.e., now also depends on *unknown* discrepancy level γ, and potentially *unknown* change point n_P. A main aim is therefore to design a procedure which, without such knowledge, still makes near optimal adaptive choices of levels at any time t.

Our adaptive strategies, detailed in Section 4, rely directly on the relative proportions of samples observed on a path from the root of a tree T down to a leaf containing X_t. Roughly, let $n_r(X_t)$ denote the covariate count in the bin containing X_t at level r (by time t) locally at X_t. We then choose, roughly, the smallest level r such that $n_r^{-1}(X_t) \leq r^2$. For intuition, this choice roughly balances regression variance (controlled by n_r^{-1}) and bias (controlled by r). Such a choice stems from prior insights on adaptive tree-based regression with fixed data distribution but unknown d (see e.g. Kpotufe and Dasgupta, 2012), which we show here to yield a regression rate similar to that of the oracle choice $r_t(\gamma, n_P)$. As a result, we can automatically adapt to unknown problem parameters d, γ, α, n_P while the algorithm remains agnostic to underlying shifts in covariate distribution.

However, such an adaptive choice introduces nontrivial book-keeping issues which complicate the analysis. Namely, the number of observed rewards for a given arm i – which drive the estimates $\hat{f}$ – might significantly differ from the number of covariates $n_r(X_t)$ in a bin, as we eliminate suboptimal arms over time. Further care is thus required for such book-keeping on observed rewards (or *arm pulls*).

Adaptive Bandits. Our main theorem considers an adaptive policy π given by the randomized procedure of Algorithm 1 from Section 4, which operates with a *confidence parameter* $\delta \in (0, 1)$. While π is randomized to simplify choices of arms to pull, we note that a deterministic variant with similar guarantees is feasible (see Remark 7).

As previously mentioned, the first theorem below is stated under a single change in distribution $P_X \to Q_X$, for simplicity of presentation, while a similar result for multiple (unknown) changes in distribution is given in Appendix C.

Just as in previous work for the stationary case, the margin parameter α needs not be known, while in addition here, we do not need to know the drift parameters n_P, γ, nor the dimension d of Q_X either. The expectation in the statement below is over the entire sequence $\mathbf{X}_n, \mathbf{Y}_n \sim P^{n_P} \times Q^{n-n_P}$, plus the randomness in π.

Theorem 1 *Let π denote the procedure of Algorithm 1, ran, with parameter $\delta \in (0, 1)$, up till time $n > n_P \geq 0$, with the change point n_P possibly unknown. Suppose P_X has unknown transfer exponent γ w.r.t. Q_X, that Q_X has (C_d, d) box dimension, and that the average reward function f satisfies a margin condition with unknown α under Q_X. Let $n_Q \doteq n - n_P$ denote the (possibly unknown) number of rounds after the drift, i.e., over the phase $X_t \sim Q_X$. We have for some $C > 0$:*

$$\mathbb{E} \, \boldsymbol{R}_n^Q(\pi) \leq C n_Q \left[\min \left(\left(\frac{K \log (K/\delta)}{n_P} \right)^{\frac{\alpha+1}{2+\alpha+d+\gamma}}, \left(\frac{K \log (K/\delta)}{n_Q} \right)^{\frac{\alpha+1}{2+\alpha+d}} \right) + \frac{K \log (K/\delta)}{n_Q} + n\delta \right]$$

The following corollary is immediate.

Corollary 1 *Under the setup of Theorem 1, letting $\delta = O(1/n^2)$ yields:*

$$\mathbb{E} \, \boldsymbol{R}_n^Q(\pi) \leq C n_Q \left[\min \left(\left(\frac{K \log(Kn)}{n_P} \right)^{\frac{\alpha+1}{2+\alpha+d+\gamma}}, \left(\frac{K \log(Kn)}{n_Q} \right)^{\frac{\alpha+1}{2+\alpha+d}} \right) + \frac{K \log(Kn)}{n_Q} \right].$$

The above rates interpolate between two terms: one involving n_P past observations and the drift parameter γ, the other involving n_Q. This second term is of the form $n_Q^{1-\frac{\alpha+1}{2+\alpha+d}}$ and is attained by the adaptive π when there is no drift, i.e., for $n_P = 0$. Note that, under the *strong density* assumptions of Rigollet and Zeevi (2010); Perchet and Rigollet (2013), the regret would have been $n_Q^{1-\frac{\alpha+1}{2+d}}$, i.e., smaller due to the easier setting (our rates for the same algorithm would in fact be of the same order under the strong density assumption; see Theorem 5 of Appendix D).

The regret of $n_Q^{1-\frac{\alpha+1}{2+\alpha+d}}$, for $n_P = 0$, is however tight under our relaxed assumptions on Q_X: this becomes evident by noticing that the corresponding average regret of $n_Q^{-\frac{\alpha+1}{2+\alpha+d}}$ matches minimax lower-bounds of Audibert and Tsybakov (2007) for classification under equivalent nonparametric conditions (see Remark 4 below).

For $n_P > 0$, the regret, interpolating both terms, can be rewritten as $n_Q \cdot \left(n_P^{d_\gamma} + n_Q \right)^{-1 \wedge \frac{\alpha+1}{2+\alpha+d}}$ for $d_\gamma = (2 + \alpha + d)/(2 + \alpha + d + \gamma)$; in other words $n_P^{d_\gamma}$ can be viewed as the effective amount of *past experience* contributed despite the drift; the quantity $n_P^{d_\gamma}$ is largest when $\gamma = 0$, lowering regret, and vanishes as $\gamma \to \infty$, i.e., with larger discrepancy between P_X and Q_X. At $\gamma = \infty$, e.g., when Q_X has sizable mass outside the support of P_X, past experience under P_X can only improve constants in the regret under Q, as the rate defaults to what it would have been under Q without distributional shift. Such intuition on adapting to increasing γ is confirmed in simulations (Figure 3).

Corollary 2 (total regret vs. R_n^Q) *Under the setup of Theorem 1, suppose also that the support of P_X, $\mathcal{X}_P$, has (C_{d_P}, d_P) box dimension and that the reward function f additionally satisfies a margin*
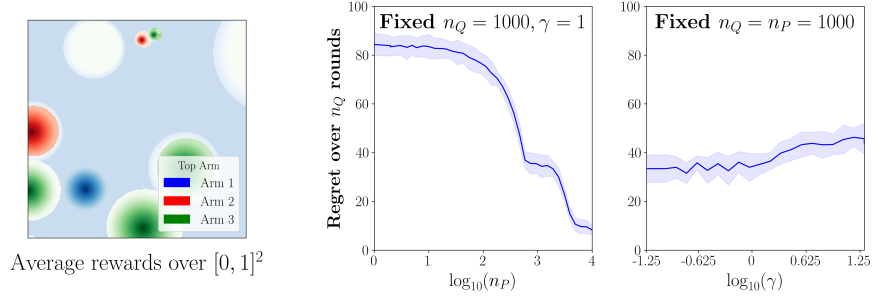
Fig. 3: Simulation Results. $Q_X \sim \mathcal{U}([0,1]^2)$, P_X has density $\propto \|x\|_2^\gamma$, $K = 3$ arms, with rewards $Y^i = f^i(X) + \mathcal{N}(0, .05)$, $i \in [K]$, where $f^i(x) \propto \sum_k \omega_{i,k} (1 - \|x - z_k\|_2 / r_k)_+$ for 25 randomly placed *bumps* with centers z_k, Rademacher signs $\omega_{i,k} \in \{\pm 1\}$, and radius r_k. A profile of f is shown on the left, with lower gradient colors corresponding to least margins (white meaning *no* margin). The right plots average 20 runs of Algorithm 1, and verify the guarantees of Theorem 1, namely that the procedure *adapts* to unknown shift parameters n_p and γ. In particular, the amount of *past experience* $n_P^{d_\gamma}$ clearly helps, and how much it helps depends on the level of shift $P \to Q$ as captured by γ. We also compare Algorithm 1 to an *adversarial* nonparametric contextual bandits algorithm from Slivkins (2014) in Appendix A.

condition with unknown α under P_X. Then, letting $\delta = O(1/n^2)$ in Algorithm 1, the **total regret** $\boldsymbol{R}_{1,n}(\pi)$ *is bounded as follows:*

$$\mathbb{E}\,\boldsymbol{R}_{1,n}(\pi) \leq C \left[K \log(Kn) + n_P \left(\frac{K \log(Kn)}{n_P} \right)^{\frac{\alpha+1}{2+\alpha+d_P}} \right.$$

$$\left. + n_Q \min \left(\left(\frac{K \log(Kn)}{n_P} \right)^{\frac{\alpha+1}{2+\alpha+d+\gamma}}, \left(\frac{K \log(Kn)}{n_Q} \right)^{\frac{\alpha+1}{2+\alpha+d}} \right) \right]$$

In contrast to the bound on the regret $\mathbf{R}_n^Q(\pi)$ considered in Theorem 1, the above result is directly comparable to the total regret bounds typically considered in the literature. We note that a total regret bound can always be recovered from a bound on $\boldsymbol{R}_n^Q(\pi)$ by setting $n_P = 0$. However, it is not true that a bound on the total regret necessarily implies a bound on the regret after the time of shift.

Remark 3 (Multiple shifts) *As previously mentioned, the results readily extend to the case of multiple changes in distribution before time n_P, with γ above replaced by a weighted average $\bar{\gamma}$ of transfer exponents between past P_X's and the current Q_X (see Theorem 2 of Appendix C).*

Remark 4 (Lower Bounds) *Finally, we note that the above rates are tight (up to $\log$ terms) in the sense that the average regret $\left(n_P^{d_\gamma} + n_Q \right)^{-\frac{\alpha+1}{2+\alpha+d}}$ matches minimax lower bounds for classification under covariate-shift of (Kpotufe and Martinet, 2018). Formally, by a simple reduction via **online-to-batch** conversion, the contextual bandit problem is at least as hard as its classification counterpart (see discussion and Corollary 6 in Appendix D).*

4. Algorithms and Analysis Overview

4.1. Algorithm Overview

All algorithms build on a dyadic partitioning tree T defined as follows.

Algorithm 1 Adaptive Bandits

1: **Requires**: upper bound on Lipschitz constant $\lambda \geq 1$, set of arms $[K]$, tree T with levels $r \in \mathcal{R}$
2: **Input Parameters**: $\delta \in (0, 1)$, covariates $X_1, X_2, \ldots$
3: **Initialization**: For any bin B at any level in T, set $\mathcal{I}_B \leftarrow [K]$ # Set of candidate arms in B.
4: **for** $t = 1, 2, \ldots$ **do**
5:    If $t \leq \lceil 8K \log(K/\delta) \rceil$, play a random arm $i \in [K]$ selected with probability $1/K$.

6:    Otherwise, for $t > \lceil 8K \log(K/\delta) \rceil$:

7:       **Choose a level $r_t \in \mathcal{R}$ for X_t:** $r_t \leftarrow \min \left\{ r \in \mathcal{R} : r \geq \sqrt{\frac{8K \log(K/\delta)}{n_r(X_t)}} \right\}$

8:       **Update candidate arms for the bin B containing X_t at level r_t:**
9:       Set $\mathcal{I}_B \leftarrow \bigcap_{B' \in T, B \subseteq B'} \mathcal{I}_{B'}$ # Discard arms previously discarded by ancestor bins

10:       Compute $\hat{f}^i(B)$ for any $i \in \mathcal{I}_B$ over B # See Definition 7 for estimate $\hat{f}^i$

11:       Refine candidate arms: $\mathcal{I}_B \leftarrow \mathcal{I}_B \setminus \{i : \hat{f}^i(B) < \hat{f}^{(1)}(B) - 8\lambda r_t\}$.

12:       **Play a random arm $i \in \mathcal{I}_B$ selected with probability $1/|\mathcal{I}_B|$.**
13: **end for**

Definition 6 (Partition Tree) *Let $\mathcal{R} \doteq \{2^{-i} : i \in \mathbb{N} \cup \{0\}\}$, and let $T_r, r \in \mathcal{R}$ denote a regular partition of $[0, 1]^D$ into hypercubes (which we refer to as **bins**) of side length (a.k.a. bin size) r. We then define the dyadic **tree** $T \doteq \{T_r\}_{r \in \mathcal{R}}$, i.e., a hierarchy of nested partitions of $[0, 1]^D$. We will refer to the **level** r of T as the collection of bins in partition T_r. The **parent** of a bin $B \in T_r, r < 1$ is the bin $B' \in T_{2r}$ containing B; **child**, **ancestor** and **descendant** relations follow naturally. The notation $T_r(x)$ will then refer to the bin at level r containing x.*

Note that, while in the above definition, T has infinite levels $r \in \mathcal{R}$, at any round t in a procedure, we implicitly only operate on the subset of T containing data. Our procedures, as in prior work on nonparametric bandits, maintain estimates $\hat{f}$ of the average reward function f over levels of T.

Definition 7 (Regression estimates and arm pull counts) *At any round $t > \lceil 8K \log(K/\delta) \rceil$, for any bin B at any level in the tree, we define the following regression estimate for arm i:*

$$\hat{f}_t^i(B) \doteq \frac{1}{m_t(B, i)} \sum_{X_s \in B, s \leq t-1, \pi_s = i} Y_s^i,$$

where $m_t(B, i)$ denotes the number of times arm i was pulled in B before time t. If $m_t(B, i) = 0$, we take $\hat{f}_t^i(B) = 0$. For any B at level r in the tree, $\hat{f}_t^i(B)$ serves as a regression estimate for any covariate $x \in B$. We often drop B or t in the above definitions, when understood from context.

Definition 8 (Covariate counts) *Let $B \doteq T_r(X_t)$. We write: $n_r(X_t) \doteq \sum_{s \in [t-1]} \mathbb{1}\{X_s \in B\}$.*

At any round t, upon observing X_t, a level r_t is chosen according to the *covariate counts* $n_r(X_t)$ along the path $\{T_r(X_t)\}_{r \in \mathcal{R}}$. Roughly, r_t is picked as the smallest $r \in \mathcal{R}$ such that $1/\sqrt{n_r(X_t)} \leq r$. The level r_t, more precisely the bin $B = T_{r_t}(X_t)$ containing X_t at that level, then determines the estimate $\hat{f}(B)$ to be used at time t.

To understand this choice, recall that a main aim is to quickly identify which arms are suboptimal – and shouldn't be pulled for X_t, and we ought to therefore use a good estimate of $f(X_t)$. We

will show that r_t indeed provides such an estimate at a near optimal regression rate in terms of unknown n_P and γ (see Lemma 6). In particular, the covariate counts $n_r(X_t)$, at any level r, account for covariates from both P_X and Q_X, whenever $t > n_P$. Intuitively, we will then expect $n_r(X_t) \approx n_P \cdot P_X(T_r(X_t)) + (t - 1 - n_P) \cdot Q_X(T_r(X_t)) \gtrsim n_P \cdot r^{d+\gamma} + (t - 1 - n_P) \cdot r^d$. The choice of r_t can then be shown to properly balance regression variance and bias in terms of unknown n_P and γ.

Once this choice is made, only those arms deemed safe are pulled for X_t at time t. These so-called *candidate arms* are maintained as $\mathcal{I}_B \subseteq [K]$ for each bin B over time, and exclude identified suboptimal arms whose average rewards are clearly below that of the unknown best arm $\pi^*(x)$ for any $x \in B$. In particular, suppose at any time t, we can ensure that $|\hat{f}_t^i(B) - f^i(x)| \lesssim r_t$ for all remaining arms i over $x \in B$. Then we can safely discard i if $\hat{f}_t^{(1)}(B) - \hat{f}_t^i(B) \gtrsim r_t$. It then makes sense to also discard such an arm in all descendants of B.

Remark 5 (Margin Adaptation) *Adaptation to the unknown margin parameter α comes through such decisions over $\mathcal{I}_B$. Namely, if the margin $f^{(1)}(x) - f^{(2)}(x) \gg r_t$ for all $x \in B$, then all suboptimal arms are discarded by time t so we suffer no regret for $X_t \in B$ at time t. Otherwise, all arms i left in $\mathcal{I}_B$ satisfy $f^{(1)}(x) - f^i(x) \lesssim r_t$, for $x \in B$, i.e., a bound on regret; on the other hand, the margin distribution ensures that the Q_X-probability of X_t landing in such a bin with low margins is small (if r_t were not a function of X_t). The main difficulty is to ensure that r_t is of the right order in terms of t and the unknown n_P, γ, even though it can significantly differ over the spatial location of X_t (as we have minimal assumptions on covariate distributions – see Section 4.2 below).*

Remark 6 (Book-keeping) *As discussed earlier, our adaptive choice of level r_t brings in additional difficulty in the book-keeping of arm pulls. In fact, the above discussion assumes that covariate counts $n_r(X_t)$ (used in choosing r_t, towards adapting to unknown n_P, γ) and arm-pull counts $m_t(B, i)$ (used in estimating $\hat{f}_t(B)$) are of similar order. However, this needs not be the case for a couple reasons. First, a single arm i is pulled whenever a covariate X_t lands in a bin B so that $m_t(B, j)$ is not updated for $j \neq i$ even though covariate counts for B are. Second, more nuanced, the following situation can happen since we can have $r_t > r_{t-1}$ as X_t, X_{t-1} fall in different regions of space: in a given bin $B \doteq T_{r_t}(X_t)$ chosen at time t, some arms in $\mathcal{I}_B$ might have been eliminated in a descendant of B at an earlier time, and therefore not pulled as much as other arms in $\mathcal{I}_B$. Such situations do not arise under more common global choice of $r_t \leq r_{t-1}$, i.e., where choices of levels are monotonic in time.*

It turns out that, in fact, enough regularity is built into our adaptive choice to mitigate this issue, namely our choices of levels are approximately monotonic in that a descendant of B cannot be chosen before the first time B is ever chosen. This fact, and its implications, are outlined below in Section 4.2.

Remark 7 (Deterministic Alternative) *We note that a deterministic alternative to Algorithm 1 should be possible, by properly scheduling arms to be played in a round-robin fashion over $\mathcal{I}_B$. We instead focus in the present work on the simpler random choice as the additional technicality appears to add no significant further insight.*

4.2. Outline for the Proof of Theorem 1

Continuing on the discussion of Section 4.1, here we give an outline of the proof of Theorem 1. Supporting lemmas, propositions, and their proofs are found in Appendix B.

● **Bounding the regression error.** First, it is easy to verify that the criterion for choosing r_t on Line 7 of Algorithm 1 is well-defined for $t > \lceil 8K \log(K/\delta) \rceil$, that is, the set being minimized over is non-empty as it contains at least $r = 1$. At any such round t with selected bin $B \doteq T_{r_t}(X_t)$, we have by standard arguments (Lemma 1 Appendix B) that, with probability at least $1 - \delta$ (over random rewards, conditioned on all past covariates):

$$\forall x \in B, i \in \mathcal{I}_B : |\hat{f}_t^i(B) - f^i(x)| \leq \sqrt{\frac{\log(2K/\delta)}{m_t(B,i)}} + 2\lambda r_t. \tag{3}$$

Note that the first term on the RHS above contains $m_t(B,i)$ while the second term r_t is chosen according to $n_{r_t}(X_t)$. Following the discussion of Remark 6, we will relate these two terms by arguing that $m_t(B,i) \gtrsim n_{r_t}(X_t)$. To this end, as we will see, it will suffice to show that the first time B is picked, say at time $s \leq t$, we indeed have $m_s(B,i) \gtrsim n_{r_s}(X_s)$ (where, by assumption, $r_s = r_t$ since B is picked).

Therefore, suppose B is first visited at round $s \leq t$. We establish in Lemma 2 that no descendant of B had been selected by time s. Thus, at time s, the second situation described in Remark 6 has not happened, i.e., no arm in $\mathcal{I}_B$ has been eliminated in a descendant bin of B. In other words, all arms in $\mathcal{I}_{\mathcal{B}}$ are *expected* to have so far been pulled the same amount of time, roughly $n_{r_s}(X_s)/|\mathcal{I}_{\mathcal{B}}|$. More precisely, as established w.h.p. in Lemma 3 via concentration, we have that $m_s(B,i) \geq n_{r_s}(X_s)/(4K)$ for all $i \in \mathcal{I}_B$.

Pulling it all together, observe that $m_t(B,i) \geq m_s(B,i)$, while by the definition of r_s (Line 7 of Algorithm 1), we have $\sqrt{1/n_{r_s}(X_s)} \lesssim r_s = r_t$, thus relating $m_t(B,i)$ to r_t.

Formally, plugging back into (3), we have:

$$\forall x \in B, i \in \mathcal{I}_B : |\hat{f}_t^i(B) - f^i(x)| \leq \sqrt{\frac{\log(2K/\delta)}{m_s(B,i)}} + 2\lambda r_t \leq \sqrt{\frac{4K \log(2K/\delta)}{n_{r_s}(X_s)}} + 2\lambda r_t \leq 4\lambda r_t, \tag{4}$$

establishing that the regression error at round t is at most $4\lambda r_t$ with high probability.

● **Relating regression error to the regret at time t.** So far we have established that (4) holds with high probability at all rounds t. It follows that, given our elimination criteria whereby we only discard an arm i if $\hat{f}_t^{(1)}(B) - \hat{f}_t^i(B) \geq 8\lambda r_t$, the best arm for any $x \in B$ is never discarded (Corollary 3), before or after the unknown shift time n_P. In particular for $t > n_P$, it follows by simple triangle inequalities that the regret $|f^{(1)}(X_t) - f^{\pi_t}(X_t)| \lesssim \lambda r_t$ (Corollary 4).

Furthermore, we can similarly argue that whenever (4) holds at some $x \in B$ with sufficient margin $f^{(1)}(x) - f^{(2)}(x) \gtrsim \lambda r_t$, the regret must be 0 at time t, since then all suboptimal arms would have been eliminated (Corollary 4). This was discussed in Remark 5, and comes in handy below as we integrate the margin parameter α into the regret bound (last bullet point).

● **Relating adaptive r_t to an oracle choice r_t^*.** Next, we show that at each round $t > n_P$, r_t is of small order in terms of n_P and t. To this end, we will show that it cannot lead to worse regression estimates than a suitable global choice r_t^* (that in turn can be shown to yield optimal regret). First, Proposition 2 establishes that r_t satisfies:

$$\lambda r_t \lesssim \min_{r \in \mathcal{R}} \{n_r(X_t)^{-1/2} + \lambda r\}, \tag{5}$$

in other words, by (3) (and the ensuing discussion leading to (4)), λr_t is at most the best variance plus bias terms achievable by *any other choice of level r*, in particular, any suitable r_t^*.

Therefore, consider $r_t^* \doteq r_t(\gamma, n_P)$, the oracle choice of level introduced in Section 3. As discussed, it will later become clear that it leads to our main (tight) upper-bound on regret. Formally, we define r_t^* as the smallest level $r \in \mathcal{R}$ greater than or equal to

$$\min\left(\left(\frac{K \log(K/\delta)}{n_P} \right)^{\frac{1}{2+\alpha+d+\gamma}}, \left(\frac{K \log(K/\delta)}{\tau} \right)^{\frac{1}{2+\alpha+d}} \right),$$

where, to further simplify notation, we let $\tau \doteq t - n_P - 1$, the time elapsed after round n_P. Clearly, as discussed before, the choice r_t^* requires knowledge of all parameters n_P, γ, α, d, and is independent of X_t (it is a global choice at time t). Following (5), we can relate r_t to $n_{r_t^*}(X_t)$ (and r_t^*). We therefore proceed to bounding $n_{r_t^*}(X_t)$ as follows.

First, we restrict attention to times t where X_t falls in a bin of sufficient mass (for concentration) at level r_t^* (since the probability of X_t not falling in such a bin is negligible). To this end, define the event

$$A_t = \left\{ \max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t))) \geq 8 \log(1/\delta) \right\}.$$

Under event A_t, by a Chernoff bound (Proposition 3), with probability at least $1 - \delta$ (over random rewards, conditioned on X_t and A_t):

$$n_{r_t^*}(X_t) \gtrsim \max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t))). \tag{6}$$

Then, combining (5) with (6), and recalling (4), gives the following bound on λr_t:

$$\lambda r_t \lesssim \sqrt{\frac{K \log(K/\delta)}{n_{r_t^*}(X_t)}} + \lambda r_t^* \lesssim \sqrt{\frac{K \log(K/\delta)}{\max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t)))}} + \lambda r_t^* \tag{7}$$

$$\doteq \sigma_t^* + \lambda r_t^*.$$

Equipped with this upper-bound in terms of r_t^*, we are now ready to take expectation (over all randomness till time t) and properly account for the margin parameter α. Now, as λr_t controls the regret (either upper-bounds it, or drives it to 0 under margin as discussed above), so does $\sigma_t^* + \lambda r_t^*$ (where σ_t^* is the first term on the R.H.S. of (7), and can be viewed as the regression *variance* induced by choosing level r_t^*).

• **Expected regret at time t: integrating over margin distribution.** The arguments below constitute a second departure (on top of the above adaptive choice of r_t) from traditional analyses of Lipschitz bandits, due to the fact that we don't constrain the covariate distribution to be near-uniform, nor even stationary (see Remark 8 on how the usual uniform conditions remove much of the technicality herein). As a result, although r_t^* is independent of location X_t, the variance term σ_t^* can vary significantly with the random choice of X_t (as different bins at level r_t^* can have significantly different mass). This has to therefore be carefully integrated into the margin distribution across space. Since we consider a fixed round t for now, let $X \doteq X_t$ to simplify notation.

Therefore, let $\delta_f(X) \doteq f^{(1)}(X) - f^{(2)}(X)$ be the margin at X. Next, we recall and condition on all the favorable events discussed thus far. First, let G_t be the event that (4) holds and, for rounds $t > n_P$, additionally that (6) holds. Let $F_t \doteq \cap_{s=1}^t G_s$. As a reminder, F_t occurs with high

probability so we can safely condition on this event in taking expectation. As discussed so far, under event F_t, a non-zero regret (at most r_t) is incurred at time t *only if* $\delta_f(X) \lesssim r_t$. In other words, we have:

$$\mathbb{E}[(f^{(1)}(X) - f^{\pi_t}(X))\mathbb{1}\{F_t\}] \leq \mathbb{E}[(f^{(1)}(X) - f^{\pi_t}(X))\mathbb{1}\{(f^{(1)}(X) - f^{\pi_t}(X)) \vee \delta_f(X) \lesssim \sigma_t^* + r_t^*\}]$$
$$\lesssim \mathbb{E}[r_t^* \cdot \mathbb{1}\{0 < \delta_f(X) \lesssim r_t^*\}] + \mathbb{E}[\sigma_t^* \cdot \mathbb{1}\{0 < \delta_f(X) \lesssim \sigma_t^*\}]. \tag{8}$$

where, in the second inequality above, we use $\mathbb{1}\{x \leq a + b\} \leq \mathbb{1}\{x \leq 2a\} + \mathbb{1}\{x \leq 2b\}$. By the margin condition (Definition 3), the first term on the R.H.S. of (8) is of order at most $(r_t^*)^{1+\alpha}$. The main technicality left is to show that the second term is of the same order, despite the instability in σ_t^* over the spatial choice of x. For this purpose, first consider the following two random quantities in the definition of σ_t^*:

$$a(X) \doteq Q_X(T_{r_t^*}(X))^{-1/2}, \text{ and } b(X) \doteq P_X(T_{r_t^*}(X))^{-1/2}.$$

Notice that σ_t^* is of order less than either of $a(X) \cdot \tau^{-1/2}$ and $b(X) \cdot n_P^{-1/2}$. As these are unstable quantities (unlike in the uniform distribution case), the main idea is to break integration into two terms, below and above a threshold $\sqrt{\epsilon}$, for some arbitrary $\epsilon > 0$, which will then be optimized over. Next we present the main argument in the case of $a(X)$, which yields a first bound on regret in terms of elapsed time τ, while adapting the same line of arguments to $b(X)$ yields a second bound in terms of unknown n_P and γ; both of these bounds are then combined into Theorem 1.

Thus, for some $\epsilon > 0$, decompose the second term on the R.H.S. of (8) as:

$$\mathbb{E}\left[\sigma_t^* \cdot \mathbb{1}\left\{0 < \delta_f(X) \lesssim \sigma_t^*\right\} \cdot \left(\mathbb{1}\{a(X) < \sqrt{\epsilon}\} + \mathbb{1}\{a(X) \geq \sqrt{\epsilon}\}\right)\right].$$

We start by handling the case where $a(X) < \sqrt{\epsilon}$. By definition, as mentioned above, we have:

$$\sigma_t^* \leq a(X) \cdot \sqrt{\frac{K \log(K/\delta)}{\tau}}. \tag{9}$$

Combining the above with the margin condition (Definition 3) gives

$$\mathbb{E}\left[\sigma_t^* \cdot \mathbb{1}\left\{0 < \delta_f(X) \lesssim \sigma_t^*, a(X) < \sqrt{\epsilon}\right\}\right] \lesssim \sqrt{\frac{\epsilon K \log(K/\delta)}{\tau}} \cdot \mathbb{E}\,\mathbb{1}\left\{0 < \delta_f(X) \leq \sqrt{\frac{\epsilon K \log(K/\delta)}{\tau}}\right\}$$
$$\lesssim \left(\sqrt{\frac{\epsilon K \log(K/\delta)}{\tau}}\right)^{\alpha+1}. \tag{10}$$

Next, in the case where $a(X) \geq \sqrt{\epsilon}$, we have again by (9) that:

$$\mathbb{E}\left[\sigma_t^* \cdot \mathbb{1}\left\{0 < \delta_f(X) \lesssim \sigma_t^*\right\} \mathbb{1}\{a(X) \geq \sqrt{\epsilon}\}\right] \leq \sqrt{\frac{K \log(K/\delta)}{\tau}} \cdot \mathbb{E}\left[a(X)\mathbb{1}\{a(X) \geq \sqrt{\epsilon}\}\right]$$
$$\lesssim \sqrt{\frac{K \log(K/\delta)}{\tau}} \cdot \frac{(r_t^*)^{-d}}{\sqrt{\epsilon}}, \tag{11}$$

where the last inequality is a technical result derived in Proposition 4, relying on the fact that $\mathbb{E}\left[a^2(X)\right] \lesssim (r_t^*)^{-d}$ (this latter fact is relatively standard under *box cover dimension* d – Definition 4).

14

Finally, (10) + (11) is optimized by setting $\epsilon \propto \tau^{\frac{\alpha+d}{2+\alpha+d}}$, showing that the second term of (8) is $\tilde{O}(\tau^{-\frac{\alpha+1}{2+\alpha+d}})$. We can similarly show an upper-bound of $\tilde{O}(n_P^{-\frac{\alpha+1}{2+\alpha+d+\gamma}})$ by decomposing expectation in terms of $b(X)$, where analogously, a term of the form $\mathbb{E}\left[b(X)\mathbb{1}\{b(X) \geq \sqrt{\epsilon}\}\right]$ is bounded via the technical result of Proposition 4, and using the (less standard but similarly obtained) fact that $\mathbb{E}\left[b^2(X)\right] \lesssim (r_t^*)^{-(\gamma+d)}$.

Thus, we have that the regret at round t, under event F_t, is $\tilde{O}(\min(\tau^{-\frac{\alpha+1}{2+\alpha+d}}, n_P^{-\frac{\alpha+1}{2+\alpha+d+\gamma}})) \propto (r_t^*)^{1+\alpha}$. Summing over time t then yields the bound of Theorem 1.

The case of multiple shifts is handled similarly by properly bounding $\mathbb{E}[n_{r_t^*}(x)]$ in the derivation of (6) to incorporate multiple previous distributions P_X^j for $j \in [N]$ (see Appendix C).

Remark 8 (Contrast with the strong density case) *We note that under the usual* strong density *condition on Q_X (see Remark 1 for a definition), bringing α into the regret bound is more direct. Namely, due to the near-uniformity of Q_X under strong density, the optimal regression rate is of the same order $r_t^{**} \propto \tau^{-1/(2+d)}$ spatially for all possible X_t. As a consequence, since nonzero regret only happens when the margin $\delta_f(X_t) \lesssim r_t^{**}$ (a non-random quantity), we almost immediately have that the regret at time t – upon optimal regression – is upper-bounded by*

$$r_t^{**} \cdot \mathbb{E}\,\mathbb{1}\{\delta_f(X_t) \lesssim r_t^{**}\} \lesssim (r_t^{**})^{\alpha+1}.$$

On the other hand, we note the following interesting but nuanced point: even in our current setting without the strong density assumption, it can be shown that the optimal choice of level w.r.t. L_2 regression (i.e. in bounding $\mathbb{E}\,(\hat{f}^i(X_t) - f^i(X_t))^2$) remains r_t^{**}, rather than than r_t^* as defined in our case. In particular, our larger (oracle) choice of level $r_t^* > r_t^{**}$ – is required to mitigate variance in lower-density regions over space – and implies that the bandit problem differs in this case more significantly from the underlying regression problem (enough that the optimal regression choice of r_t^{**} is now suboptimal for bandits).

Finally, the case of strong density with unknown d, γ and α is handled in Theorems 5 and 6.

Remark 9 (Beyond Covariate Shift) *Under more severe shifts in the rewards, for instance ones where previously discarded arms now become optimal, our procedure will be sub-optimal as it has no mechanism to detect such changes. However, the regret rates of Theorem 1 can still be attained by Algorithm 1 under mild changes in the reward function $f_P \to f_Q$. Let $\epsilon(x) \doteq \max_{i \in [K]} |f_P^i(x) - f_Q^i(x)|$, which roughly quantifies the amount of shift in the rewards. Intuitively, if $\epsilon(x) \ll \sqrt{1/n_P(r_{n_P})}$ for all $x \in \mathcal{X}_Q$ where $n_P(r)$ is the covariate count in bin $T_r(x)$ from P_X, then the amount of change is so small that the new best arm under Q, $i_Q^*(x)$, must have been retained as a candidate arm at time n_P. Moreover, we can still accurately estimate f^i using $\hat{f}_t^i$ for any candidate arm i as if there was no shift. To see this, consider decomposing $\hat{f}_t^i = \alpha \cdot \hat{f}_P^i + (1-\alpha) \cdot \hat{f}_Q^i$ as a weighted sum of oracle estimates $\hat{f}_P^i$ (resp. $\hat{f}_Q^i$) using only the observed data from P (resp. Q) with weight $\alpha \doteq n_P(r_t)/(n_P(r_t) + n_Q(r_t))$ ($n_Q(r)$ defined analogously):*

$$\max_{i \in \mathcal{I}} |\hat{f}_t^i(x) - f_Q^i(x)| \lesssim \alpha \cdot |\hat{f}_P^i(x) - f_P^i(x)| + \alpha \cdot \epsilon(x) + (1-\alpha) \cdot |\hat{f}_Q^i(x) - f_Q^i(x)|$$

$$\lesssim r_t + \sqrt{\frac{1}{n_P(r_t) + n_Q(r_t)}} \lesssim r_t.$$

On the other hand, a different approach is required to handle the more challenging setting of severe shifts in the rewards.

15

Acknowledgements

References

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

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016.

Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In *Advances in neural information processing systems*, 2013.

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153, 2012.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, 2014.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Nicoló Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions*, 50(9):2050–2057, 2004.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: efficient, optimal, and parameter-free. In *32nd Annual Conference on Learning Theory*, 2019.

Yining Chen, Haipeng Luo, Tengyu Ma, and Chicheng Zhang. Active online domain adaptation. *arXiv preprint arXiv:2006.14481*, 2020.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Hedging the drift: learning to optimize under non-stationarity. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Kenneth L Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-neighbor methods for learning and vision: theory and practice*, pages 15–59, 2006.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Melody Y Guan and Heinrich Jiang. Nonparametric stochastic contextual bandits. *AAAI*, 2018.

Negar Hariri, Bamshad Mobasher, and Robin Burke. Adapting to user preference changes in interactive recommendation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Elad Hazan and Nimrod Megiddo. Online learning with prior knowledge. In *International Conference on Computational Learning Theory*, pages 499–513. Springer, 2007.

Zohar S Karnin and Oren Anava. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems*, pages 199–207, 2016.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. *COLT*, 2018.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

Fang Liu, Joohyun Lee, and Ness Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *31st Annual Conference on Learning Theory (COLT)*, 2018.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *ICML*, pages 1977–1985, 2016.

Henry W. J. Reeve, Joe Mellor, and Gavin Brown. The k-nearest neighbour ucb algorithm for multi-armed bandits with covariates. *JMLR*, 2018.

Phillipe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *COLT*, 2010.

Clayton Scott and Robert D Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE transactions on information theory*, 52(4):1335–1353, 2006.

Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

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

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 495–504, 2018.

Yuhong Yang, Dan Zhu, et al. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

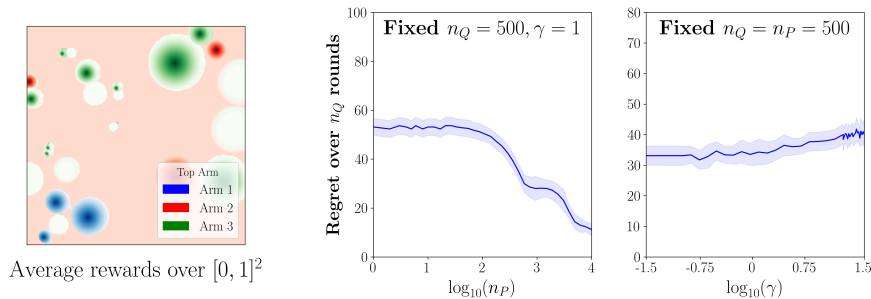# Appendix A. Additional Experiments

Average rewards over $[0,1]^2$

Fig. 4: Additional simulations with a different choice of bump centers than in Figure 3: here, the bump centers are chosen uniformly in $[0,1]^2$. Under this setup, we see that the behavior of the regret in the center and right plots are similar to that of Figure 3.

For all our experiments, we fix a covariate space $\mathcal{X} = [0,1]^2$. Our terminal covariate distribution is $Q_X \sim \mathcal{U}([0,1]^2)$, the uniform on $[0,1]^2$, and our initial covariate distribution P_X has density $\propto \|x\|_2^\gamma$ so that P_X, Q_X satisfy Definition 5 with transfer exponent γ.

The reward function common to both P and Q is constructed as the sum of 25 bump functions, each with a circular support disjoint from the other bumps, in the following manner:

1. The bump centers $\{z_k\}_{k=1}^{25}$ are first randomly sampled from the Gaussian $\mathcal{N}\left((0.5, 0.5), 0.5 \cdot \text{Id}_2\right)$ in Figure 3 and from the uniform $\mathcal{U}([0,1]^2)$ in Figure 4.

2. The bumps' radii $\{r_k\}_{k=1}^{25}$ are then chosen in a random order to maximize the bump areas.

3. Then, for each of the $K = 3$ arms, we determined the sign of each of the 25 bumps randomly and independently.

4. To introduce additional heterogeneity in the top arm identity $\pi^*(x)$, each reward function f^i was further raised or lowered by a randomly selected height in the range $[-0.3, 0.3]$, in the area outside of the bumps.

The fourth step determines a unique top arm (Arm 1 in Figure 3 and Arm 2 in Figure 4) in the region outside of the bumps. To summarize the above, the reward functions f^i can be written as

$$f^i(x) \propto \sum_k \omega_{i,k}(1 - \|x - z_k\|_2/r_k)_+,$$

where $\omega_{i,k}$ are independent random Rademacher variables. Furthermore, Gaussian noise was added to each f^i to produce the observed rewards Y^i according to $Y^i = f^i(X) + \mathcal{N}(0, 0.05)$ for each $i \in [K]$.

Now, having determined f, we considered a range of different values for the parameters n_Q, n_P, γ, shown in the horizontal axes of Figures 3 and 4. Then, Algorithm 1 was run on 20 simulations of data $(\mathbf{X}_n, \mathbf{Y}_n)$ for each choice of parameters, so that the plots show the mean and standard deviation of the regret $\mathbf{R}_n^Q$ across 20 trials.

The first plot in each of Figures 3 and 4 exhibits the guarantee of increasing past experience n_P improving the regret, for fixed n_Q, γ. The second plot in each figure shows the effect of increasing γ

worsening the regret, for fixed n_P, n_Q. Together, the two plots in each of Figure 3 and 4 demonstrate the guarantees of Theorem 1.

We also compare Algorithm 1 to the *adversarial* nonparametric contextual bandits algorithm "ContextualBandit" of Slivkins (2014) under the same setting as Figure 4. We instantiate their algorithm using the static bandits algorithm Exp3 with a common learning rate $\eta = \sqrt{\log K / TK}$. These simulations show that adversarial design can be too conservative in leveraging past experience: for small n_P we see no benefit for either procedure, while our procedure better leverages the past for sufficiently large n_P.
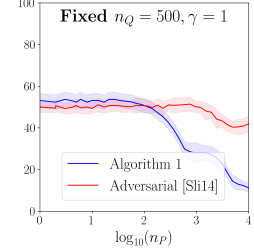
Appendix B. Proof of Theorem 1

Throughout the proof, $c_0, c_1, c_2, \ldots$ will denote positive constants not depending on t, n_P, n_Q, K.

First, it is straightforward to verify that the criterion for choosing r_t on Line 7 of Algorithm 1 is well-defined for $t > \lceil 8K \log(K/\delta) \rceil$, i.e. $r = 1$ satisfies the minimization criterion.

Bias-Variance Bound.   This first proposition establishes a standard bias-variance bound on the error of a regression function estimate $|\hat{f}_t^i(B) - f^i(x)|$ for a bin B, a round t, an arm $i \in \mathcal{I}_B$, and a covariate $x \in B$.

We use Z_t to denote the randomness of Algorithm 1 at round t in choosing the particular arm π_t to play.

Lemma 1   *Consider any round $t > 8K \log(K/\delta)$ with observed covariate X_t, and fix any bin B containing X_t. Consider the estimate $\hat{f}_t^i(B)$ as in Definition 7, and let $m_t(B, i)$ be defined therein (i.e., the number of times arm i is pulled in B by time t). We then have at round t, that with probability at least $1 - \delta$ with respect to the conditional distribution $\mathbf{Y}_{t-1} | X_t, \{Z_s\}_{s<t}$:*

$$\forall x \in B, i \in [K] : |\hat{f}_t^i(B) - f^i(x)| \leq \sqrt{\frac{\log(2K/\delta)}{m_t(B, i)}} + 2\lambda r. \tag{12}$$

Proof Fix bin B and let $r \in \mathcal{R}$ be its side length. If $m_t(B, i) = 0$, then the desired bound is vacuously true. So, suppose $m_t(B, i) > 0$. Now, recall from Definition 7,

$$\hat{f}_t^i(B) \doteq \frac{1}{m_t(B, i)} \sum_{X_s \in B, s \leq t-1, \pi_s = i} Y_s^i.$$

For the sake of introducing a bias term of $\hat{f}_t^i(B)$, define

$$\tilde{f}_t^i(B) \doteq \mathbb{E}_{\mathbf{Y}_{t-1} | \mathbf{X}_{t-1}} [\hat{f}_t^i(B)] = \frac{1}{m_t(B, i)} \sum_{X_s \in B, s \leq t-1, \pi_s = i} \mathbb{E}[Y_s^i | X_s].$$

Triangle inequality then yields

$$|\hat{f}_t^i(B) - f^i(x)| \leq |\hat{f}_t^i(B) - \tilde{f}_t^i(B)| + |\tilde{f}_t^i(B) - f^i(x)|.$$

20

The second term on the R.H.S. above is at most $2\lambda r$ by the Lipschitz assumption (Assumption 2). Now, fix the values of $\{m_t(B, i)\}_{i \in [K]}$. By Hoeffding inequality and union bound, the first term on the R.H.S. above satisfies with probability at least $1 - \delta$ w.r.t. the distribution of $\mathbf{Y}_{t-1}|\mathbf{X}_t, \{m_t(B, i)\}_{i \in [K]}$,

$$\forall i \in [K] : |\hat{f}_t^i(B) - \tilde{f}_t^i(B)| \leq \sqrt{\frac{\log(2K/\delta)}{m_t(B, i)}}.$$

In fact, by the tower property, the above holds with probability at least $1 - \delta$ w.r.t. the distribution of $\mathbf{Y}_{t-1}|\mathbf{X}_t.\{Z_s\}_{s<t}$. ∎

In particular, the bound of (12) holds for $r = r_t$ and $B = T_{r_t}(X_t)$. At $r = r_t$, the first term on the R.H.S. of (12) depends on the arm-pull count $m_t(B, i)$, while the second term is chosen according to $n_{r_t}(X_t)$ (Line 7 of Algorithm 1). We next relate the counts $m_t(B, i)$ and $n_{r_t}(X_t)$ to show that the R.H.S. of (12) is $O(r_t)$.

Relating Arm-Pull Counts $m_t(B, i)$ to Covariate Counts $n_r(X_t)$. We start by showing that the chosen level r_t and bin $B \doteq T_{r_t}(X_t)$ cannot "skip" levels in the sense that B is selected before any of its descendants. Thus, each arm in $\mathcal{I}_B$ had a chance of being pulled once (by the randomization in Line 12 of Algorithm 1) for each of the $n_{r_t}(X_t)$ covariates.

This will ensure that each arm in $\mathcal{I}_B$ has been played roughly the same number of times, so that $m_t(B, i) \gtrsim n_{r_t}(X_t)/K$ (Lemma 3 further below), thus relating the two counts.

Lemma 2 *Fix a bin B, and suppose t is the first round that B is selected. Then, no descendant bin of B was selected in a round previous to t.*

Proof For contradiction, suppose a descendant $B' \subset B$ was selected at round $s < t$. W.L.O.G., let s be the first round that any descendant of B was selected. Then, by the criterion for choosing r_s on Line 7 of Algorithm 1, we have:

$$r_s \geq \sqrt{\frac{8K \log(K/\delta)}{n_{r_s}(X_s)}} \implies n_{r_s}(X_s) \geq \frac{8K \log(K/\delta)}{r_s^2}.$$

However, since $r_t \geq 2r_s$, we have

$$\frac{8K \log(K/\delta)}{r_t^2} \leq \frac{8K \log(K/\delta)}{(2r_s)^2} \leq \left\lceil \frac{n_{r_s}(X_s)}{4} \right\rceil < n_{r_s}(X_s).$$

The above implies that there was an earlier round $s' < s$ and a covariate $X_{s'} \in B'$ such that

$$n_{r_s}(X_{s'}) \geq \frac{8K \log(K/\delta)}{r_t^2} \implies r_t \geq \sqrt{\frac{8K \log(K/\delta)}{n_{r_s}(X_{s'})}} \geq \sqrt{\frac{8K \log(K/\delta)}{n_{r_t}(X_{s'})}}.$$

According to the minimization criterion for choosing $r_{s'}$ on Line 7 of Algorithm 1, we must have $r_{s'} \leq r_t$, a contradiction to either t being the first round that B is selected or s being the first round that a descendant of B was selected. ∎

We next use a concentration argument to show that $m_t(B, i) \gtrsim n_{r_t}(X_t)/K$.

Lemma 3 *Fix a round $t > \lceil 8K \log(K/\delta) \rceil$ with observed covariate X_t and selected bin B. Suppose that t is the first round that B is selected. Then, with probability at least $1 - \delta$ with respect to the distribution of $\mathbf{Y}_{t-1}, \{Z_s\}_{s<t} | X_t$, we have*

$$\forall i \in \mathcal{I}_B : m_t(B, i) \geq \frac{n_{r_t}(X_t)}{4K}.$$

Proof Fix the values of $\mathbf{X}_t, \mathbf{Y}_{t-1}, \mathcal{I}_B$ and fix some $i \in \mathcal{I}_B$. Recall:

$$m_t(B, i) = \sum_{X_s \in B, s \leq t-1} \mathbb{1}\{\pi_s = i\}.$$

We first handle the case where $r_t = 1$ and thus $B = [0, 1]^D$. Then, we have

$$n_{r_t}(X_t) = n_1(X_t) = \lceil 8K \log(K/\delta) \rceil + 1.$$

By Line 5 of Algorithm 1, for each of the t rounds elapsed thus far, we pulled arm i with probability $1/K$. Thus, we have

$$\mathbb{E}[m_t(B, i)] = n_1(X_t) \cdot \frac{1}{K} \geq 8K \log(K/\delta).$$

Then, by a Chernoff bound, since $\{Z_s\}_{s<t}$ are independent:

$$\mathbb{P}\left(m_t(B, i) \leq \frac{n_1(X_t)}{2K}\right) \leq \mathbb{P}\left(m_t(B, i) \leq \frac{\mathbb{E}[m_t(B, i)]}{2}\right) \leq \delta/K.$$

This gives us the desired result for $r_t = 1$.

The general case of $r_t < 1$ will follow similarly from a Chernoff bound. However, more care is required in that the sequence $\{Z_s\}_{s<t}$ is no longer independent since each Z_s depends on $\mathcal{I}_B$ (Line 12 of Algorithm 1), a random object possibly varying with time and depending on $\{Z_s\}_{s<t}, \mathbf{Y}_{t-1}, \mathbf{X}_{t-1}$. To overcome this, we relate $m_t(B, i)$ to a smaller count of *independently randomized* arm-pulls so that we can use concentration.

First, let t_0 be the first round that the parent $B' \doteq T_{2r_t}(X_t)$ of bin B was used. Let $m_{[t_0,t]}(B, i)$ be the number of arm pulls of arm i in bin B between rounds t_0 and t. Then, it suffices to show $m_{[t_0,t]}(B, i) \geq n_{r_t}(X_t)/(4K)$.

Let $\mathcal{A}_{t_0}$ be the set of candidate arms determined in B' at time t_0 so that $\mathcal{A}_{t_0} \supseteq \mathcal{I}_B$. Next, let $\tilde{m}_{[t_0,t]}(B, i)$ be a draw from a Binomial$(n_{[t_0,t]}(B), 1/|\mathcal{A}_{t_0}|)$ distribution where $n_{[t_0,t]}(B)$ is the number of covariates observed in B between times t_0 and t. More plainly, $\tilde{m}_{[t_0,t]}(B, i)$ is the number of pulls of arm i in B between times t_0 and t if Algorithm 1 was *prohibited* from eliminating any arms in $\mathcal{A}_{t_0}$ between times t_0 and t in bin B' – in other words, $\tilde{m}_{[t_0,t]}(B, i)$ counts the number of times arm i is selected with probability $1/|\mathcal{A}_{t_0}|$ instead of with probability $1/|\mathcal{I}_B|$, as in Line 12 of Algorithm 1.

Then, since $|\mathcal{I}_B| \leq |\mathcal{A}_{t_0}|$, we have that $m_{[t_0,t]}(B, i) \geq \tilde{m}_{[t_0,t]}(B, i)$ so that it remains to show $\tilde{m}_{[t_0,t]}(B, i) \geq n_{r_t}(X_t)/(4K)$.

We first show that the number $n_{[t_0,t]}(B)$ of covariates observed in B between t_0 and t is at least $n_{r_t}(X_t)/2$. Since t is the first time that B is chosen, we must have by Line 7 of Algorithm 1

$$r_t \geq \sqrt{\frac{8K \log(K/\delta)}{n_{r_t}(X_t)}} \implies n_{r_t}(X_t) \geq \frac{8K \log(K/\delta)}{r_t^2}. \tag{13}$$

By similar reasoning, since t_0 is the first time that B's parent at level $2r_t$ is chosen, we must have

$$\sqrt{\frac{8K \log(K/\delta)}{n_{2r_t}(X_{t_0}) - 1}} > 2r_t \implies n_{2r_t}(X_{t_0}) \leq \frac{8K \log(K/\delta)}{4r_t^2} + 1. \tag{14}$$

Then, putting (13) and (14) together, we have $n_{r_t}(X_t) - n_{[t_0,t]}(B) \leq n_{2r_t}(X_{t_0}) \leq \frac{n_{r_t}(X_t)}{2}$. This implies that $n_{[t_0,t]}(B) \geq n_{r_t}(X_t)/2$.

Then, by Lemma 2, at every round $s \in [t_0, t) \cap \mathbb{N}$ during which a covariate was observed in B, we pulled arm i with probability at least $1/K$. Thus, we have

$$\mathbb{E}\left[\tilde{m}_{[t_0,t]}(B,i) \mid \mathbf{X}_{t_0}, \mathbf{Y}_{t_0}, \mathcal{A}_{t_0}\right] \geq n_{[t_0,t]}(B) \cdot \frac{1}{K} \geq \frac{n_{r_t}(X_t)}{2K}.$$

Furthermore, we have from (13) and the fact that $r_t \leq 1/2$ that the above R.H.S. is further lower bounded by:

$$\frac{8 \log(K/\delta)}{2r_t^2} \geq 8 \log(K/\delta).$$

Then, since $\tilde{m}_{[t_0,t]}(B,i)$ is a sum of *independent* Bernoulli's conditioned on $\mathbf{X}_{t_0}, \mathbf{Y}_{t_0}, \mathcal{A}_{t_0}$, by a Chernoff bound, we have:

$$\mathbb{P}\left(\tilde{m}_{[t_0,t]}(B,i) \leq \frac{n_{r_t}(X_t)}{4K}\right) \leq \mathbb{P}\left(\tilde{m}_{[t_0,t]}(B,i) \leq \frac{\mathbb{E}\left[\tilde{m}_{[t_0,t]}(B,i) \mid \mathbf{X}_{t_0}, \mathbf{Y}_{t_0}, \mathcal{A}_{t_0}\right]}{2}\right) \leq \delta/K.$$

Re-tracing our previous steps, we have with probability at least $1 - \delta/K$:

$$m_t(B,i) \geq m_{[t_0,t]}(B,i) \geq \tilde{m}_{[t_0,t]}(B,i) > \frac{n_{r_t}(X_t)}{4K}.$$

By a union bound and the tower property, we have that the event $\{\forall i \in \mathcal{I}_B : m_t(B,i) \geq \frac{n_{r_t}(X_t)}{4K}\}$ holds with probability at least $1 - \delta$ w.r.t. the distribution of $\mathbf{Y}_{t-1}, \{Z_s\}_{s<t} | \mathbf{X}_t$. ∎

Now, consider a round $t > \lceil 8K \log(K/\delta) \rceil$ with observed covariate X_t and selected bin B. Suppose $s \leq t$ is the first round when bin B is selected. Then, the set of candidate arms determined at round s in B must contain any arm currently retained in B at round t.

Furthermore, by Lemma 3, we have that with probability at least $1 - \delta$, $m_t(B,i) \geq m_s(B,i) > n_{r_s}(X_s)/(4K)$. Thus, combining this with our earlier bias-variance bound (12):

$$|\hat{f}_t^i(B) - f^i(x)| \leq \sqrt{\frac{4K \log(2K/\delta)}{n_{r_s}(X_s)}} + 2\lambda r_t \leq 2\lambda r_s + 2\lambda r_t = 4\lambda r_t.$$

This shows that the regression error at round t is at most $4\lambda r_t$.

Justifying Arm Eliminations. For each round $t > 8K \log(K/\delta)$, define the event G_t on which Lemma 1 and Lemma 3 hold or

$$G_t \doteq \left\{\forall i \in \mathcal{I}_B, x \in B : |\hat{f}_t^i(B) - f^i(x)| \leq 4\lambda r_t, B = T_{r_t}(X_t)\right\}.$$

This is the "good" event on which our regression function estimates $\hat{f}_t^i(B)$ are accurate enough to be able to discern which arms have low and high rewards. For the sake of brevity, from here on, let B be the selected bin at round t. This first proposition asserts that an eliminated arm cannot have a better reward than the best candidate arm.

Proposition 1 *Suppose at round t, under event G_t, we select bin B. Then for any two arms $i, j \in \mathcal{I}_B$ and any $x \in B$,*

$$\hat{f}_t^i(B) - \hat{f}_t^j(B) > 8\lambda r_t \implies f^i(x) > f^j(x)$$

Proof Using the definition of G_t, we have

$$f^i(x) - f^j(x) \geq \hat{f}_t^i(B) - \hat{f}_t^j(B) - 8\lambda r_t > 0.$$

∎

The first implication of Proposition 1 is that under event $\cap_{s=1}^t G_s$, the best arm $i^*(x)$ at any covariate $x \in B$ is always retained in $\mathcal{I}_B$. This is immediate since $i^*(x)$ can never be discarded for any $x \in B$ as long as the regression bounds of G_s hold.

Corollary 3 *Suppose at round t, under event $(\cap_{s=1}^t G_s)$, we select bin B. Then $\mathcal{I}_B$ contains the best arm $i^*(x) = \operatorname{argmax}_{j \in [K]} f^j(x)$ for all $x \in B$.*

The next corollary gives us that the margin and regret of playing any candidate arm at any point in B is bounded by λr_t. Following the discussion of Section 4.2, it will then suffice to bound r_t in terms of n_P and t.

Corollary 4 *Suppose at round t, under event $(\cap_{s=1}^t G_s)$, we select bin B. Then, both of the following hold for all $x \in B$:*

(1) $|f^{(1)}(x) - f^j(x)| \leq 16\lambda r_t$ *for all $j \in \mathcal{I}_B$.*

(2) Either $0 < |f^{(1)}(x) - f^{(2)}(x)| \leq 16\lambda r_t$ *or* $f^j(x) = f^{(1)}(x)$ *for all $j \in \mathcal{I}_B$*

Proof Fix $x \in B$, and let $i^*(x) = \operatorname*{argmax}_{j \in [K]} f^j(x)$. Using the definition of G_t and the fact that $i^*(x) \in \mathcal{I}_B$ (Corollary 3), we first establish (1):

$$\begin{aligned} f^{i^*(x)}(x) - f^j(x) &\leq \hat{f}^{i^*(x)}(x) - \hat{f}^j(x) + 8\lambda r_t \\ &\leq \hat{f}^{(1)}(x) - \hat{f}^j(x) + 8\lambda r_t \\ &\leq 16\lambda r_t \text{ (because } j \text{ was not eliminated)} \end{aligned}$$

To show (2), we have if $\mathcal{I}_B$ contains a sub-optimal arm $j \in \mathcal{I}_B$ at x, then by (1),

$$|f^{(1)}(x) - f^{(2)}(x)| = f^{(1)}(x) - f^{(2)}(x) \leq f^{(1)}(x) - f^j(x) \leq 16\lambda r_t.$$

On the other hand, if $\mathcal{I}_B$ does not contain a sub-optimal arm at x, then $f^j(x) = f^{(1)}(x)$ for all $j \in \mathcal{I}_B$. ∎

Relating adaptive r_t to an oracle choice r_t^* Following the outline of Section 4.2, we relate r_t to the oracle choice $r_t^* \doteq r_t(\gamma, n_P)$ for rounds $t > n_P$. We do this by first establishing that r_t leads to near optimal regression estimates. Let

$$\phi_t(r) \doteq \sqrt{\frac{8K \log(K/\delta)}{n_r(X_t)}} + \lambda r.$$

This is essentially the bias-variance decomposition from earlier bounding the regression error $|f^i(x) - \hat{f}_t^i(B)|$. Next, we show that r_t, the high probability regret bound achieved thus far (Corollary 4), is less than $\phi_t(r)$ for any $r \in \mathcal{R}$.

Proposition 2 (r_t minimizes $\phi_t(\cdot)$) *We have $r_t \leq 2 \cdot \min_{r \in \mathcal{R}} \phi_t(r)$.*

Proof Fix $r \in \mathcal{R}$. If $r_t \leq r$, then using the fact that $\lambda \geq 1$:

$$r_t \leq r \leq \lambda r \leq \phi_t(r).$$

If $r_t > r \implies r_t/2 \geq r$, then using the definition of r_t on Line 7 of Algorithm 1:

$$r_t = 2(r_t/2) \leq 2\sqrt{\frac{8K \log(K/\delta)}{n_{r_t/2}(X_t)}} \leq 2\sqrt{\frac{8K \log(K/\delta)}{n_r(X_t)}} \leq 2 \cdot \phi_t(r).$$

$\blacksquare$

Proposition 2 directly gives us that $\lambda r_t \leq 2\lambda \cdot \min_{r \in \mathcal{R}} \phi_t(r)$. Thus, using the level r_t at time t achieves regression error *no worse than that of any other choice of level r.*

Recall from Section 4, we let $\tau \doteq t - n_P - 1$, the time elapsed after round n_P, to further simplify notation. Also, let r_t^* be the oracle choice of level introduced in Section 3, or the smallest level of $\mathcal{R}$ greater than or equal to

$$\min\left(\left(\frac{K \log(K/\delta)}{n_P}\right)^{\frac{1}{2+\alpha+d+\gamma}}, \left(\frac{K \log(K/\delta)}{\tau}\right)^{\frac{1}{2+\alpha+d}} \right).$$

Then, by Proposition 2, we have $\lambda r_t \leq 2\lambda \cdot \phi_t(r_t^*)$. From here, it remains to integrate $\phi_t(r_t^*)$ over the covariate space.

Expected regret at time t: integrating over margin distribution. To later use the margin condition (Definition 3) for all relevant bounds Δ on the margin, we first need to ensure that $\Delta \leq \delta_0$, where δ_0 is as in Definition 3. It turns out that this will only amount to constraining our analysis to rounds for which $\tau \gtrsim K \log(K/\delta)$, which will not pose an issue since the regret of any fixed $O(K \log(K/\delta))$ rounds among rounds $\{n_P + 1, \ldots, n\}$ is of the right order with respect to Theorem 1. Formally, let τ_0 be the largest positive integer such that

$$\tau_0 \leq \left(\frac{c_0 \vee c_3}{\delta_0}\right)^{2+\alpha+d} K \log(K/\delta), \tag{15}$$

25

where $c_0, c_3 \geq 1$ are constants to be determined (see Lemma 4 and (22), respectively, for where they arise). Rearranging (15), we obtain:

$$c_0 \left(\frac{K \log(K/\delta)}{\tau_0} \right)^{\frac{1}{2+\alpha+d}} \geq \delta_0.$$

Thus, for rounds t such that $\tau > \tau_0$, from the above, we have $c_0 r_t^* \leq \delta_0$.

Similarly, again rearranging (15), we obtain for $\tau > \tau_0$:

$$c_3 \sqrt{\left(\frac{K \log(K/\delta)}{\tau} \right)^{1 - \frac{\alpha+d}{2+\alpha+d}}} \leq \delta_0. \tag{16}$$

The above inequality will later be useful in integrating over low-margin regions of $\mathcal{X}$ with respect to the variance term in $\phi_t(r_t^*)$.

Next, following the outline of Section 4.2, we consider bins of sufficient mass at level r_t^* to use concentration. So, define the event A_t:

$$A_t = \left\{ \max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t))) \geq 8 \log(1/\delta) \right\}.$$

Then, in the following proposition, we use concentration to relate the variance term of $\phi_t(r_t^*)$ to the masses $Q_X(T_{r_T^*}(X_t)), P_X(T_{r_t^*}(X_t))$, under event A_t.

Proposition 3 *Consider any round t with $\tau > K \log(K/\delta)$ and with observed covariate $X \doteq X_t$. We then have, at round t, that with probability at least $1 - \delta$ w.r.t. the conditional distribution $\mathbf{X}_{t-1}|X_t, A_t$:*

$$\phi_t(r_t^*) \leq 2 \sqrt{ \frac{8K \log(K/\delta)}{\max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t)))} } + 2\lambda r_t^*. \tag{17}$$

Proof We note that

$$\mathbb{E}\left(n_{r_t^*}(X_t) \right) = n_P P_X(T_{r_t^*}(X)) + \tau Q_X(T_{r_t^*}(X)) \geq \max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t))).$$

Then, we have, under event A_t, by a Chernoff bound that:

$$\mathbb{P}\left(n_{r_t^*}(X_t) \leq \frac{1}{2}\mathbb{E}[n_{r_t^*}(X_t)] \right) \leq \exp\left(-\frac{1}{8}\mathbb{E}[n_{r_t^*}(X_t)] \right) \leq \delta.$$

Thus, with probability at least $1 - \delta$:

$$\phi_t(r_t^*) \leq 2 \left(\sqrt{ \frac{8K \log(K/\delta)}{\max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t))} } + \lambda r_t^* \right).$$

∎

Combining Proposition 3 with Proposition 2, we have with probability at least $1 - \delta$:

$$\lambda r_t \leq 2\lambda \sqrt{ \frac{8K \log(K/\delta)}{\max(\tau Q_X(T_{r_t^*}(X_t)), n_P P_X(T_{r_t^*}(X_t)))} } + 2\lambda^2 r_t^* \tag{18}$$

From here on, fix a round t with corresponding $\tau > \tau_0$.

Notation We recall some of the notation from Section 4.2, which we make more precise here. Denote the first term on the R.H.S. of (18) by σ_t^*. Let $\delta_f(X_t) \doteq f^{(1)}(X_t) - f^{(2)}(X_t)$ be the margin at X_t. Define the event E_t as the event on which the bound of Proposition 3 holds or:

$$E_t \doteq \left\{ n_{r_t^*}(X_t) > \frac{1}{2}\mathbb{E}[n_{r_t^*}(X_t)] \right\}.$$

Finally, let $F_t \doteq A_t \cap E_t \cap \left(\cap_{s=1}^t G_s \right)$, which is the event on which all the high-probability bounds established thus far hold.

Lemma 4 *For $\tau > K\log(K/\delta)$:*

$$\mathbb{E}[(f^{(1)}(X_t) - f^{\pi_t}(X_t))\mathbb{1}\{F_t\}] \leq c_0 \left(\mathbb{E}[r_t^* \cdot \mathbb{1}\{0 < \delta_f(X_t) \leq c_0 r_t^*\}] + \mathbb{E}[\sigma_t^* \cdot \mathbb{1}\{0 < \delta_f(X_t) \leq c_0\sigma_t^*\}] \right). \tag{19}$$

Proof As in Section 4.2, to simplify notation, let $X \doteq X_t$. From Corollary 4, we have that a non-zero regret (of at most $16\lambda r_t$) is incurred only if $\delta_f(X_t) \leq 16r_t$. Combining this fact with (18), we have the regret at round t of pulling arm j under event F_t is

$$\mathbb{E}[(f^{(1)}(X) - f^{\pi_t}(X))\mathbb{1}\{F_t\}] \leq \mathbb{E}[(f^{(1)}(X) - f^{\pi_t}(X))\mathbb{1}\{0 < (f^{(1)}(X) - f^{\pi_t}(X)) \\ \vee \delta_f(X) \leq c_1(\sigma_t^* + r_t^*)\}].$$

Next, using $\mathbb{1}\{x \leq a + b\} \leq \mathbb{1}\{x \leq 2a\} + \mathbb{1}\{x \leq 2b\}$, we decompose the above R.H.S.:

$$\mathbb{E}[(f^{(1)}(X) - f^{\pi_t}(X))\mathbb{1}\{F_t\}] \leq \mathbb{E}\left[(f^{(1)}(X) - f^{\pi_t}(X)) \cdot \mathbb{1}\{0 < \delta_f(X) \vee (f^{(1)}(X) - f^{\pi_t}(X)) \leq c_0 r_t^*\} \right] \\ + \mathbb{E}\left[(f^{(1)}(X) - f^{\pi_t}(X)) \cdot \mathbb{1}\{0 < \delta_f(X) \vee (f^{(1)}(X) - f^{\pi_t}(X)) \leq c_0\sigma_t^*\} \right] \\ \leq c_0 \left(\mathbb{E}\left[r_t^* \cdot \mathbb{1}\{0 < \delta_f(X) \leq c_0 r_t^*\}\right] + \mathbb{E}\left[\sigma_t^* \cdot \mathbb{1}\{0 < \delta_f(X) \leq c_0\sigma_t^*\}\right] \right).$$

∎

We next show that the R.H.S. of (19) is of order at most $(r_t^*)^{1+\alpha}$. For the first term on the R.H.S. of (19), this is immediate: we use the margin assumption (Definition 3) along with (15) to write:

$$\mathbb{E}_{X_t}[r_t^* \cdot \mathbb{1}\{0 < \delta_f(X_t) \leq c_0 r_t^*\}] \leq c_2(r_t^*)^{1+\alpha}.$$

Thus, it remains to analyze the second term, involving σ_t^*, on the R.H.S. of (19). Now, since r_t^* is a minimum of two terms, one involving τ and one involving n_P, it suffices to show separately that $\mathbb{E}[\sigma_t^* \cdot \mathbb{1}\{0 < \delta_f(X_t) \leq c_0\sigma_t^*\}]$ is $\tilde{O}(\tau^{-\frac{\alpha+1}{2+\alpha+d}})$ and is also $\tilde{O}(n_P^{-\frac{\alpha+1}{2+\alpha+d+\gamma}})$. We show the first bound involving τ; the bound involving n_P will follow from the same arguments with the appropriate modifications (which we will make explicit hereafter).

Proceeding with the bound involving τ, first consider the quantity $a(X_t) \doteq Q_X(T_{r_t^*}(X_t))^{-1/2}$. We note that

$$\sigma_t^* \leq 2\lambda \cdot a(X_t) \cdot \sqrt{\frac{8K\log(K/\delta)}{\tau}} \tag{20}$$

Thus, it suffices to bound the expectation of $a(X_t)$ over regions of low-margin. We achieve this by splitting the integral into two terms, conditioned on the value of $a(X_t)$ being above or below a

threshold $\sqrt{\epsilon}$ for some fixed $\epsilon > 0$ (to be optimized over later). More precisely, for some $\epsilon > 0$, we decompose the second term on the R.H.S. of (19) as:

$$\mathbb{E}\left[\sigma_t^* \cdot \mathbb{1}\{0 < \delta_f(X_t) \le c_0 \sigma_t^*\} \cdot (\mathbb{1}\{a(X_t) < \sqrt{\epsilon}\} + \mathbb{1}\{a(x_t) \ge \sqrt{\epsilon}\})\right]. \tag{21}$$

This gives us two cases to consider: $a(X_t) < \sqrt{\epsilon}$ and $a(X_t) \ge \sqrt{\epsilon}$. We start by handling the case where $a(X_t) < \sqrt{\epsilon}$. Plugging (20) into the first term above gives (using c_3 to collect constants):

$$c_3 \cdot \sqrt{\frac{K \log(K/\delta)}{\tau}} \cdot \mathbb{E}\left[a(X_t)\mathbb{1}\left\{0 < \delta_f(X_t) \le c_3 \cdot a(X_t) \cdot \sqrt{\frac{K \log(K/\delta)}{\tau}}\right\} \cdot \mathbb{1}\{a(X_t) < \sqrt{\epsilon}\}\right].$$

Next, bounding $a(X_t)$ by $\sqrt{\epsilon}$ inside the expectation and using the margin assumption (justified by (16) for our eventual optimal value $\epsilon = \left(\frac{\tau}{K \log(K/\delta)}\right)^{\frac{\alpha+d}{2+\alpha+d}}$), the above is bounded by:

$$c_4 \left(\sqrt{\frac{\epsilon K \log(K/\delta)}{\tau}}\right)^{\alpha+1}. \tag{22}$$

For the case where $a(X_t) \ge \sqrt{\epsilon}$, we again have by (20) that:

$$\mathbb{E}\left[\sigma_t^* \cdot \mathbb{1}\{0 < \delta_f(X_t) \le c_0 \sigma_t^*\} \cdot \mathbb{1}\{a(x_t) \ge \sqrt{\epsilon}\}\right] \le c_5 \sqrt{\frac{K \log(K/\delta)}{\tau}} \mathbb{E}_{X_t}[a(X_t)\mathbb{1}\{a(X_t) \ge \sqrt{\epsilon}\}]. \tag{23}$$

Note that the expectation on the R.H.S. is only over $X_t \sim Q_X$ since $a(\cdot)$ depends only on X_t. We use the following proposition to bound this integral over X_t.

Proposition 4 *For some $c_6.c_7 > 0$, we have for any $r \in \mathcal{R}$:*

$$\mathbb{E}_{X \sim Q_X}\left[\frac{1}{\sqrt{Q_X(T_r(X))}} \cdot \mathbb{1}\left\{\frac{1}{\sqrt{Q_X(T_r(X))}} \ge \sqrt{\epsilon}\right\}\right] \le c_6 \frac{r^{-d}}{\sqrt{\epsilon}} \tag{24}$$

$$\mathbb{E}_{X \sim Q_X}\left[\frac{1}{\sqrt{P_X(T_r(X))}} \cdot \mathbb{1}\left\{\frac{1}{\sqrt{P_X(T_r(X))}} \ge \sqrt{\epsilon}\right\}\right] \le c_7 \frac{r^{-d-\gamma}}{\sqrt{\epsilon}} \tag{25}$$

Proof We first show (24); (25) will be shown nearly identically. In an abuse of notation, let $a(X) \doteq Q_X(T_r(X))^{-1/2}$. We first use the tail probability formula for expectations:

$$\mathbb{E}_X[a(X) \cdot \mathbb{1}\{a(X) \ge \sqrt{\epsilon}\}] = \int_0^\infty Q_X(a(X) \cdot \mathbb{1}\{a(X) \ge \sqrt{\epsilon}\} \ge s)\, ds$$

Next, we observe that

$$a(X)\mathbb{1}\{a(X) \ge \sqrt{\epsilon}\} \ge s \iff a(X) \ge \sqrt{\epsilon} \vee s.$$

This gives us

$$\int_0^\infty Q_X(a(X)\mathbb{1}\{a(X) \ge \sqrt{\epsilon}\} \ge s)\, ds = \int_{\sqrt{\epsilon}}^\infty Q_X(a(X) \ge s)\, ds + \int_0^{\sqrt{\epsilon}} Q_X(a(X) \ge \sqrt{\epsilon})\, ds.$$

$$\le \int_{\sqrt{\epsilon}}^\infty \frac{\mathbb{E}_{Q_X}(a(X)^2)}{s^2}\, ds + \sqrt{\epsilon} \cdot \frac{\mathbb{E}_{Q_X}(a(X)^2)}{\epsilon}, \tag{26}$$

28

where we used Chebyshev's inequality above to bound the masses $Q_X(\cdot)$. To bound the second moments in (26), we observe a fairly standard implication of the *box cover dimension d* (Definition 4): for any $r \in \mathcal{R}$:

$$\mathbb{E}_{Q_X}(a^2(X)) = \mathbb{E}_{Q_X}\left(\frac{1}{Q_X(T_r(X))}\right) \le C_d \cdot r^{-d}.$$

Plugging the above into (26) gives us that

$$\mathbb{E}_X[a(X) \cdot \mathbb{1}\{a(X) \ge \sqrt{\epsilon}\}] \le c_6 \frac{r^{-d}}{\sqrt{\epsilon}}.$$

To show (25), we repeat the argument above using $b(X) \doteq P_X(T_r(X))^{-1/2}$ and the following analogous bound on $\mathbb{E}_{Q_X}(b^2(X))$ (using the definition of transfer exponent in Assumption 5):

$$\mathbb{E}_{Q_X}\left(\frac{1}{P_X(T_r(X))}\right) \le \mathbb{E}_{Q_X}\left(\frac{1}{C_\gamma \cdot r^\gamma \cdot Q_X(T_r(X))}\right) \le \frac{C_d}{C_\gamma} \cdot r^{-d-\gamma}.$$

■

Thus, by (24) of Proposition 4, (23) is bounded by

$$c_8 \sqrt{\frac{K \log(K/\delta)}{\tau}} \cdot \frac{(r_t^*)^{-d}}{\sqrt{\epsilon}}. \tag{27}$$

Then, (22) and (27) are balanced by setting $\epsilon \doteq \left(\frac{\tau}{K \log(K/\delta)}\right)^{\frac{\alpha+d}{2+\alpha+d}}$, which makes (19), and hence the regret at round t under event F_t, order $O\left(\left(\frac{K \log(K/\delta)}{\tau}\right)^{\frac{\alpha+1}{2+\alpha+d}}\right)$.

Next, we claim that (19) is bounded by

$$O\left(\left(\frac{K \log(K/\delta)}{n_P}\right)^{\frac{\alpha+1}{2+\alpha+d+\gamma}}\right).$$

For this, we can follow an identical line of reasoning as above, with the following substitutions:

(a) Replace the measure Q_X for P_X, and, thus, the quantity $a(X_t)$ by $b(X_t) \doteq P_X(T_{r_t^*}(X_t))^{-1/2}$.

(b) Use instead the threshold value $\epsilon \doteq \left(\frac{n_P}{K \log(K/\delta)}\right)^{\frac{\alpha+d+\gamma}{2+\alpha+d+\gamma}}$ in the analogue of (21).

(c) Use (25) instead of (24) from Proposition 4.

Thus, (19) is of order $O\left(\min\left(\left(\frac{K \log(K/\delta)}{n_P}\right)^{\frac{\alpha+1}{2+\alpha+d+\gamma}}, \left(\frac{K \log(K/\delta)}{\tau}\right)^{\frac{\alpha+1}{2+\alpha+d}}\right)\right)$.

Bounding the Regret on Bad Events and Summing the Regrets over t It remains to bound the regret under the low-probability event F_t^c and sum our bounds over $\tau \in [n_Q]$. By an integral approximation, we have

$$\sum_{\tau=K\log(K/\delta)}^{n_Q} \left(\frac{K\log(K/\delta)}{\tau}\right)^{\frac{\alpha+1}{2+\alpha+d}} \leq c_9 \int_{K\log(K/\delta)}^{n_Q} \left(\frac{K\log(K/\delta)}{z}\right)^{\frac{\alpha+1}{2+\alpha+d}} dz$$

If $\alpha \leq d + \alpha + 1$, this integral, for some $c_{10} > 0$, is bounded by

$$c_{10} n_Q \left(\frac{K\log(K/\delta)}{n_Q}\right)^{\frac{\alpha+1}{2+d}}.$$

Otherwise, it is bounded by $O(K\log(K/\delta))$.

Since the regret at round t is bounded by 1 on F_t^c, it remains to show $\mathbb{P}(F_t^c)$ is appropriately small. First, by the definition of F_t, the definition of event A_t, and Proposition 3, we have:

$$\mathbb{P}(F_t^c) \leq \mathbb{P}(A_t^c) + \mathbb{P}(E_t^c) + \mathbb{P}\left(\cup_{s=1}^t G_s^c\right) \leq \mathbb{P}(A_t^c) + \delta + 2 \cdot t \cdot \delta,$$

Summing $\delta + 2 \cdot t \cdot \delta$ over t accounts for the $O(n_Q n\delta)$ term in the desired regret bound of Theorem 1. Finally, to handle the term $\mathbb{P}(A_t^c)$, we use the definition of the transfer exponent (Assumption 5) and the definition of the support complexity (Definition 4):

$$\begin{aligned}
\mathbb{P}(A_t^c) &= \mathbb{P}\left(Q_X(T_{r_t^*}(X_t)) < \frac{8\log(1/\delta)}{\tau} \cap P_X(T_{r_t^*}(X_t)) < \frac{8\log(1/\delta)}{n_P}\right) \\
&\leq \mathbb{P}\left(Q_X(T_{r_t^*}(X_t)) < \frac{8\log(1/\delta)}{\tau} \cap C_\gamma \cdot (r_t^*)^\gamma \cdot Q_X(T_{r_t^*}(X_t)) < \frac{8\log(1/\delta)}{n_P}\right) \\
&= \sum_{B \in T_{r_t^*}} \mathbb{P}\left(Q_X(B) < \min\left(\frac{8\log(1/\delta)}{\tau}, \frac{8\log(1/\delta)}{n_P \cdot C_\gamma \cdot (r_t^*)^\gamma}\right) \cap B = T_{r_t^*}(X_t)\right) \\
&\leq C_d \cdot (r_t^*)^{-d} \cdot \min\left(\frac{8\log(1/\delta)}{\tau}, \frac{8\log(1/\delta)}{n_P \cdot C_\gamma \cdot (r_t^*)^\gamma}\right) \\
&\leq c_{11} \min\left(\left(\frac{K\log(K/\delta)}{\tau}\right)^{\frac{\alpha+2}{2+\alpha+d}}, \left(\frac{K\log(K/\delta)}{n_P}\right)^{\frac{\alpha+2}{2+\alpha+d+\gamma}}\right).
\end{aligned}$$

This is clearly smaller than the bound on the regret at time t we derived in the previous section. Thus, summing the above bound over τ using an integral approximation in the same manner as before, we see that $\sum_{t=n_P+1}^n \mathbb{P}(A_t^c)$ is of the right order. This concludes the proof of Theorem 1. ∎

Appendix C. Multiple Shifts

In this section, we give an extension of Theorem 1 to multiple distribution shifts. Let $\mathcal{P} \doteq \{P_j\}_{j=1}^N$ be a sequence of N initial distributions on the covariate-reward pair (X, Y). Suppose each P_j satisfies covariate shift with respect to the terminal distribution Q. We then consider bandits with a sequence of shifts

$$P_1 \to P_2 \to \cdots \to P_N \to Q.$$

In this setup, data from each P_j is observed for n_j consecutive rounds. Then, there are $n_P \doteq \sum_{j=1}^{N} n_j$ total rounds played under distributions from the class $\mathcal{P}$. We then consider the regret $\mathbf{R}_n^Q(\pi)$ of a policy π playing $n = n_P + n_Q$ total rounds, over the last n_Q rounds of data observed from Q.

Theorem 2 *Let π denote the procedure of Algorithm 1, ran, with parameter $\delta \in (0, 1)$, up till time $n > n_P \geq 0$, with $n_P, N, \{n_j\}_{j=1}^{N}$ all possibly unknown. Suppose the marginal of the covariate X under each P_j has unknown transfer exponent γ_j w.r.t. Q_X, that Q_X has (C_d, d) box dimension and that the average reward function f satisfies a margin condition with unknown α under Q_X. Let $n_Q \doteq n - n_P$ denote the (possibly unknown) number of rounds after the drifts, i.e., over the phase $X_t \sim Q_X$. Let $\overline{\gamma} = \sum_{j=1}^{N} \gamma_j \cdot \frac{n_j}{n_P}$. We have for some constant $C > 0$:*

$$\mathbb{E}\,\mathbf{R}_n^Q(\pi) \leq Cn_Q \left[\min\left(\left(\frac{K \log(K/\delta)}{n_P} \right)^{\frac{\alpha+1}{2+\alpha+d+\overline{\gamma}}}, \left(\frac{K \log(K/\delta)}{n_Q} \right)^{\frac{\alpha+1}{2+\alpha+d}} \right) + \frac{K \log(K/\delta)}{n_Q} + n\delta \right]$$

Proof [Proof Outline] The proof is almost identical to that of Theorem 1. Let r_t^* be the oracle level or the smallest level of $\mathcal{R}$ greater than or equal to

$$\min\left(\left(\frac{\log(K/\delta)}{n_P} \right)^{\frac{1}{2+\alpha+d+\overline{\gamma}}}, \left(\frac{\log(K/\delta)}{\tau} \right)^{\frac{1}{2+\alpha+d}} \right),$$

where, recall the notation $\tau \doteq t - n_P - 1$. Consider a round t for $\tau > K \log(K/\delta)$. For a fixed level $r \in \mathcal{R}$, let $n_P(r)$ be the covariate count in $T_r(X_t)$ from any of the distribution P_j for $j \in [N]$ or

$$n_P(r) \doteq \sum_{s \in [n_P]} \mathbb{1}\{X_s \in T_r(X)\}.$$

Then, similar to Proposition 3, we have that, at round t, with probability at least $1 - \delta$ with respect to the conditional distribution $\mathbf{X}_{t-1}|X_t$:

$$\phi_t(r_t) \leq 2\sqrt{\frac{8K \log(K/\delta)}{\max(\tau Q_X(T_{r_t^*}(X_t)), \mathbb{E}[n_P(r_t^*)])}} + 2\lambda r_t^*.$$

Next, observe that if P_j^X is the covariate marginal of P_j, then:

$$\mathbb{E}[n_P(r_t^*)] = n_P \cdot \left[\sum_{j=1}^{N} \frac{n_j}{n_P} \cdot P_j^X(T_{r_t^*}(X_t)) \right].$$

Then, following the same steps and notation as the proof of Theorem 1, let

$$a(X_t) \doteq \sqrt{\frac{1}{\sum_{j=1}^{N} \frac{n_j}{n_P} \cdot P_j^X(T_{r_t^*}(X_t))}}.$$

Then, using the definition of the transfer exponent (Definition 5) and Definition 4, we have an analogous inequality to that used in deriving (25) in Proposition 4

$$
\begin{aligned}
\mathbb{E}\left[a^2(X_t)\right] &= \mathbb{E}\left[\frac{1}{\sum_{j=1}^N \frac{n_j}{n_P} P_j^X(T_{r_t^*}(X_t))}\right] \\
&\leq c_{12}\mathbb{E}\left[\frac{1}{Q_X(T_{r_t^*}(X_t))\sum_{j=1}^N \frac{n_j}{n_P}(r_t^*)^{\gamma_j}}\right] \\
&\leq c_{13}\frac{1}{\sum_{j=1}^N \frac{n_j}{n_P}(r_t^*)^{\gamma_j}}\cdot(r_t^*)^{-d}.
\end{aligned}
$$

Next, since the function $x \mapsto r^x$ is convex for any $r > 0$, by Jensen's inequality we have

$$
\sum_{j=1}^N \frac{n_j}{n_P}\cdot(r_t^*)^{\gamma_j} \geq (r_t^*)^{\frac{1}{n_P}\sum_{j=1}^N n_j\gamma_j} = (r_t^*)^{\overline{\gamma}}.
$$

Thus, $\mathbb{E}[a^2(X_t)] \leq c_{13}(r_t^*)^{-d-\overline{\gamma}}$. This yields essentially the same bound as (25), except $\overline{\gamma}$ replaces γ. Using this in place of (25) in the proof of Theorem 1, we obtain the result. ∎

Appendix D. Lower Bound

Here, we establish that the bound of Theorem 1 is minimax optimal, up to $\log$ terms in the case where $K = 2$, over a continuum of regimes of choices of n_P, n_Q, γ, α.

Our strategy is to use *online-to-batch* conversion to convert an online algorithm with regret $R_{n_P,n}$ during the last n_Q rounds to a classifier with excess risk of order $R_{n_P,n}/n_Q$. This then implies a conversion from classification lower-bounds to bandits lower-bounds.

We note that *online-to-batch* conversion results – which we call as a black-box – are usually given for i.i.d. sequences of covariate-reward pairs, while we instead consider a setting with a shift in distribution $P \to Q$. Therefore, in much of what follows, we treat the first phase $\{(X_t, Y_t)\}_{t=1}^{n_P} \sim P^{n_P}$ as a separate input randomness Z, and apply conversion arguments to the second phase $\{(X_t, Y_t)\}_{t=n_P+1}^{n} \sim Q^{n_Q}$.

First, we claim a bandit policy π can be converted to an online classification algorithm where $\pi_t \in [K]$ indicates the predicted label for covariate X_t. This requires defining a reward Y^i for each label $i \in [K]$, which is done in Definition 9 below. To simplify notation, we will denote the set of $K = 2$ arms as $\{0, 1\}$.

Definition 9 (Conversion from Labels to Rewards) *In the case of binary classification with covariate $X \in \mathcal{X}$ and label $\tilde{Y} \in \{0, 1\}$, we define the reward of arm $i \in \{0, 1\}$ as $Y^i \doteq \mathbb{1}\{\tilde{Y} = i\}$. We use $\mathcal{T}$ to denote a class of tuples (P, Q) of distributions on the covariate-label pair $(X, \tilde{Y})$. Each distribution on $(X, \tilde{Y})$ then induces a distribution on the covariate-reward pair (X, Y). Let $\mathcal{T}'$ be the class of tuples of distributions on (X, Y), induced by $\mathcal{T}$.*

To simplify notation, in what follows, tuples (P, Q) will refer to either tuples in $\mathcal{T}$ or their one-to-one mapping to tuples in $\mathcal{T}'$, which will be clear from context.

We will also let $\{(X_t, \tilde{Y}_t)\}_{t=1}^{m}$ be a sequence of covariate-label pairs and let $\{(X_t, Y_t)\}_{t=1}^{m}$ be the sequence of corresponding covariate-reward pairs.

In this constructed bandits problem, the regression function of arm i is $f^i(x) = \mathbb{P}(\tilde{Y} = i | X = x)$. Next, let $h^*(x) \doteq \mathbb{1}\{f^1(x) \geq 1/2\} = \pi^*(x)$ be the Bayes classifier; the excess risk of a classifier h w.r.t. distribution Q is then given as:

$$\mathcal{E}_Q(h) \doteq \mathbb{E}_Q \left[\mathbb{1}\{h(X) \neq \tilde{Y}\} - \mathbb{1}\{h^*(X) \neq \tilde{Y}\} \right].$$

Consider an arbitrary online learner $\Lambda = \Lambda(Z)$, based on additional randomness Z independent of the training data. We let $\Lambda_1, \Lambda_2, \ldots$ denote the sequentially generated classifiers of Λ. We also define the *mistake count* $\mathcal{M}_m(\Lambda)$ over m rounds of Λ as:

$$\mathcal{M}_m(\Lambda) \doteq \sum_{t=1}^{m} \mathbb{1}\{\Lambda_t(X_t) \neq \tilde{Y}_t\} - \mathbb{1}\{h^*(X_t) \neq \tilde{Y}_t\}.$$

In expectation, we have that the mistake count $\mathcal{M}_m(\Lambda)$ is equal to the regret $\mathbf{R}_n^Q(\pi)$ of a bandits policy π when Λ is the online learner induced by π, via the conversion of Definition 9. Thus, in lower bounding $\mathcal{M}_m(\Lambda)$, we obtain a lower bound on the regret.

The next few definitions and results will be stated in terms of an arbitrary online learner Λ and, in Corollary 6, we will specialize Λ to the online learner induced by policy π.

First, we formalize the types of black-box guarantees on online-to-batch conversion our arguments will rely on. In what follows, let $\mathbf{X}_m \doteq \{X_t\}_{t=1}^m$ and $\mathbf{Y}_m \doteq \{Y_t\}_{t=1}^m$.

Definition 10 *In what follows, let $a \doteq \{a_m\}, b \doteq \{b_m\}$ denote bounded sequences in $[0, 1]$, indexed over $m \in \mathbb{N}$. An **online to batch conversion rate** is a mapping F from sequences $a \mapsto b$ such that the following holds:*

If there exists an online learner $\Lambda = \Lambda(Z)$, for additional randomness Z, which achieves expected mistake count $\mathbb{E}_{Z, \mathbf{X}_m, \mathbf{Y}_m} (\mathcal{M}_m(\Lambda)) \leq m \cdot a_m$ for some sequence a, then there exists a classifier $\hat{h} = \hat{h}(\Lambda)$ with excess risk $\mathbb{E}_{Z, \mathbf{X}_m, \mathbf{Y}_m} (\mathcal{E}_Q(\hat{h})) \leq (F(a))_m$.

*Now, for any $b = \{b_m\}$, define the **pseudo-inverse** $F^\dagger(b) \doteq \inf\{a \doteq \{a_m\} : (F(a))_m > b_m\}$, where the inf over a set of sequences is defined pointwise over $m \in \mathbb{N}$ (that is, $(F^\dagger(b))_m = \inf\{a_m : (F(a))_m > b_m\}$ for $m \in \mathbb{N}$).*

Next, we formally define the notion of a minimax lower bound for offline and online classification problems in terms of a rate $\{a_m\}$.

Definition 11 *Fix $n_P \in \mathbb{N}$. We say that the class $\mathcal{T}$ (of distribution pairs (P, Q)) has a **classification minimax lower bound** of $b = \{b_m\}$ if the following holds:*

For any $m \in \mathbb{N}$ and any classifier $\hat{h}$ learned on data $\{(X_t, \tilde{Y}_t)\}_{t=1}^m \sim Q^m$, and additional randomness $Z = \{(X_t', \tilde{Y}_t')\}_{t=1}^{n_P} \sim P^{n_P}$,

$$\sup_{(P,Q) \in \mathcal{T}} \mathbb{E}_{Z, \mathbf{X}_m, \mathbf{Y}_m}[\mathcal{E}_Q(\hat{h})] > b_m.$$

*Similarly, a class $\mathcal{T}$ has an **online minimax lower bound** of $a = \{a_m\}$ if the following holds:*

For any $m \in \mathbb{N}$ and any online learner $\Lambda = \Lambda(Z)$ trained on data $\{(X_t, \tilde{Y}_t)\}_{t=1}^m \sim Q^m$, and additional randomness $Z = \{(X_t', \tilde{Y}_t')\}_{t=1}^{n_P} \sim P^{n_P}$, we have:

$$\sup_{(P,Q) \in \mathcal{T}} \mathbb{E}_{Z, \mathbf{X}_m, \mathbf{Y}_m} [\mathcal{M}_m(\Lambda)] > m \cdot a_m.$$

Given an online to batch conversion rate, the next lemma allows us to deduce an online minimax lower bound from a classification minimax lower bound.

Lemma 5 (Minimax Lower Bound Conversion) *Suppose $b \doteq \{b_m\}$ denotes a classification minimax lower bound for the class $\mathcal{T}$. Then, if there exists an online to batch conversion rate F with $(F^\dagger(b))_m > 0$ for all $m \in \mathbb{N}$, we have that $\frac{1}{2} \cdot F^\dagger(b)$ is an online minimax lower bound for the class $\mathcal{T}$.*

Proof Consider an online learner $\Lambda = \Lambda(Z)$, with additional randomness $Z = \{(X_t', \tilde{Y}_t')\}_{t=1}^{n_P}$, with mistake count $a = \{a_m\}$. For contradiction, suppose there exists $m \in \mathbb{N}$ such that:

$$\sup_{(P,Q)\in\mathcal{T}} \mathbb{E}_{Z,\mathbf{X}_m,\mathbf{Y}_m}[\mathcal{M}_m(\Lambda)] \leq m \cdot a_m \leq m \cdot \frac{1}{2} \cdot (F^\dagger(b))_m < m \cdot (F^\dagger(b))_m.$$

Then, by the definition of F and the pseudo-inverse $F^\dagger$, there exists a classifier $\hat{h} = \hat{h}(\Lambda)$ such that:

$$\sup_{(P,Q)\in\mathcal{T}} \mathbb{E}_{Z,\mathbf{X}_m,\mathbf{Y}_m}[\mathcal{E}_Q(\hat{h})] \leq (F(a))_m \leq b_m.$$

This is a contradiction on b being a classification minimax lower bound for the class $\mathcal{T}$. ∎

We next specify the online to batch conversion rate F that we will use with Lemma 5.

Theorem 3 (Theorem 4 of (Cesa-Bianchi et al., 2004), paraphrased) *Let $\Lambda = \Lambda(Z)$ be an arbitrary online learner, trained on $\{(X_t, \tilde{Y}_t)\}_{t=1}^m$, with additional randomness Z. Then, for any $\delta \in (0, 1]$, there exists a classifier $\hat{h} = \hat{h}(\Lambda)$, trained on $\{(X_t, \tilde{Y}_t)\}_{t=1}^m$, such that:*

$$\mathbb{P}\left(\mathbb{P}_{(X,\tilde{Y})\sim Q}(\hat{h}(X) \neq \tilde{Y}) \geq \frac{1}{m}\sum_{t=1}^m \mathbb{1}(\Lambda_t(X_t) \neq \tilde{Y}_t) + 6\sqrt{\frac{\log(2(m+1)/\delta)}{m}} \,\middle|\, Z\right) \leq \delta.$$

Corollary 5 *Let $\Lambda = \Lambda(Z)$ be an online learner trained on data $\{(X_t, \tilde{Y}_t)\}_{i=1}^m$ with additional input Z. Then, there exists a classifier $\hat{h} = \hat{h}(\Lambda)$ such that for any distribution on $\mathbf{X}_m, \mathbf{Y}_m, Z$:*

$$\mathbb{E}_{Z,\mathbf{X}_m,\mathbf{Y}_m}[\mathcal{E}_Q(\hat{h})] \leq \frac{\mathbb{E}_{Z,\mathbf{X}_m,\mathbf{Y}_m}[\mathcal{M}_m(\Lambda)]}{m} + 6\sqrt{\frac{2\log(m\sqrt{3})}{m}} + \frac{1}{m}.$$

Proof Fix a value of Z and let the event A be as in Theorem 3:

$$A = \left\{\mathbb{P}_{(X,\tilde{Y})\sim Q}(\hat{h}(X) \neq \tilde{Y}) \geq \frac{1}{m}\sum_{t=1}^m \mathbb{1}(\Lambda_t(X_t) \neq \tilde{Y}_t) + 6\sqrt{\frac{\log(2(m+1)/\delta)}{m}} \,\middle|\, Z\right\}.$$

Then, letting $\delta = 1/m$ and conditioning on the event A, we have:

$$\mathbb{E}_{\mathbf{X}_m,\mathbf{Y}_m}[\mathcal{E}_Q(\hat{h}) \mid Z] \leq \mathbb{E}_{\mathbf{X}_m,\mathbf{Y}_m}\left[\frac{1}{m}\sum_{t=1}^m \mathbb{1}(\Lambda_t(X_t) \neq \tilde{Y}_t) - \mathbb{1}(h^*(X_t) \neq \tilde{Y}_t) \,\middle|\, Z\right]$$
$$+ 6\sqrt{\frac{2\log(m\sqrt{3})}{m}} + \frac{1}{m}$$
$$= \frac{\mathbb{E}_{\mathbf{X}_m,\mathbf{Y}_m}[\mathcal{M}_m(\Lambda) \mid Z]}{m} + 6\sqrt{\frac{2\log(m\sqrt{3})}{m}} + \frac{1}{m}.$$

34

Taking a further expectation with respect to Z on both sides of the inequality gives the desired result. ∎

Theorem 1 of Kpotufe and Martinet (2018) provides us the classification minimax lower-bound, which we restate it here.

Theorem 4 (Theorem 1 of Kpotufe and Martinet (2018)) *Let $\mathcal{T}'$ be the class of all tuples (P, Q) of distributions satisfying Assumption 2 and Definition 4, and Definitions 3 and 5, with some fixed parameters $(\lambda, C_d, d, C_\alpha, \alpha, \delta_0, C_\gamma, \gamma)$. In what follows, let $\mathcal{T}$ be the one-to-one mapping of $\mathcal{T}'$ to tuples of distributions on covariate-label pairs as in Definition 9. Then, there exists a constant $c > 0$ such that for any $n_P, n_Q \in \mathbb{N}$ and classifier $\hat{h}$ learned on $\{(X_t, \tilde{Y}_t)\}_{t=1}^{n_P} \sim P^{n_P}$ and $\{(X_t, \tilde{Y}_t)\}_{t=n_P+1}^{n_P+n_Q} \sim Q^{n_Q}$, we have:*

$$\sup_{(P,Q)\in\mathcal{T}} \mathbb{E}\left(\mathcal{E}_Q(\hat{h})\right) > c\left(n_P^{\frac{2+\alpha+d}{2+\alpha+d+\gamma}} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}}.$$

Next, we will take $b_m \doteq c\left(n_P^{\frac{2+\alpha+d}{2+\alpha+d+\gamma}} + m\right)^{-\frac{\alpha+1}{2+\alpha+d}}$ as our classification minimax lower-bound where m here stands for n_Q. Combining Lemma 5, Corollary 5, and Theorem 4, we obtain the following minimax lower bound for bandits:

Corollary 6 (Matching Lower Bounds over Given Regimes) *Let the class $\mathcal{T}'$ and the constant $c > 0$ be as in Theorem 4. Suppose that n_P, n_Q satisfy:*

$$6\sqrt{\frac{2\log(n_Q\sqrt{3})}{n_Q}} + \frac{1}{n_Q} < \frac{c}{2}\left(n_P^{\frac{2+\alpha+d}{2+\alpha+d+\gamma}} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}}. \tag{28}$$

Then, for any fixed such n_P, n_Q and any contextual bandits policy π, we have:

$$\sup_{(P,Q)\in\mathcal{T}'} \mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n}\left[\mathbf{R}_n^Q(\pi)\right] \geq \frac{c}{4}n_Q\left(n_P^{\frac{2+\alpha+d}{2+\alpha+d+\gamma}} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}}.$$

Proof Fix n_P, n_Q satisfying the inequality in (28) and let $n = n_P + n_Q$. Let π_O be the online learner induced by a policy π restricted to the Q phase $\{(X_t, Y_t)\}_{t=n_P+1}^{n}$ with additional randomness $Z = \{(X_t, Y_t)\}_{t=1}^{n_P}$. Then, by Corollary 5, we have there exists a classifier $\hat{h}$ such that:

$$\mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n}[\mathcal{E}_Q(\hat{h})] \leq \frac{\mathbb{E}_{\mathbf{X}_n, \mathbf{Y}_n}[\mathcal{M}_{n_Q}(\pi_O)]}{n_Q} + 6\sqrt{\frac{2\log(n_Q\sqrt{3})}{n_Q}} + \frac{1}{n_Q}.$$

We then have that the map F, defined below on a sequence $a = \{a_m\}$, is an online to batch conversion rate:

$$(F(a))_m \doteq a_m + 6\sqrt{\frac{2\log(m\sqrt{3})}{m}} + \frac{1}{m},$$

Let $b_m \doteq c \left(n_P^{\frac{2+\alpha+d}{2+\alpha+d+\gamma}} + m \right)^{-\frac{\alpha+1}{2+\alpha+d}}$ be as in Theorem 4. Then, by Theorem 4 and Lemma 5, we have:

$$\sup_{(P,Q)\in\mathcal{T}} \mathbb{E}_{\mathbf{X}_n,\mathbf{Y}_n}[\mathcal{M}_{n_Q}(\pi_O)] \geq \frac{n_Q \cdot (F^\dagger(b))_{n_Q}}{2}.$$

Next, we observe:

$$\mathbb{E}_{\mathbf{X}_n,\mathbf{Y}_n}[\mathcal{M}_{n_Q}(\pi_O)] = \mathbb{E}_{\mathbf{X}_n,\mathbf{Y}_n} \left(\sum_{t=1}^{n_Q} \mathbb{1}(\pi_{O,t}(X_t) \neq \tilde{Y}_t) - \mathbb{1}(h^*(X_t) \neq \tilde{Y}_t) \right)$$

$$= \mathbb{E}_{\mathbf{X}_n,\mathbf{Y}_n} \left(\sum_{t=1}^{n_Q} Y_t^{\pi_t^*(X_t)} - Y_t^{\pi_t(X_t)} \right)$$

$$= \mathbb{E}_{\mathbf{X}_n,\mathbf{Y}_n} \left[\mathbf{R}_n^Q(\pi) \right].$$

Thus,

$$\sup_{(P,Q)\in\mathcal{T}'} \mathbb{E}_{\mathbf{X}_n,\mathbf{Y}_n} \left[\mathbf{R}_n^Q(\pi) \right] \geq \frac{n_Q \cdot (F^\dagger(b))_{n_Q}}{2}$$

$$\geq \frac{n_Q}{2} \cdot \left(b_{n_Q} - 6\sqrt{\frac{2\log(n_Q\sqrt{3})}{n_Q}} - \frac{1}{n_Q} \right)$$

$$\geq \frac{n_Q}{2} \cdot \frac{c}{2} \left(n_P^{\frac{2+\alpha+d}{2+\alpha+d+\gamma}} + n_Q \right)^{-\frac{\alpha+1}{2+\alpha+d}}.$$

∎

Remark 10 *The inequality in* (28) *corresponds to the regime* $n_P = \tilde{O} \left(n_Q^{\frac{2+\alpha+d+\gamma}{2+2\alpha}} \right)$ *with* $\alpha < d$. *In particular this includes the following subregimes.*

- **Performance on Q depends mostly on covariates** $X_t \sim Q_X, t > n_P$. *This is the sub-regime where* $n_P \lesssim n_Q^{\frac{2+\alpha+d+\gamma}{2+\alpha+d}}$, *roughly, that is when (in the upper-bound of Theorem 1)* $\min(n_P^{-\frac{\alpha+1}{2+\alpha+d+\gamma}}, n_Q^{-\frac{\alpha+1}{2+\alpha+d}}) = n_Q^{-\frac{\alpha+1}{2+\alpha+d}}$, *i.e.,* past experience *under P is too short to significantly influence regret under Q. The lower-bound of Corollary 6 then is of the form*

$$\sup_{(P,Q)\in\mathcal{T}'} \mathbb{E} \left[\mathbf{R}_n^Q(\pi) \right] \gtrsim n_Q \cdot n_Q^{-\frac{\alpha+1}{2+\alpha+d}},$$

which confirms that the threshold $n_P = \tilde{O} \left(n_Q^{\frac{2+\alpha+d+\gamma}{2+\alpha+d}} \right)$ *(on when past experience is too short) is indeed tight.*

- **Performance on Q depends mostly on covariates** $X_t \sim P_X, t \leq n_P$. *This is the subregime where*

$$n_Q^{\frac{2+\alpha+d+\gamma}{2+\alpha+d}} \lesssim n_P \lesssim n_Q^{\frac{2+\alpha+d+\gamma}{2+2\alpha}}.$$

In other words $\min(n_P^{-\frac{\alpha+1}{2+\alpha+d+\gamma}}, n_Q^{-\frac{\alpha+1}{2+\alpha+d}}) = n_P^{-\frac{\alpha+1}{2+\alpha+d+\gamma}}$, *i.e., past experience under P significantly influences regret under Q. The lower-bound of Corollary 6 then is of the form*

$$\sup_{(P,Q)\in\mathcal{T}'} \mathbb{E}\left[\mathbf{R}_n^Q(\pi)\right] \gtrsim n_Q \cdot n_P^{-\frac{\alpha+1}{2+\alpha+d+\gamma}},$$

that is, the upper-bounds of Theorem 1 are again tight up to $\log$ *factors.*

Appendix E. Bounded Mass Assumption

Here, we consider the *strong density condition* of Remark 1, differing from our more relaxed condition on the marginal distribution Q_X in Definition 4. The strong density assumption ensures that Q_X has good coverage of $[0,1]^d$. It holds, for instance, if Q_X has lower-bounded Lebesgue density on $[0,1]^d$. We note that, unlike the notion of support dimension used in previous sections (Definition 4), the "dimension" of our support now coincides with the ambient dimension, denoted D in previous sections. This is defined formally as follows:

Assumption 3 (Mass under Q) $\exists\, C_d > 0$ *s.t.,* $\forall\, \ell_\infty$ *balls* $B \subset [0,1]^d$ *of diameter* $r \in (0,1]$:

$$Q_X(B) \geq C_d \cdot r^d.$$

Under this assumption, we obtain similar regret rates as Theorem 1 with d now being defined as in Assumption 3. However, as discussed in Remark 8, the analysis differs heavily from that of Theorem 1 and is closer in spirit to that of Perchet and Rigollet (2013). We first consider the case of a single shift $P \to Q$ and handle multiple shifts in Theorem 6, by a similar extension to that made in Appendix C.

Theorem 5 *Let π denote the procedure of Algorithm 1, ran, with parameter $\delta \in (0,1)$, up till time $n > n_P \geq 0$, with n_P possibly unknown. Suppose P_X has unknown transfer exponent γ w.r.t. Q_X, that Q_X satisfies Assumption 3 with (C_d, d), and that the average reward function f satisfies a margin condition with unknown α under Q_X. Let $n_Q \doteq n - n_P$ denote the (possibly unknown) number of rounds after the drift, i.e., over the phase $X_t \sim Q_X$. We have for some constant $C > 0$:*

$$\mathbb{E}\, \boldsymbol{R}_n^Q(\pi) \leq Cn_Q\left[\min\left(\left(\frac{K\log(K/\delta)}{n_P}\right)^{\frac{\alpha+1}{2+d+\gamma}}, \left(\frac{K\log(K/\delta)}{n_Q}\right)^{\frac{\alpha+1}{2+d}}\right) + \frac{K\log(K/\delta)}{n_Q} + n\delta\right]$$

Corollary 7 *Under the setup of Theorem 1, letting $\delta = O(1/n^2)$ yields:*

$$\mathbb{E}\, \boldsymbol{R}_n^Q(\pi) \leq Cn_Q\left[\min\left(\left(\frac{K\log(Kn)}{n_P}\right)^{\frac{\alpha+1}{2+d+\gamma}}, \left(\frac{K\log(Kn)}{n_Q}\right)^{\frac{\alpha+1}{2+d}}\right) + \frac{K\log(Kn)}{n_Q}\right].$$

E.1. Outline of the Proof of Theorem 5

The first part of the proof is nearly identical to the proof of Theorem 1. Lemmas 1, 2, and 3 remain true without modification. Thus, we have, at round t with chosen bin $B \doteq T_{r_t}(X_t)$, that with probability at least $1 - \delta$ with respect to the distribution $\mathbf{Y}_{t-1}|\mathbf{X}_t$:

$$\forall x \in B, i \in \mathcal{I}_B : |\hat{f}_t^i(B) - f^i(x)| \leq 4\lambda r_t.$$

Additionally, Proposition 1 and Corollaries 3 and 4 still hold so that the regret and margin at round t are both bounded by $16\lambda r_t$. Thus, following the intuition of Remark 8, it suffices to show r_t is of optimal regression order.

Showing r_t is of Optimal Regression Order. As in Section 4.2, for each round t, define the event G_t on which the high-probability bound on the regression error holds or

$$G_t = \left\{\forall i \in \mathcal{I}_B, x \in B : |\hat{f}_t^i(B) - f^i(x)| \leq 4\lambda r_t, B = T_{r_t}(X_t)\right\}.$$

Recall that this is the "good" event on which we can identify unfavorable arms. Recall also, from Section 4.2, that we define $\tau \doteq t - 1 - n_P$ to simplify notation.

Lemma 6 *Fix a round t with observed covariate X_t. Then, for some $c_{14} > 0$, with probability at least $1 - \delta$ w.r.t. the distribution of $\mathbf{X}_{t-1}|X_t$, we have*

$$r_t \leq c_{14} \min\left(\left(\frac{K\log(K/\delta)}{n_P}\right)^{\frac{1}{2+d+\gamma}}, \left(\frac{K\log(K/\delta)}{\tau}\right)^{\frac{1}{2+d}}\right).$$

Proof It suffices to show

$$r_t \leq c_{14} \begin{cases} \left(\frac{K\log(K/\delta)}{n_P}\right)^{\frac{1}{2+d+\gamma}} & \text{when } n_P > K\log(K/\delta) \\ \left(\frac{K\log(K/\delta)}{\tau}\right)^{\frac{1}{2+d}} & \text{when } \tau > K\log(K/\delta) \end{cases},$$

since this implies the desired result for any n_P, τ. We first show that when $\tau > K\log(K/\delta)$:

$$r_t \leq c_{14}\left(\frac{K\log(K/\delta)}{\tau}\right)^{\frac{1}{2+d}}.$$

The other bound on r_t involving n_P will follow similarly, with the appropriate modifications. First, we simplify notation for the sake of this proof: at round t, it will be understood that the observed covariate is represented by $X \doteq X_t$. We also let $n(r) \doteq n_r(X)$ be the covariate count for *a fixed arbitrary* level r.

First, by Assumption 3 and the definition of transfer exponent (Definition 5), we have for some $c_{15} > 0$:

$$\mathbb{E}[n(r)] = n_P P_X(T_r(X)) + \tau Q_X(T_r(X)) \geq c_{15}(n_P r^{d+\gamma} + \tau r^d). \tag{29}$$

In fact, without loss of generality, we can assume

$$c_{15} \leq C_d^{\frac{2+d}{d}}. \tag{30}$$

38

Thus, by a Chernoff bound similar to that used in the proof of Proposition 3, for any fixed level r satisfying $n(r) \geq K \log(1/\delta)$, we have with probability at least $1 - \delta$ that

$$\sqrt{\frac{K \log(K/\delta)}{n(r)}} \leq \sqrt{\frac{8K \log(K/\delta)}{c_{15} \tau r^d}}. \tag{31}$$

Now, let $r_t^* \in \mathcal{R}$ be the smallest level greater than or equal to

$$c_{15}^{-\frac{1}{2+d}} \left(\frac{8K \log(K/\delta)}{\tau} \right)^{\frac{1}{2+d}}.$$

To put things simply, $r_t^* \propto \left(\frac{K \log(K/\delta)}{\tau} \right)^{\frac{1}{2+d}}$. Then, it suffices to show $r_t \leq r_t^*$. For the next part of the proof, we define $n_Q(r)$ as the covariate count coming exclusively from Q_X for any given level r:

$$n_Q(r) \doteq \sum_{s \in [\tau]} \mathbb{1}\{X_{n_P + s} \in T_r(X)\}.$$

Then, we claim r_t^* satisfies $n_Q(r_t^*) \geq \log(1/\delta)$ with probability at least $1 - \delta$, so that (31) holds with probability at least $1 - 2\delta$ for $r = r_t^*$, by the aforementioned Chernoff bound. Since $\tau > 8K \log(K/\delta)$ by hypothesis, we have by (30) and Assumption 3 that

$$\mathbb{E}[n_Q(r)] \geq C_d \cdot \tau \cdot (r_t^*)^d \geq C_d \cdot \tau \cdot c_{15}^{-\frac{d}{2+d}} \left(\frac{8K \log(K/\delta)}{\tau} \right)^{\frac{d}{2+d}} \geq 8 \log(1/\delta).$$

Thus, by a Chernoff bound, we have

$$Q_X(n_Q(r_t^*) < \log(1/\delta)) \leq Q_X \left(n_Q(r_t^*) < \frac{1}{2} \mathbb{E}[n_Q(r)] \right) \leq \exp \left(-\frac{1}{8} \mathbb{E}[n_Q(r)] \right) \leq \delta.$$

Then, we have that by virtue of how r_t^* is defined, with probability at least $1 - 2\delta$:

$$r_t^* \geq \sqrt{\frac{8K \log(K/\delta)}{c_{15} \tau (r_t^*)^d}} \geq \sqrt{\frac{8K \log(K/\delta)}{n(r_t^*)}}.$$

This gives us that $r_t \leq r_t^*$ by the minimization criteria for selecting r_t on Line 7 of Algorithm 1, as desired.

The other inequality

$$r_t \leq c_{14} \left(\frac{K \log(K/\delta)}{n_P} \right)^{\frac{1}{2+d+\gamma}},$$

can be shown in an identical fashion to the case above with the appropriate modifications: specifically, τ is replaced with n_P, (29) is further lower bounded by $c_{15} n_P r^{d+\gamma}$, and $n_Q(r)$ is replaced with $n_P(r)$ which is defined as the bin covariate counts from distribution P:

$$n_P(r) \doteq \sum_{s \in [n_P]} \mathbb{1}\{X_s \in T_r(X)\}.$$

$\blacksquare$

Cumulative Regret Bound. Next, we put the previous conclusions together to bound the cumulative regret by bounding the regret accrued at each round t and then summing over $t \in \{n_P+1, \dots, n\}$. Similarly to the proof of Theorem 1, for rounds $s < t$, define the event E_s as the event on which the bound of Lemma 1 holds or $E_s \doteq G_s$. For round t, define the event E_t as the event on which the bounds of Lemma 1 and Lemma 6 hold or:

$$E_t \doteq G_t \cap \left\{ r_t \le c_{14} \min \left(\left(\frac{K \log(K/\delta)}{n_P} \right)^{\frac{1}{2+d+\gamma}}, \left(\frac{K \log(K/\delta)}{\tau} \right)^{\frac{1}{2+d}} \right) \right\}.$$

To sum the regrets across time t, the argument will involve conditioning on the event $\cap_{s=1}^{t} E_s$, on which (a) Algorithm 1 correctly eliminates arms and (b) r_t is of the optimal order.

To this end, let $F_t \doteq \cap_{s=1}^{t} E_s$. Also, to simplify notation, let U_t denote

$$U_t \doteq c_{14} \min \left(\left(\frac{K \log(K/\delta)}{n_P} \right)^{\frac{1}{2+d+\gamma}}, \left(\frac{K \log(K/\delta)}{\tau} \right)^{\frac{1}{2+d}} \right).$$

Recall U_t is the earlier high-probability upper bound on r_t in the definition of E_t. If $n_Q \le K \log(K/\delta)$, we are already done since the regret is then bounded by $K \log(K/\delta)$, which is the right order. Assume for the rest of the proof that $K \log(K/\delta) < n_Q$.

To later use the margin condition (Definition 3), we require that $U_t \lesssim \delta_0$ (where δ_0 is as in Definition 3). To ensure this, we need only constrain our analysis to rounds for which $\tau \gtrsim K \log(K/\delta)$, which is not an issue since the regret of any $O(K \log(K/\delta))$ rounds is of the desired order. More precisely, let τ_0 be the largest positive integer satisfying

$$c_{16} \left(\frac{K \log(K/\delta)}{\tau_0} \right)^{\frac{1}{2+d}} > \delta_0,$$

where $c_{16} > 0$ will be determined later. The regret for the first τ_0 rounds among rounds $\{n_P + 1, \dots, n\}$ is $O(K \log(K/\delta))$ which is always of the right order. For the rest of the proof, we now assume that the round t is such that $\tau > \tau_0$ and $c_{16} U_t \le \delta_0$.

Next, let the event H_t be

$$H_t \doteq \{ |\hat{f}_t^{(1)}(B) - \hat{f}_t^{(2)}(B)| \le 8\lambda r_t, B = T_{r_t}(X_t) \}.$$

Conditioned on $\mathbf{X}_t$, H_t is the event where one arm remains in contention at round t according to Line 11 of Algorithm 1. For the remainder of the proof, let B be the bin that was selected at round t given an understood value of $\mathbf{X}_t$. To further simplify notation, let $\mathbf{X} \doteq \mathbf{X}_{t-1}$ and $\mathbf{Y} \doteq \mathbf{Y}_{t-1}$ since we are fixing t momentarily here.

Consider the expected regret of pulling arm π_t at round t, and decomposing it by conditioning the events F_t and H_t:

$$\mathbb{E} f^{(1)}(X_t) - f^{\pi_t}(X_t) = \mathbb{E}_{X_t} \left[\mathbb{E}_{\mathbf{X},\mathbf{Y}|X_t} (f^{(1)}(X_t) - f^{\pi_t}(X_t)) \cdot (\mathbb{1}\{F_t\} + \mathbb{1}\{F_t^c\}) \cdot (\mathbb{1}\{H_t\} + \mathbb{1}\{H_t^c\}) \right].$$

The above gives us three different cases depending on whether event $F_t \cap H_t$, or $F_t \cap H_t^c$, or F_t^c holds.

1) Suppose event $F_t \cap H_t$ holds. Suppose also that there is a suboptimal arm $i \in \mathcal{I}_B$ for which $f^i(X_t) < f^{(1)}(X_t)$, lest the regret at time t be zero. Then, by Corollary 4, we have for $c_{16} = c_{14} \cdot 16\lambda$ and any $j \in \mathcal{I}_B$:

$$|f^{(1)}(X_t) - f^j(X_t)| \leq 16\lambda r_t \leq c_{16} U_t.$$

Furthermore:

$$0 < |f^{(1)}(X_t) - f^{(2)}(X_t)| \leq 16\lambda r_t \leq c_{16} U_t.$$

This above inequality happens with probability at most $C_\alpha (c_{16} U_t)^\alpha$, under $X_t \sim Q_X$, by the margin condition (Definition 3). Thus, we have

$$\mathbb{E}_{X_t} \mathbb{E}_{\mathbf{X},\mathbf{Y}|X_t} \left(f^{(1)}(X_t) - f^j(X_t) \right) \cdot \mathbb{1}\{F_t \cap H_t\} \leq C_\alpha (c_{16})^{\alpha+1} U_t^{\alpha+1}.$$

2) Next, on $F_t \cap H_t^c$, the pointwise regret is zero by Corollary 3 since $\mathcal{I}_B$ must contain the optimal arm at X_t and no other arms.

3) On F_t^c, the pointwise regret is trivially bounded above by 1. By Lemma 6, this happens with probability at most

$$\mathbb{P}(F_t^c) \leq \mathbb{P}\left(\cup_{s=1}^t E_s^c \right) \leq \sum_{s=1}^t 2\delta.$$

Thus, $\mathbb{E}_{X_t} \mathbb{E}_{\mathbf{X},\mathbf{Y}|X_t} \left(f^{(1)}(X_t) - f^j(X_t) \right) \cdot \mathbb{1}\{F_t^c\} \leq 2t\delta.$

Next, we put the three cases above together. We have that, for some $c_{17} > 0$, the cumulative regret over the rounds $\{\tau_0 + 1, \ldots, n\}$ is then at most

$$c_{17} \left[\sum_{\tau=\tau_0+1}^{n_Q} \min \left(\left(\frac{K \log(K/\delta)}{n_P} \right)^{\frac{\alpha+1}{2+d+\gamma}}, \left(\frac{K \log(K/\delta)}{\tau} \right)^{\frac{\alpha+1}{2+d}} \right) + (n_P + \tau)\delta \right].$$

Taking the sum over the last term on the R.H.S. above, we have $\sum_{\tau=1}^{n_Q}(n_P + \tau)\delta = O(n_Q n \delta)$. For the remaining term in the sum, it suffices to bound

$$\sum_{\tau=\tau_0+1}^{n_Q} \left(\frac{K \log(K/\delta)}{\tau} \right)^{\frac{\alpha+1}{2+d}}.$$

As in the proof of Theorem 1, by an integral approximation, we have since $\tau_0 \propto K \log(K/\delta)$:

$$\sum_{\tau=\tau_0+1}^{n_Q} \left(\frac{K \log(K/\delta)}{t} \right)^{\frac{\alpha+1}{2+d}} \leq c_{18} \int_{K \log(K/\delta)}^{n_Q} \left(\frac{K \log(K/\delta)}{z} \right)^{\frac{\alpha+1}{2+d}} dz$$

If $\alpha \leq d + 1$, the above integral, for some $c_{18} > 0$, is bounded by

$$c_{19} n_Q \left(\frac{K \log(K/\delta)}{n_Q} \right)^{\frac{\alpha+1}{2+d}}.$$

Otherwise, said integral is bounded by $O(K \log(K/\delta))$. This concludes the proof of Theorem 5. ∎

E.2. Multiple Shifts under Bounded Mass

Here, we derive an analogue to Theorem 2 under the bounded mass assumption and the same setup as Appendix C. The proof is nearly identical as that of Theorem 2, relying on the same bound on $r^{\sum_{j=1}^{N} \frac{n_j}{n_P} \cdot \gamma_j}$.

Theorem 6 *Let π denote the procedure of Algorithm 1, ran, with parameter $\delta \in (0, 1)$, up till time $n > n_P \geq 0$, with $n_P, N, \{n_j\}_{j=1}^{N}$ all possibly unknown (see definitions in Appendix C). Suppose the marginal of the covariate X under each P_j has unknown transfer exponent γ_j w.r.t. Q_X, that Q_X satisfies Assumption 3 with (C_d, d), and that the average reward function f satisfies a margin condition with unknown α under Q_X. Let $n_Q \doteq n - n_P$ denote the (possibly unknown) number of rounds after the drifts, i.e., over the phase $X_t \sim Q_X$. Let $\overline{\gamma} = \sum_{j=1}^{N} \gamma_j \cdot \frac{n_j}{n_P}$. We have for some constant $C > 0$:*

$$\mathbb{E}\, \boldsymbol{R}_n^Q(\pi) \leq C n_Q \left[\min \left(\left(\frac{K \log (K/\delta)}{n_P} \right)^{\frac{\alpha+1}{2+\alpha+d+\overline{\gamma}}}, \left(\frac{K \log (K/\delta)}{n_Q} \right)^{\frac{\alpha+1}{2+\alpha+d}} \right) + \frac{K \log (K/\delta)}{n_Q} + n\delta \right]$$

Proof [Proof Idea] Consider a round $t > n_P$ and any level $r \in \mathcal{R}$ such that $n_r(X_t) > \log(1/\delta)$. Similarly to the proof of Lemma 6, by a Chernoff bound and the definition of the transfer exponent (Definition 5), we have that the covariate count $n_r(X_t)$, with probability at least $1 - \delta$ satisfies

$$n_r(X_t) \gtrsim \left[(t - 1 - n_P)r^d + \sum_{j=1}^{N} n_j \cdot r^{d+\gamma_j} \right].$$

Next, by Jensen's inequality we have

$$\sum_{j=1}^{N} \frac{n_j}{n_P} \cdot r^{d+\gamma_j} \geq r^{d+\overline{\gamma}}.$$

Thus, combining the above two inequalities, $n_r(X_t) \gtrsim n_P \cdot r^{d+\overline{\gamma}}$. Using this bound, we can derive a generalization of Lemma 6 where γ is replaced by $\overline{\gamma}$. All other parts of the proof of Theorem 5 remain the same in showing Theorem 6. ∎

E.3. Lower Bound under Bounded Mass Assumption

Finally, we present an analogue to Corollary 6 under the bounded mass assumption, thus showing the regret bound of Theorem 5 is minimax optimal in a certain regime. The proof is identical to that of Corollary 6, relying on the same online-to-batch conversion procedure. The same theorem cited in Appendix D (Theorem 1 of Kpotufe and Martinet, 2018) provides us the classification minimax lower-bound required for the online-to-batch conversion.

Corollary 8 (Matching Lower Bounds over Given Regimes) *Let $\mathcal{T}'$ be the class of all tuples (P, Q) of distributions satisfying Assumptions 2 and 3, and Definitions 3 and 5, with some fixed parameters $(\lambda, C_d, d, C_\alpha, \alpha, \delta_0, C_\gamma, \gamma)$. Suppose that n_P, n_Q satisfy:*

$$6\sqrt{\frac{2 \log(n_Q\sqrt{3})}{n_Q}} + \frac{1}{n_Q} < \frac{c}{2} \left(n_P^{\frac{2+d}{2+d+\gamma}} + n_Q \right)^{-\frac{\alpha+1}{2+d}}. \tag{32}$$

Then, for any fixed such n_P, n_Q and any contextual bandits policy π, we have:

$$\sup_{(P,Q)\in\mathcal{T}'} \mathbb{E}_{\boldsymbol{X}_n,\boldsymbol{Y}_n}\left[\boldsymbol{R}_n^Q(\pi)\right] \geq \frac{c}{4}n_Q\left(n_P^{\frac{2+d}{2+d+\gamma}} + n_Q\right)^{-\frac{\alpha+1}{2+d}}.$$

Remark 11 *Similarly to Remark 10, the inequality in (32) corresponds to the regime $n_P = \tilde{O}\left(n_Q^{\frac{2+d+\gamma}{2+2\alpha}}\right)$ with $\alpha < d/2$.*