The Annals of Statistics
2024, Vol. 52, No. 1, 207–232
https://doi.org/10.1214/23-AOS2341

TRANSFER LEARNING FOR CONTEXTUAL MULTI-ARMED BANDITS

BY CHANGXIAO CAI[1,a], T. TONY CAI[2,b] AND HONGZHE LI[3,c]

[1]*Department of Industrial and Operations Engineering, University of Michigan,* [a]*cxcai@umich.edu*
[2]*Department of Statistics and Data Science, University of Pennsylvania,* [b]*tcai@wharton.upenn.edu*
[3]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,* [c]*hongzhe@pennmedicine.upenn.edu*

Motivated by a range of applications, we study in this paper the problem of transfer learning for nonparametric contextual multi-armed bandits under the covariate shift model, where we have data collected from source bandits before the start of the target bandit learning. The minimax rate of convergence for the cumulative regret is established and a novel transfer learning algorithm that attains the minimax regret is proposed. The results quantify the contribution of the data from the source domains for learning in the target domain in the context of nonparametric contextual multi-armed bandits.

In view of the general impossibility of adaptation to unknown smoothness, we develop a data-driven algorithm that achieves near-optimal statistical guarantees (up to a logarithmic factor) while automatically adapting to the unknown parameters over a large collection of parameter spaces under an additional self-similarity assumption. A simulation study is carried out to illustrate the benefits of utilizing the data from the source domains for learning in the target domain.

1. Introduction. Inspired by the human intelligence of leveraging prior experiences to tackle novel problems, *transfer learning*, which aims to improve the learning performance in a target domain by transferring the knowledge contained in different but related source domains, has become an active and promising area of research in machine learning. Transfer learning has achieved significant success in a wide range of practical applications such as computer vision (Quattoni, Collins and Darrell (2008), Kulis, Saenko and Darrell (2011), Li et al. (2013)), genomic and genetic studies (Wang et al. (2019), Peng et al. (2021)) and medical imaging (Raghu et al. (2019), Yu et al. (2022)) to mention a few. We refer interested readers to Pan and Yang (2009), Weiss, Khoshgoftaar and Wang (2016) for detailed surveys on transfer learning. Motivated by the success in these applications, substantial progress has also been made recently in the theoretical quantification for transfer learning in supervised and unsupervised settings. A partial list of examples includes classification (Cai and Wei (2021), Kpotufe and Martinet (2021), Maity, Sun and Banerjee (2020), Reeve, Cannings and Samworth (2021)), high-dimensional linear regression (Li, Cai and Li (2022)), graphical model (Li, Cai and Li (2023)) and nonparametric regression (Cai and Pu (2021a), Ma, Pathak and Wainwright (2023), Pathak, Ma and Wainwright (2022)).

In this paper, we consider transfer learning for nonparametric contextual multi-armed bandits. Since the seminal formulation in Robbins (1952), the multi-armed bandit (MAB) and its various extensions have been widely used in numerous fields related to sequential decision-making, including personalized medicine (Tewari and Murphy (2017), Rabbi et al. (2018), Zhou et al. (2019), Shrestha and Jain (2021), Demirel, Celik and Tekin (2022)), recommendation system (Li et al. (2010), Kallus and Udell (2020)) and dynamic pricing (Rothschild

Received November 2022; revised November 2023.
MSC2020 subject classifications. Primary 62G08; secondary 62L12.
Key words and phrases. Contextual multi-armed bandit, transfer learning, covariate shift, minimax rate, regret bounds, adaptivity, self-similarity.

(1974), Kleinberg and Leighton (2003), Wang, Chen and Simchi-Levi (2021)). In the classical nonparametric contextual multi-armed bandit problem, a decision maker sequentially and repeatedly chooses an arm from a set of available arms and receives a random reward generated by the selected arm. The goal is to develop an arm selection policy that maximizes the expected cumulative rewards over a finite time horizon. Motivated by the common scenario where the decision maker often has access to side information to assist arm selection, covariates are introduced to encode the features that affect the reward yielded by each arm at each time step. The expected reward of each arm conditioned on the context is assumed to follow a nonparametric form, allowing for a more flexible and robust formulation in real-world applications.

However, collecting enough reward feedback to design an optimal arm selection strategy is often challenging in practice. For instance, the contextual multi-armed bandit framework is widely used in precision medicine that aims to tailor the medical care to each patient (Rindtorff et al. (2019), Zhou et al. (2019)). Every time a patient visits, a healthcare provider needs to determine a treatment based on the patient's profile, including genetics, biomarkers, environment and demographic information. The objective is to optimize the post-treatment health outcomes of all patients. In this case, patient profiles, treatments and health outcomes correspond to covariates, arms and rewards, respectively. However, there is no shortage of cases where biomedical data of minority populations are underrepresented in certain healthcare institutions (Sudlow et al. (2015)). Given this limited availability of data in clinical research, it is common to resort to the healthcare records of other patients with similar characteristics. In such a scenario, the task of transfer learning for contextual multi-armed bandits naturally arises.

In addition to applications in precision medicine, contextual multi-armed bandits are also frequently used in online recommendation systems that seek to learn dynamically the preferences of an individual customer for a collection of products based on the demographics and purchase histories (Agrawal et al. (2019), Kallus and Udell (2020)). Since each user can only purchase a small set of products, the availability of transactional data is often limited in practice. Therefore, it is natural to explore the information of different but related customers, in order to better predict the possibility of an individual customer purchasing a specific product. Similarly, anomaly detection systems often rely on a limited number of interactions with human experts for verification to maximize accurate anomaly detection. Due to its tradeoff between exploration (e.g., investigation of various anomalies to improve prediction) and exploitation (e.g., queries of the most suspicious one), anomaly detection has also been formulated by a contextual multi-armed bandit framework (Ding, Li and Liu (2019), Soemers et al. (2018)). In credit card fraud identification systems, if the account history of a single card holder is short, it would be advisable to utilize the information of similar types of transactions and customers to increase the detection accuracy of fraudulent transactions.

In this work, we consider the following setting of transfer learning for contextual multi-armed bandits. Let Q and P be two probability distributions over $[0,1]^d \times [0,1]^K$ that generate a sequence of independent random vectors $(X_t, Y_t^{(1)}, \ldots, Y_t^{(K)})_{t\geq 1}$ associated with a contextual K-armed bandit. Here, $(X_t)_{t\geq 1}$ is a sequence of i.i.d. random vectors in $\mathcal{X} := [0,1]^d$ representing the covariates. For each $1 \leq k \leq K$ and $t \geq 1$, $Y_t^{(k)}$ is a random variable in $[0,1]$ indicating the reward yielded by arm k at time t, with conditional expectation given by $\mathbb{E}[Y_t^{(k)} \mid X_t] = f_k(X_t)$, where $f_k : \mathcal{X} \to [0,1]$ is referred to as a reward function. The bandit game operates as follows: at each time step t, given side information X_t, the decision maker pulls one of the K arms, denoted by π_t and receives the random reward $Y_t^{(\pi_t)}$. Suppose that before the Q-bandit game starts, we have access to precollected samples $\{(X_i^P, \pi_i^P, Y_i^{(\pi_i^P)})\}_{i=1}^{n_P}$, generated from a source bandit with underlying distribution P.

Throughout the paper, Q refers to the target distribution about which we wish to make statistical inferences; P stands for the source distribution from which we have collected data to improve the decision-making under Q. We use the Q-bandit (resp., P-bandit) to represent the bandit with distribution Q (resp., P). In addition, we use superscripts Q and P to refer to quantities associated with the Q-bandit and the P-bandit, respectively. Our goal is to design a policy $\pi = \{\pi_t\}_{t\geq 1}$ that maximizes the expected cumulative rewards under the target distribution Q. The performance of the policy π can be measured by its cumulative regret over n_Q time steps, given by

$$\mathbb{E}\big[R_{n_Q}(\pi)\big] := \mathbb{E}\bigg[\sum_{t=1}^{n_Q}\big(Y_t^{Q,(\pi^\star(X_t^Q))}\big(X_t^Q\big) - Y_t^{Q,(\pi_t(X_t^Q))}\big(X_t^Q\big)\big)\bigg], \tag{1}$$

where $\pi^\star$ is the oracle policy with complete knowledge of the reward functions $\{f_k\}_{k=1}^K$. One can expect that as long as distributions P and Q are similar, the source data from the P-bandit can improve the decision-making in the Q-bandit. Therefore, it is natural to quantify the improvement in the cumulative regret, which can be viewed as the amount of information transferred from the source distribution P to the target distribution Q.

This paper focuses on transfer learning under the covariate shift model, where the marginal distributions of covariates P_X and Q_X differ (i.e., $P_X \neq Q_X$), but the conditional reward distributions of $Y^{(k)}$ given X are identical under P and Q (i.e., $P_{Y^{(k)}|X} = Q_{Y^{(k)}|X}$) for all $1 \leq k \leq K$. This framework is well motivated by many practical applications, typically arising from the scenarios when the same study is conducted among different populations. For instance, healthcare providers in a hospital may utilize medical records from other healthcare centers to better guide medical treatments. While the patient characteristics (captured by the marginal context distributions) tend to differ across different hospitals, given the same patient profile, the effects of the same treatment (which can be modeled as the conditional reward distributions) in various medical institutions can be identical in many cases. Therefore, it is natural to model this scenario as a covariate shift.

The covariate shift model hinges on the characterization of the similarity between the marginal distributions P_X and Q_X. Various assumptions on the similarity have been proposed in the literature. In the present paper, we adopt the concept of *transfer exponent*—introduced in Kpotufe and Martinet (2021) to study transfer learning for nonparametric classification—that measures the discrepancy between P_X and Q_X in terms of the ball mass ratio of the respective distributions. It is assumed that there exists a transfer exponent $\gamma \geq 0$ such that $P_X(B(x,r)) \gtrsim r^\gamma Q_X(B(x,r))$ holds for any ℓ_∞-ball of center $x \in \mathcal{X}$ and radius $r \in (0,1]$ (see Definition 1). Informally, the transfer exponent γ gauges how locally singular P_X is with respect to Q_X. When γ is small, this condition ensures that the source data adequately covers important regions under the target distribution Q (i.e., with large Q_X mass), thus facilitating the transfer of information from the source domain. In a recent study by Suk and Kpotufe (2021), contextual multi-armed bandits were considered in a setting where the distribution of covariates changes over time. The authors employed a similar framework to capture the distribution shift. They proposed an algorithm that provably achieves the near-optimal minimax regret while automatically adapting to the unknown change point in time and covariate shift level. However, the results therein suffer some deficiencies. First, the proposed algorithm is designed for Lipschitz reward functions (i.e., smoothness parameter $\beta = 1$, see Section 2 for the formal definition). Therefore, it falls short of accommodating the general settings $\beta \neq 1$. Moreover, this limitation raises a more challenging question—is it possible to design a data-driven procedure that can adapt to the smoothness level, which is typically *a priori* unknown in practice?

Moving beyond concerns about smoothness, it is noteworthy that Suk and Kpotufe (2021) focused on a scenario where the decision maker is allowed to explore the bandit game amid a covariate shift actively. However, it is more common in practice that one does not have the freedom to interact with the source bandit. Instead, one has to rely on a fixed precollected batch data set. In this setup, it is natural to expect a potentially larger regret, as the arm selection policy in the source data set might be uninformative (e.g., a suboptimal arm is predominantly selected in the source data set). To account for this challenge, we introduce an *exploration coefficient* κ (see Definition 2) that quantifies the extent to which each arm is explored in the source bandit. Specifically, the exploration coefficient κ ensures that each arm is pulled with probability at least κ/K across the covariate space in the source data set. Consequently, when κ is not vanishingly small, each arm is explored more or less sufficiently, thereby enabling the extraction of valuable information about the reward functions from the source domain. A natural question arising from such a scenario is: what is the minimax regret in the transfer learning setting when dealing with precollected batch data? Finally, we note a logarithmic gap exists between the upper and lower bounds in Suk and Kpotufe (2021), and it remains unknown whether the minimax lower bound on the regret can be attained.

In the present paper, we aim to address the following questions: given a precollected source data set, what is the minimax rate of convergence of the regret for nonparametric contextual multi-armed bandits in the covariate shift setting? Can we design a rate-optimal policy that achieves the minimax regret? Moreover, is it possible to develop a data-driven procedure that achieves near-optimal statistical guarantees while at the same time automatically adapting to the unknown smoothness of the reward functions and the covariate shift of the source distribution? Encouragingly, the answers to these questions are affirmative.

1.1. *Main contribution*. Our main contribution is twofold. We first establish the minimax rate of convergence of the cumulative regret for nonparametric contextual multi-armed bandits under the covariate shift model. In addition to the standard assumptions that the reward functions are β-Hölder (see Assumption 1) and the target distribution Q satisfies a margin assumption with parameter α (see Assumption 2), we assume that the source and target distributions P and Q satisfy transfer learning conditions with transfer exponent γ and exploration coefficient κ (see Definitions 1 and 2). Given n_P source samples collected from the source bandit, we show that the minimax regret over n_Q time steps in the target Q-bandit is of order $n_Q(n_Q + (\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}})^{-\frac{\beta(1+\alpha)}{d+2\beta}}$. In the classical setting where one has no auxiliary data from the source domain, that is, $n_P = 0$, the minimax regret is known to be of order $n_Q^{1-\frac{\beta(1+\alpha)}{d+2\beta}}$. Therefore, the term $(\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}}$ in the minimax regret captures the contribution from the source data to the target bandit, which depends on the amount of covariate shift between P_X and Q_X as well as the degree of arm exploration in the source data set.

We also develop a novel transfer learning algorithm and prove that it achieves the minimax regret. However, the constructed procedure depends on the knowledge of the smoothness and transfer learning parameters. Unfortunately, it has been widely recognized that adaptation to unknown smoothness is generally infeasible in nonparametric bandit problems (Locatelli and Carpentier (2018), Gur, Momeni and Wager (2022), Cai and Pu (2022b)). To this end, we choose to focus on the bandits with reward functions that satisfy the self-similarity condition—an assumption extensively used in the statistics literature (Picard and Tribouley (2000), Giné and Nickl (2010)). We develop a data-driven algorithm and show that it simultaneously achieves the near-optimal minimax regret at the penalty of an additional logarithmic factor over a large class of parameter spaces. Moreover, we demonstrate that the self-similarity assumption does not decrease the complexity of the problem by establishing the minimax lower bound that remains the same as in the general case.

1.2. *Related works*.

Contextual multi-armed bandits. The framework of contextual multi-armed bandits was first introduced in Woodroofe (1979). Among the parametric approaches, an important line of work assumes a linear reward function (Abe and Long (1999), Auer (2002), Bastani and Bayati (2020), Goldenshluger and Zeevi (2013)). In this setup, Bastani, Bayati and Khosravi (2021) proposed a rate-optimal greedy strategy with regret logarithmic in the length of the time horizon. For nonparametric contextual multi-armed bandits, it is typical to assume Hölder smooth reward functions. Yang and Zhu (2002) developed a greedy policy and showed that its regret goes to zero as the time horizon tends to infinity. Rigollet and Zeevi (2010) studied the two-armed bandits and proposed an upper-bound-confidence (UCB) type policy that attains a near-optimal minimax regret. This result was further refined in Perchet and Rigollet (2013) where a rate-optimal policy called ABSE was proposed in the multi-armed setting. Additionally, Reeve, Mellor and Brown (2018) combined a UCB-type policy with the nearest neighbor method to design a near-optimal algorithm capable of adapting to the low intrinsic dimension of contexts. It is worth noting that all the aforementioned methods are tailored for β-Hölder smooth reward with $\beta \in (0, 1]$. Hu, Kallus and Mao (2022) extended the theory to accommodate smoother reward functions with $\beta > 1$.

Adaptivity. It has been demonstrated in various bandit settings that adaptation to the unknown smoothness of reward functions is generally impossible. This means that no policy can achieve the minimax regrets simultaneously over different classes of reward functions (Locatelli and Carpentier (2018), Gur, Momeni and Wager (2022), Cai and Pu (2022b)). We note that this phenomenon is closely related to the impossibility of constructing adaptive confidence intervals in nonparametric function estimation (Low (1997), Cai and Low (2004), Cai (2012)). Fortunately, adaptive statistical inference can be accomplished under certain shape constraints, such as monotonicity and convexity (Cai, Low and Xia (2013), Hengartner and Stark (1995), Dümbgen (1998), Genovese and Wasserman (2005)). Self-similarity—first introduced by Picard and Tribouley (2000) for adaptive nonparametric confidence intervals—is another widely used condition that allows adaptivity. This concept finds applications in various fields, including density estimation (Giné and Nickl (2010)), sparse regression (Nickl and van de Geer (2013)) and ℓ_p-confidence sets (Nickl and Szabó (2016)). It was first introduced to the nonparametric contextual bandit setting by Qian and Yang (2016), where a UCB-type policy based on Lepski's method (Lepski, Mammen and Spokoiny (1997)) was proposed and shown to achieve the minimax regret up to a logarithmic factor. The drawback, however, is that its cost of adaptation tends to infinity as the covariate dimension grows. Gur, Momeni and Wager (2022) improved upon this result by reducing the adaptation cost to a logarithmic factor independent of the dimension.

Transfer learning. Transfer learning has been explored using information measures such as KL-divergence and total variation to quantify the distinction between target and source distributions (Ben-David et al. (2006), Blitzer et al. (2007), Mansour, Mohri and Rostamizadeh (2009)). Generalization bounds are then established based on these metrics. Despite its generality, such results are often not tight when applied to specific statistical models. Recent work has imposed more structured assumptions on the similarity between target and source distributions, such as covariate shift and posterior drift (Cai and Wei (2021), Cai and Pu (2021a), Hanneke and Kpotufe (2019), Kpotufe and Martinet (2021), Maity, Sun and Banerjee (2020), Reeve, Cannings and Samworth (2021)), thereby leading to more refined theoretical guarantees. Finally, our work is also closely related to hybrid reinforcement learning that aims to combine offline data sets with online interaction to improve statistical/computational efficiency (Ross and Bagnell (2012), Xie et al. (2021), Song et al. (2022), Wagenmaker and Pacchiano (2023), Li et al. (2023), Nakamoto et al. (2023)).

1.3. *Organization.* The rest of the paper is organized as follows. Section 2 formulates the problem and introduces definitions and assumptions. We then establish the minimax optimal rate of the regret and develop a rate-optimal algorithm in Section 3. In Section 4, we propose a data-driven adaptive procedure that achieves the minimax regret up to logarithmic factors. The proofs of our theorems, technical lemmas and numerical experiments are deferred to the Supplementary Material (Cai, Cai and Li (2024)). We conclude with a discussion of future directions in Section 5.

1.4. *Notation.* For any $a, b \in \mathbb{R}$, we define $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. We denote by $\|\cdot\|_2$ and $\|\cdot\|_\infty$ the ℓ_2 norm and the ℓ_∞ norm, respectively. The notation $B_\infty(x, r) := \{y : \|y - x\|_\infty \le r\}$ refers to the ℓ_∞ ball of center x and radius r, and we define the shorthand $B(x, r) := B_\infty(x, r)$. Denote by $[K] := \{1, 2, \ldots, K\}$. We use $\mathbb{1}\{\cdot\}$ to represent the indicator function, and we define $\log^+(x) := \log(x) \vee 1$. Let $\mathsf{supp}(\cdot)$ denote the support of any probability distribution. For any distributions P, Q, the notation $\mathsf{KL}(P\|Q)$ stands for the KL-divergence. For any $a \in \mathbb{R}$, denote by $\lfloor a \rfloor$ (resp., $\lceil a \rceil$) the largest (resp., smallest) integer that is strictly smaller (resp., larger) than a. The notation $\mathbb{N}$ stands for the set of the natural numbers, and we denote $\mathcal{X} := [0, 1]^d$. Throughout the paper, we denote by C or c some constants independent of n_P and n_Q, which may vary from line to line.

For any two functions $f(n), g(n) > 0$, the notation $f(n) \lesssim g(n)$ (resp., $f(n) \gtrsim g(n)$) means that there exists a constant $C > 0$ such that $f(n) \le Cg(n)$ (resp., $f(n) \ge Cg(n)$). The notation $f(n) \asymp g(n)$ means that $C_0 f(n) \le g(n) \le C_1 f(n)$ holds for some constants $C_0, C_1 > 0$. In addition, $f(n) = o(g(n))$ means that $\limsup_{n\to\infty} f(n)/g(n) = 0$, $f(n) \ll g(n)$ means that $f(n) \le c_0 g(n)$ for some small constant $c_0 > 0$, and $f(n) \gg g(n)$ means that $f(n) \ge c_1 g(n)$ for some large constant $c_1 > 0$.

2. Problem formulation.

2.1. *Transfer learning for nonparametric contextual multi-armed bandits.* Let Q be a probability distribution over $\mathcal{X} \times [0, 1]^K$ that generates a sequence of independent random vectors $(X^Q, Y^{Q,(1)}, \ldots, Y^{Q,(K)})$. At each time point t, based on the covariate $X_t^Q \in \mathcal{X}$ drawn from the marginal distribution Q_X, a decision maker selects an arm $k \in [K]$ and receives a random reward $Y_t^{Q,(k)} \in [0, 1]$ associated with the chosen arm according to the conditional distribution $Q_{Y^{(k)}|X_t^Q}$. We assume that for any $k \in [K]$ and $t \ge 1$, the random reward $Y_t^{Q,(k)}$ is a random variable with conditional expectation given by

$$\mathbb{E}[Y_t^{Q,(k)} \mid X_t^Q] = f_k^Q(X_t^Q),$$

where $f_k^Q : \mathcal{X} \to [0, 1]$ is an unknown function called a reward function. A policy π is a collection of functions $\{\pi_t\}_{t\ge1}$ where $\pi_t : \mathcal{X} \to [K]$ prescribe the arm to pull at time t.

In the context of transfer learning, we assume that the decision maker is given a batch data set $\mathcal{D}^P := \{(X_i^P, \pi_i^P, Y_i^{P,(\pi_i^P)})\}_{i=1}^{n_P}$. This data set is collected from a contextual K-armed bandit over n_P rounds, of which the underlying probability distribution P generates a sequence of independent random vectors $(X^P, Y^{P,(1)}, \ldots, Y^{P,(K)}) \in \mathcal{X} \times [0, 1]^K$. Here, $X_i^P \in \mathcal{X}$ represents the covariate observed at time i, policy $\pi_i^P : \mathcal{X} \to [K]$ denotes the selected arm at time i, and $Y_i^{P,(\pi_i^P)}$ corresponds to the observed random reward at time i. Similar to the Q-bandit, it is assumed that for any $k \in [K]$ and $i \ge 1$, the random reward $Y_i^{P,(k)}$ of the P-bandit obeys

$$\mathbb{E}[Y_i^{P,(k)} \mid X_i^P] = f_k^P(X_i^P),$$

for an unknown function $f_k^P : \mathcal{X} \to [0, 1]$. Throughout the paper, we denote $n := n_Q \vee n_P$.

As mentioned in the Introduction, this paper focuses on the covariate shift model. To be specific, it is assumed that the marginal distributions of covariates in the P-bandit and Q-bandit are different (i.e., $P_X \neq Q_X$) while the distributions of rewards conditioned on the covariate value are identical (i.e., $P_{Y^{(k)}|X} = Q_{Y^{(k)}|X}$ for all $1 \leq k \leq K$). In particular, the latter implies that the reward functions of the two bandits are also identical. We denote these common reward functions as $f_k(x) := f_k^P(x) \equiv f_k^Q(x)$ for all $k \in [K]$ and $x \in \mathcal{X}$.

Recall that $\pi^\star$ is the oracle policy with access to full knowledge of the reward functions $\{f_k\}_{k=1}^K$. It is straightforward to see that given a covariate value x, the oracle policy $\pi^\star$ selects any arm with the largest expected reward, with ties broken arbitrarily. In other words,

$$\pi^\star(x) \in \operatorname*{argmax}_{k\in[K]} f_k(x).$$

Therefore, for any policy $\pi = \{\pi_t\}_{t\geq 1}$, the regret of π in the Q-bandit defined in (1) has the following expression:

$$\begin{aligned}\mathbb{E}\big[R_{n_Q}(\pi)\big] &= \mathbb{E}\left[\sum_{t=1}^{n_Q}\big(f_{\pi^\star(X_t^Q)}\big(X_t^Q\big) - f_{\pi_t(X_t^Q)}\big(X_t^Q\big)\big)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n_Q}\Big(\max_{k\in[K]} f_k\big(X_t^Q\big) - f_{\pi_t(X_t^Q)}\big(X_t^Q\big)\Big)\right].\end{aligned} \tag{2}$$

In the remainder of the paper, we may drop the subscript n_Q whenever there is no confusion.

Finally, we would like to emphasize that the policy π_t at time t depends on both the observations of the Q-data prior to time t (i.e., $\{(X_i^Q, \pi_i, Y_i^{Q,(\pi_i)})\}_{i=1}^{t-1} \cup \{X_t^Q\}$) and the complete P-data (i.e., $\{(X_i^P, \pi_i^P, Y_i^{P,(\pi_i^P)})\}_{i=1}^{n_P}$).

2.2. *Assumptions*. It is noteworthy that one cannot hope to distinguish the optimal arm of a contextual multi-armed bandit with arbitrary covariate and reward distributions. In order to guarantee provably small cumulative regrets, we impose the following model assumptions, which have become standard in the literature on nonparametric contextual multi-armed bandits (Rigollet and Zeevi (2010), Perchet and Rigollet (2013)).

We begin by imposing a Hölder smoothness assumption on the reward functions $\{f_k\}_{k=1}^K$ as follows.

ASSUMPTION 1 (Smoothness). The reward functions $\{f_k\}_{k=1}^K$ are (β, C_β)-Hölder continuous for some constants $0 < \beta \leq 1$, $C_\beta > 0$, that is, for any $k \in [K]$,

$$\big|f_k(x) - f_k\big(x'\big)\big| \leq C_\beta \big\|x - x'\big\|_\infty^\beta \quad \forall x, x' \in \mathcal{X}.$$

REMARK 1. By the equivalence of ℓ_p norms ($p \geq 1$) in $\mathcal{X}$, the results in this work continue to hold if ℓ_∞ norm is replaced with any ℓ_p norm ($p \geq 1$).

REMARK 2. Given the primary focus of this work is to illustrate the potential for reducing cumulative regrets through the utilization of source data, we confine our attention to the case $0 < \beta \leq 1$ for simplicity of presentation. Notably, the insights and findings here can be extended to accommodate the case $\beta > 1$. A detailed discussion of this generalization is deferred to Section 5.

Next, it is natural to expect that the gap between the reward functions is a pivotal measure of a contextual multi-armed bandit problem's complexity. To this end, let $f_{(1)}$ (resp., $f_{(2)}$) denote the pointwise maximum (resp., the second pointwise maximum) of the reward functions

$\{f_k\}_{k=1}^K$, namely

$$f_{(1)}(x) := \max_{k\in[K]} f_k(x),$$

and

$$f_{(2)}(x) := \begin{cases} \max_{k\in[K]} \{f_k(x) : f_k(x) < f_{(1)}(x)\} & \text{if } \min_{k\in[K]} f_k(x) \neq \max_{k\in[K]} f_k(x), \\ f_{(1)}(x) & \text{otherwise.} \end{cases}$$

Equipped with these notation, we introduce the following margin assumption to quantify the interplay between the reward gap and the covariate distribution in the target bandit Q.

ASSUMPTION 2 (Margin). There exist constants $\alpha \geq 0$, $C_\alpha > 0$ such that the reward functions $\{f_k\}_{k=1}^K$ and marginal distribution Q_X satisfy

$$Q_X\big(0 < f_{(1)}(X) - f_{(2)}(X) \leq \delta\big) \leq C_\alpha \delta^\alpha \quad \forall 0 < \delta \leq 1.$$

Assumption 2 bears a resemblance to the margin condition initially introduced in classification (Mammen and Tsybakov (1999), Tsybakov (2004), Audibert and Tsybakov (2007)), and has been widely used in contextual multi-armed bandits (Goldenshluger and Zeevi (2009), Perchet and Rigollet (2013)) and dynamic treatment regimes (Qian and Murphy (2011), Luedtke and van der Laan (2016), Shi, Lu and Song (2020)). Roughly speaking, the margin condition encodes the distribution behavior of the contexts near the decision boundary. It is easy to see that the margin condition is inherently satisfied for $\alpha = 0$ and holds for $\alpha = 1$ when $f_{(1)}(X) - f_{(2)}(X)$ has a bounded probability density near zero. If the margin parameter α is large, it implies that with low probability the reward gap between the optimal arm and other arms is small but bounded away from zero. This means that the reward functions of different arms are well separated over a region of large probability mass, which in turn, reduces the difficulty of distinguishing between the arms.

REMARK 3. As discussed in Perchet and Rigollet ((2013), Proposition 3.1), when $\alpha\beta > d$, there exists a single arm that dominates others across the entire covariate space. In such a case, a contextual multi-armed bandit problem degenerates into a static multi-armed bandit problem that falls beyond the scope of interest for this work. Therefore, we shall assume $\alpha\beta \leq d$ in the remainder of the paper.

REMARK 4. Assumptions 1 and 2 are commonly found in the nonparametric contextual multi-armed bandit literature. However, the assumption that all reward functions are smooth may not be valid in certain practical applications. In such cases, it might be possible to relax Assumption 1 by imposing the smoothness assumption solely on the best few arms while introducing a more delicate condition on the reward gap to replace Assumption 2. We leave the development of suitable models for such settings to future investigation.

In addition, we impose a regularity condition on the marginal distribution Q_X. It ensures that the support of Q_X is regular and that the density is bounded away from zero and infinity on the support.

ASSUMPTION 3 (Bounded density). There exist constants $\overline{q} > \underline{q} > 0$ such that $\underline{q} r^d \leq Q_X(B(x, r)) \leq \overline{q} r^d$ for any $x \in \mathsf{supp}(Q_X)$ and $r \in (0, 1]$.

With these conditions pertaining to the target bandit Q in place, let us turn to the assumptions that enable reliable transfer learning. As previously discussed in Section 1, we focus on the covariate shift setting and deploy the concept of the transfer exponent. This notion was originally introduced in Kpotufe and Martinet (2021), and numerous variants have emerged in the transfer learning literature (Hanneke and Kpotufe (2019), Pathak, Ma and Wainwright (2022), Cai and Wei (2021), Suk and Kpotufe (2021)).

DEFINITION 1 (Transfer exponent). Define the transfer exponent $\gamma \in \mathbb{R}_+ \cup \{0, \infty\}$ of P_X with respect to Q_X to be the smallest constant such that

$$P_X\big(B(x,r)\big) \geq c_\gamma r^\gamma Q_X\big(B(x,r)\big) \quad \forall x \in \mathsf{supp}(Q_X), r \in (0,1], \tag{3}$$

for some constant $0 < c_\gamma \leq 1$.

Note that for an arbitrary probability distribution pair (P, Q), condition (3) always holds with $\gamma = \infty$. Also, given that the radius r is always upper bounded by one in $[0,1]^d$, the probability mass $P_X(B(x,r))$ increases as γ approaches to 0. Intuitively, this implies that the source data cover a larger subset of the covariate regime of interest, allowing more effective information to be transferred from the source distribution P to the target distribution Q.

We now give an example for Definition 1. Let Q_X be the uniform distribution over $[0,1]$. Suppose the density function $p_X(x)$ of P_X takes the form $p_X(x) = Cx^\gamma$ for some normalization constant $C > 0$. Then it is easy to verify that the transfer exponent of P_X with respect to Q_X equals γ. We refer to Kpotufe and Martinet (2021) for a more in-depth discussion of this transfer exponent.

In addition, the covariate-arm pairs $\{(X_i^P, \pi_i^P)\}_{i=1}^{n_P}$ in the precollected source data set are assumed to be generated i.i.d. according to $X_i^P \sim P_X$ and $\pi_i^P(X_i^P) \sim \mu(\cdot \mid X_i^P)$, where $\{\mu(\cdot \mid x)\}_{x \in \mathcal{X}}$ is a collection of probability distributions over the arm set $[K]$. We make a note that this i.i.d. assumption prevails in the literature on bandits and reinforcement learning where one seeks to exploit offline data (Rashidinejad et al. (2022)), which is well motivated by the data randomization procedure in experience replay (Mnih et al. (2015)). To gauge the degree of exploration over the arm set in the source data set, we introduce the exploration coefficient as defined below.

DEFINITION 2 (exploration coefficient). Define the exploration coefficient $\kappa \in [0,1]$ of a collection of distributions over the arm set $\{\mu(\cdot \mid x)\}_{x \in \mathcal{X}}$ with respect to Q_X as

$$\kappa := \inf_{k \in [K],\, x \in \mathsf{supp}(Q_X)} K\mu(k \mid x). \tag{4}$$

Note that κ/K is the lowest probability of an arm being selected over the support of Q_X in the P-data. Intuitively, when the exploration coefficient κ is not vanishingly small, each arm has been extensively tested by the source policy within the regions of interest. This, in turn, provides the decision maker with greater confidence regarding the reward function associated with each arm, thereby facilitating the decision-making in the target bandit. We make a note that Definition 2 exhibits close ties to the positivity assumption in dynamic treatment regimes (Shi, Lu and Song (2020)), as well as the notion of uniformly bounded concentrability coefficient in offline reinforcement learning (Munos (2007), Farahmand, Szepesvári and Munos (2010), Chen and Jiang (2019), Xie and Jiang (2021), Wagenmaker and Pacchiano (2023)).

Finally, we assume the number of arms K is constant throughout this paper.

Denote by $\Pi(K, \beta, C_\beta, \alpha, C_\alpha, \underline{q}, \overline{q}, \gamma, c_\gamma, \kappa)$ the class of nonparametric contextual K-armed bandits that satisfy Assumptions 1–3 and Definitions 1–2. Here and throughout, the paper, we may use the shorthands $\Pi(K, \beta, \alpha, \gamma, \kappa)$ and Π if there is no confusion.

3. Minimax rate of convergence. In this section, we establish the minimax regret for transfer learning under the covariate shift model and develop a rate-optimal procedure Algorithm 1 to achieve the minimax regret.

3.1. *Algorithm.* The key to solving nonparametric contextual multi-armed bandit problems lies in accurately estimating the values of the reward functions $\{f_k\}_{k=1}^K$ at each observed point X_t^Q. Inspired by the success of Rigollet and Zeevi (2010), Perchet and Rigollet (2013) in the classical setting, the high-level idea of Algorithm 1 is fairly straightforward. It dynamically partitions the covariate space $\mathcal{X}$ into a set of hypercubes (bins), and uses local constant estimators for the reward functions in each bin. This reduces the original contextual multi-armed bandit into a collection of (static) multi-armed bandits (without covariates). Subsequently, we can apply a successive elimination algorithm within each bin separately and independently. To be more specific, Algorithm 1 generates a sequence of nested partitions $\{\mathcal{L}_t\}_{t\geq 1}$ of the covariate space $\mathcal{X}$ over time, where the partition $\mathcal{L}_t$ at time t consists of a set of bins (of potentially different side lengths) in $\mathcal{X}$. Here, for any nonnegative integer $l \geq 0$, we define a collection of bins $\mathcal{B}_l := \{B_{\mathsf{k}}\}_{\mathsf{k}\in[2^l]^d}$ where

$$B_{\mathsf{k}} := \big\{x \in \mathcal{X} : (k_i - 1)2^{-l} \leq x_i \leq k_i 2^{-l},\ k_i \in \big[2^l\big],\ i \in [d]\big\} \quad \forall \mathsf{k} = (k_1, \ldots, k_d). \tag{5}$$

Throughout the paper, we use $|B|$ to denote the side length of any bin B, that is, $|B| = 2^{-l}$ for any $B \in \mathcal{B}_l$. As an important observation, for any bin $B \in \mathcal{L}_t$ with $Q_X(B) > 0$, if we restrict our focus to samples of which the covariates fall in bin B, it is not hard to see that the corresponding observed rewards $(Y_s^{Q,(k)}(B))_{s\geq 1}$ generated by arm k are i.i.d. random variables with expectation equal to the conditional expectation of the reward of arm k over bin B, namely

$$\bar{f}_k^Q(B) := \mathbb{E}\big[f_k\big(X_t^Q\big) \,|\, X_t^Q \in B\big] = \frac{1}{Q_X(B)} \int_B f_k(x)\, \mathrm{d}Q_X(x). \tag{6}$$

As a consequence, at each time t, given the covariate X_t^Q, we first find the bin B in the current partition $\mathcal{L}_t$ that contains X_t^Q. We then invoke Procedure 1—a transfer learning procedure tailored to multi-armed bandits that yields a policy $\{\widetilde{\pi}_s(B)\}_{s\geq 1}$ for bin B—to determine π_t, that is, the arm to pull at time t.

In order to present the policy $\pi = \{\pi_t\}_{t\geq 1}$ generated by Algorithm 1, we introduce several notation. First, for any $x \in \mathcal{X}$ and $t \geq 1$, let $B_t(x) \in \mathcal{L}_t$ denote the bin in the partition $\mathcal{L}_t$ at time t such that $x \in B_t(x)$. If there are multiple bins, we choose $B_t(x)$ to be the one whose center is closest to the origin. Next, for any bin B and time $t \geq 1$, denote by $N_t(B)$ the number of times the covariate fell into B prior to time t, that is, $N_t(B) := \sum_{1\leq s\leq t} \mathbb{1}\{X_s^Q \in B\}$. With these definitions in place, the policy $\pi = \{\pi_t\}_{t\geq 1}$ yielded by Algorithm 1 can be described by $\pi_t = \widetilde{\pi}_{N_t(B)}(B)$ with $B = B_t(X_t^Q)$ for any time $t \geq 1$.

Before delving into the details of Algorithm 1, we pause to introduce some additional notation. As Algorithm 1 maintains an adaptive partition of the covariate space $\mathcal{X}$ over time, it is convenient to describe the partition using a tree. To this end, denote by $\mathcal{T}^{(l)}$ the perfect tree with root node $\mathcal{X}$ and depth $l \geq 0$, where there are 2^{id} nodes in each depth $0 \leq i < l$ and each node B represents a bin in set $\mathcal{B}_i$. The set of children of any bin $B \in \mathcal{B}_i$ with $i \geq 0$ is defined as $\mathsf{child}(B) := \{B' \in \mathcal{B}_{i+1} : B' \subset B\}$. Then at each time t, the partition $\mathcal{L}_t$ induced by Algorithm 1 can be described as a set of leaf nodes of a subtree of $\mathcal{T}^{(l)}$ for some $l > 0$. Throughout the paper, the terms bin and node are used interchangeably.

Next, given a subset $\mathcal{D}$ of the source data set $\mathcal{D}^P$, for any bin B and arm k, let $n_k^P(B; \mathcal{D})$ denote the number of samples in data set $\mathcal{D}$ such that the covariate falls in bin B and arm k

Algorithm 1 Transfer learning algorithm for contextual multi-armed bandits

Input: arm set $\mathcal{I}$, horizon length n_Q, smoothness parameters β, C_β, transfer parameters γ, exploration coefficient κ, P-data $\mathcal{D}^P$.
Initialize $\mathcal{L}_1 \leftarrow \{\mathcal{X}\}$, $\mathcal{I}(\mathcal{X}) \leftarrow \mathcal{I}$. ▷ initialize partition and arm set
Initialize the policy $\widetilde{\pi}(\mathcal{X})$ by Procedure 1$(\mathcal{X}, \mathcal{I}(\mathcal{X}), U, \mathcal{D}^P)$.
Initialize $N(\mathcal{X}) \leftarrow 0$. ▷ initialize time for policy $\widetilde{\pi}(\mathcal{X})$
Initialize $\tau_k(\mathcal{X}) \leftarrow 0$, and $\tau_k^\star(\mathcal{X}; \mathcal{D}^P)$ as in (10), $\forall k \in \mathcal{I}(\mathcal{X})$. ▷ initialize rounds and set round upper bounds
for $t = 1, \ldots, n_Q$ **do**
  Draw a sample $X_t^Q \sim Q_X$.
  Find the bin $B \in \mathcal{L}_t$ such that $X_t^Q \in B$.
  while $|\mathcal{I}(B)| > 1$ and $\tau_k(B) \geq \tau_k^\star(B; \mathcal{D}^P)$, $\forall k \in \mathcal{I}(B)$ **do** ▷ keep partitioning B until reaching suitable scale
    if $\tau_k^\star(B; \mathcal{D}^P) = 0$, $\forall k \in \mathcal{I}(B)$ **then** ▷ no exploration needed in B: discard suboptimal arms
      Set $\underline{Y}^\star(B; \mathcal{D}^P) \leftarrow \max_{k \in \mathcal{I}(B)} \{\overline{Y}_k^P(B; \mathcal{D}^P) - U_k(0, B; \mathcal{D}^P)\}$. ▷ set largest reward lower bound
      Set $\mathcal{I}(B) \leftarrow \{k \in \mathcal{I}(B) : \overline{Y}_k^P(B; \mathcal{D}^P) + U_k(0, B; \mathcal{D}^P) \geq \underline{Y}^\star(B; \mathcal{D}^P)\}$. ▷ update arm set
    end if
    for $B' \in \mathsf{child}(B)$ **do**
      Set $\mathcal{I}(B') \leftarrow \mathcal{I}(B)$. ▷ assign remaining arms in B as initial arms in its children
      Initialize the policy $\widetilde{\pi}(B')$ by Procedure 1 $(B', \mathcal{I}(B'), U, \mathcal{D}^P)$.
      Set $N(B') \leftarrow 0$. ▷ initialize time for policy $\widetilde{\pi}(B')$
      Set $\tau_k(B') \leftarrow 0$ and $\tau_k^\star(B'; \mathcal{D}^P)$ as in (10), $\forall k \in \mathcal{I}(B')$. ▷ initialize rounds and set round upper bounds
    end for
    Set $\mathcal{L}_t \leftarrow (\mathcal{L}_t \setminus B) \cup \mathsf{child}(B)$. ▷ replace B with its children in partition
    Find the bin $B \in \mathcal{L}_t$ such that $X_t^Q \in B$.
  end while
  Set $N(B) \leftarrow N(B) + 1$. ▷ update times $X_t^Q \in B$
  Set $\pi_t \leftarrow \widetilde{\pi}_{N(B)}(B)$. ▷ choose arm by policy $\widetilde{\pi}(B)$
  Set $\mathcal{I}(B) \leftarrow \widetilde{\mathcal{I}}_{N(B)}(B)$. ▷ update arm set by policy $\widetilde{\pi}(B)$
  Set $\tau_k(B) \leftarrow \widetilde{\tau}_{N(B),k}(B)$, $\forall k \in \mathcal{I}(B)$. ▷ update numbers of rounds by policy $\widetilde{\pi}(B)$
end for
Output: policy $\{\pi_t\}_{t \geq 1}$.

is pulled, that is,

$$n_k^P(B; \mathcal{D}) := \sum_{(X_i, \pi_i, Y_i) \in \mathcal{D}} \mathbb{1}\{X_i \in B,\, \pi_i = k\}. \tag{7}$$

Let $\overline{Y}_k^P(B; \mathcal{D})$ denote the empirical mean of the reward of arm k over bin B in data set $\mathcal{D}$, namely

$$\overline{Y}_k^P(B; \mathcal{D}) := \begin{cases} \dfrac{1}{n_k^P(B; \mathcal{D})} \displaystyle\sum_{(X_i, \pi_i, Y_i) \in \mathcal{D}} Y \mathbb{1}\{X_i \in B,\, \pi_i = k\} & \text{if } n_k^P(B; \mathcal{D}) \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In addition, for any nonnegative integer $\tau \geq 0$, bin B and arm k, let us define

$$U_k(\tau, B; \mathcal{D}) := \begin{cases} 2\sqrt{\dfrac{2}{\tau + n_k^P(B;\mathcal{D})} \log^+\left(\dfrac{n_Q|B|^d}{\tau}\right)} \vee 2C_\beta|B|^\beta & \text{if } \tau > 0, \\ 2\sqrt{\dfrac{2}{n_k^P(B;\mathcal{D})} \log^+\left(n_Q|B|^{d+2\beta} \vee \kappa n_P|B|^{d+2\beta+\gamma}\right)} \vee 2C_\beta|B|^\beta & \text{if } \tau = 0, \end{cases} \tag{9}$$

where we recall the notation $\log^+(x) := \log(x) \vee 1$ and use the convention $1/0 = \infty$. The $U_k(\tau, B; \mathcal{D})$ can be essentially viewed as a confidence bound that quantifies the uncertainty of the reward function estimator used in Algorithm 1.

With these notation in place, the transfer learning algorithm for nonparametric contextual multi-armed bandits is summarized in Algorithm 1. Let us discuss its details, along with some intuition. As briefly mentioned earlier, Algorithm 1 aims to segment the covariate space based on the local margins of the reward functions. Smaller bins are employed in areas where the gaps between the reward functions of different arms are small, whereas coarser partitioning is used in regions where arms are easily distinguishable. Once this partition is established, each bin is treated as an index for a sequence of static multi-armed bandit problems, and Procedure 1 is executed in each bin with parameters specific to that bin.

In order to achieve such an adaptive partition, for each bin B and arm k, we assign a nonnegative upper bound on the number of pulls, given by

$$\tau_k^\star(B; \mathcal{D}) := \min_{\tau \in \{0\} \cup \mathbb{N}} \{\tau : U_k(\tau, B; \mathcal{D}) \leq 2C_\beta|B|^\beta\}, \tag{10}$$

where we recall the definition of $U_k(\tau, B; \mathcal{D})$ in (9). Note that the confidence bound $U_k(\tau, B; \mathcal{D})$ is composed of two terms. The first term represents the standard deviation of the reward function estimator owing to finite samples, while the second component stands for the bias term, as we attempt to approximate the reward function using its conditional expectation over bin B. Roughly speaking, the value of $\tau_k^\star(B; \mathcal{D})$ is chosen to ensure that, after arm k has been pulled $\tau_k^\star(B; \mathcal{D})$ times in bin B, the standard deviation and bias of its reward function estimator is balanced. In particular, if the conditional mean reward of arm k over bin B is low, Procedure 1 executed in bin B can identify and eliminate it by the end of $\tau_k^\star(B; \mathcal{D})$ rounds with high probability. Combined with the smoothness assumption, this procedure guarantees that the eliminated arms are uniformly suboptimal over bin B and that none of the remaining arms dominates the others. Therefore, if multiple arms remain active in bin B after each arm k has been pulled $\tau_k^\star(B; \mathcal{D})$ times, one knows that the reward functions of the remaining arms are locally close to each other, and hence need more refined estimation. Consequently, we split the node B by replacing B with its children $\mathsf{child}(B)$ in the partition tree, and the set of the active arms in node B is passed on to each $B' \in \mathsf{child}(B)$. An illustration of Algorithm 1 can be found in Figure 1.

Next, let us move on to take a closer look at Procedure 1. Since it is designed for each bin, in addition to the bin index B, set of arms $\mathcal{I}$ and confidence bound function U, Procedure 1 also requires the information of the source samples that fall in the bin. Specifically, Procedure 1 needs the sample size n_k^P, empirical mean of the reward $\overline{Y}_k^P$ and upper bound on the play rounds $\tau_k^\star$, for each arm $k \in \mathcal{I}$.

Before any arm reaches its play round limit, Procedure 1 runs similar to a standard successive elimination algorithm (see, e.g., Auer and Ortner (2010), Perchet and Rigollet (2013)). It operates in rounds and maintains a set of active arms that are potentially optimal. In each round, each arm in the active arm set is pulled once. Given access to the source data set, the observed rewards from the Q-bandit and P-bandit are combined to calculate the empirical

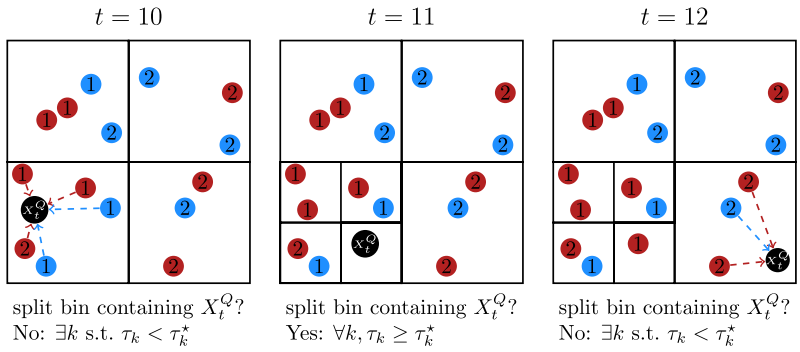

FIG. 1. *An illustration of Algorithm 1 for $d = 2$ and $K = 2$. The target samples (resp., source samples) are represented by the red (resp., blue) points. The coordinates of each point correspond to the covariate X, and the number in the point stands for the arm that is pulled. In each time step t, Algorithm 1 first assesses if the bin containing X_t^Q requires splitting. It then utilizes the samples located in the same bin as X_t^Q to execute a static MAB procedure to select an arm. For example, at time $t = 11$, one has (say) $\tau_1^\star = 3$, $\tau_2^\star = 1$, and both arms are active. In this case, we need to split the lower left bin and run Procedure 1 in the bin containing X_t^Q to choose an arm.*

average of the reward of each active arm. It then seeks to eliminate suboptimal arms from the set of active arms by comparing their upper and lower confidence bounds of the sample mean rewards.

However, once arm k has been pulled for $\tau_k^\star$ times, Procedure 1 stops selecting it temporarily. In fact, $\tau_k^\star$ can be viewed as the maximum horizon length of the exploration phase for arm k. The rationale is simple: given the additional information from the source data, one should be able to gain more certainty about whether arm k is optimal compared to the standard case. Consequently, we can reduce the cumulative regret by exiting from the exploration stage earlier. Once each active arm reaching its play round limit, Procedure 1 advances to the second phase, where we select only the arm with the highest empirical average reward. In this case, a previously suspended arm might be pulled again if it happens to be the only active arm or has the highest empirical mean award. Therefore, one critical difference between this work and previous results on classical contextual multi-armed bandits (without source data), such as Perchet and Rigollet (2013), is that our successive elimination procedure in each bin involves a more complicated early stopping stage, hence requiring much a more sophisticated analysis. In addition, achieving the exact minimax regret demands more carefully designed uncertainty estimates that take account of both the source sample size and transfer learning parameters.

It is noteworthy that Algorithm 1 has three critical steps to integrate the source data. First, we combine the target and source data to estimate the reward function of each arm, and the increased sample size leads to an improvement in estimation accuracy. Second, note that the upper bound on the play rounds $\tau_k^\star(B; \mathcal{D})$ in (10) incorporates the information from the source data $\mathcal{D}$. It is easy to see that $\tau_k^\star(B; \mathcal{D})$ is a decreasing function of the source sample size $n_k^P(B; \mathcal{D})$. Therefore, if a suboptimal arm has been pulled sufficiently many times in the P-bandit so that one is certain about its suboptimality, we no longer need to test it in the Q-bandit. This implies a shorter exploration phase, and thus reduces the regret. Finally, closely related to the second point, the source data allows us to build a deeper partition tree. We can then discretize the covariate space more finely, thus facilitating the identification of suboptimal arms and incurring a smaller regret.

We would like to remark that the covariate shift of the source data is characterized by the transfer exponent γ, which plays a crucial role in Algorithm 1. For instance, as γ decreases, the partition tree becomes deeper, resulting in a more refined partition of the covariate space.

Procedure 1 Successive elimination procedure for a static bandit with source data

Input: bin B, arm set $\mathcal{I}$, confidence bound function U, source data $\mathcal{D}$.
Set $n_k^P(B;\mathcal{D})$, $\overline{Y}_k^P(B;\mathcal{D})$, $\tau_k^\star(B;\mathcal{D})$ as in (7), (8), (10), respectively, $\forall k \in \mathcal{I}$.
Set $n_k^P \leftarrow n_k^P(B;\mathcal{D})$, $\overline{Y}_k \leftarrow \overline{Y}_k^P(B;\mathcal{D})$, and $\tau_k^\star \leftarrow \tau_k^\star(B;\mathcal{D})$, $\forall k \in \mathcal{I}$.
Initialize $t \leftarrow 0$.
Initialize $\tau_k \leftarrow 0, \forall k \in \mathcal{I}$. ▷ initialize pull counts
Initialize $\underline{Y}^\star \leftarrow \max_{k\in\mathcal{I}}\{\overline{Y}_k - U_k(0, B;\mathcal{D})\}$. ▷ initialize largest reward lower bound
loop
  if $\tau_k \geq \tau_k^\star, \forall k \in \mathcal{I}$ **then** ▷ condition to stop exploration
    Set $t \leftarrow t+1$.
    Select arm $\widetilde{\pi}_t \leftarrow \underset{k\in\mathcal{I}}{\operatorname{argmax}} \overline{Y}_k$ (with ties broken arbitrarily) and receive reward $Y^{Q,(\widetilde{\pi}_t)}$.
    Set $\tau_{\widetilde{\pi}_t} \leftarrow \tau_{\widetilde{\pi}_t} + 1$. ▷ update pull count
    Set $\tau_{t,k} \leftarrow \tau_k, \forall k \in \mathcal{I}$, and $\mathcal{I}_t \leftarrow \mathcal{I}$. ▷ record pull count and active arm set
    Set $\overline{Y}_{\widetilde{\pi}_t} \leftarrow \frac{1}{n^P_{\widetilde{\pi}_t}+\tau_{\widetilde{\pi}_t}}(Y^{Q,(\widetilde{\pi}_t)} + (n^P_{\widetilde{\pi}_t} + \tau_{\widetilde{\pi}_t} - 1)\overline{Y}_{\widetilde{\pi}_t})$. ▷ update estimated reward
  else
    for $k \in \mathcal{I}$ such that $\tau_k < \tau_k^\star$ **do** ▷ only explore arms s.t. pull counts less than upper bounds
      if $\overline{Y}_k + U_k(\tau_k, B;\mathcal{D}) \geq \underline{Y}^\star$ **then** ▷ eliminate arm s.t. reward upper bound is smaller than largest reward lower bound
        Set $t \leftarrow t+1$.
        Select arm $\widetilde{\pi}_t \leftarrow k$ and receive reward $Y^{Q,(\widetilde{\pi}_t)}$.
        Set $\tau_{\widetilde{\pi}_t} \leftarrow \tau_{\widetilde{\pi}_t} + 1$. ▷ update pull count
        Set $\tau_{t,k} \leftarrow \tau_k, \forall k \in \mathcal{I}$, and $\mathcal{I}_t \leftarrow \mathcal{I}$. ▷ record pull count and active arm set
        Set $\overline{Y}_{\widetilde{\pi}_t} \leftarrow \frac{1}{n^P_{\widetilde{\pi}_t}+\tau_{\widetilde{\pi}_t}}(Y^{Q,(\widetilde{\pi}_t)} + (n^P_{\widetilde{\pi}_t} + \tau_{\widetilde{\pi}_t} - 1)\overline{Y}_{\widetilde{\pi}_t})$. ▷ update estimated reward
        Set $\underline{Y}^\star \leftarrow \max_{k\in\mathcal{I}}\{\overline{Y}_k - U_k(\tau_k, B;\mathcal{D})\}$. ▷ update largest reward lower bound
      else
        Eliminate arm k from active arm set: $\mathcal{I} \leftarrow \mathcal{I} \setminus \{k\}$.
      end if
    end for
  end if
end loop
Output: policy $\{\widetilde{\pi}_t\}_{t\geq 1}$, arm pull counts $\{(\tau_{t,k})_{k\in\mathcal{I}_t}\}_{t\geq 1}$, sets of active arms $\{\mathcal{I}_t\}_{t\geq 1}$.

Therefore, when the covariate shift is mild, Algorithm 1 can construct more accurate local estimators for the reward functions. This aids in distinguishing the optimal arm, thereby reducing the cumulative regret. Additionally, the confidence bound $U_k(\tau, B;\mathcal{D})$ in (9) for arm k in bin B depends on the number of source samples $n_k^P(B;\mathcal{D})$ such that the covariate falls within bin B and arm k is pulled. This number depends on the source marginal distribution P_X and, consequently, the transfer exponent γ. As γ approaches zero, $n_k^P(B;\mathcal{D})$ tends to increase with high probability, leading to a tighter confidence bound $U_k(\tau, B;\mathcal{D})$. According to the elimination criterion in Algorithm 1 and Procedure 1, this enhances the accuracy of distinguishing the optimal arm, resulting in a reduction in the cumulative regret. Moreover, Algorithm 1 guarantees that in each bin B, each arm k is played for a maximum of $\tau_k^\star(B;\mathcal{D})$ times. As the upper bound on the number of pulls $\tau_k^\star(B;\mathcal{D})$ in (10) depends on the confidence

bound $U_k(\tau, B; \mathcal{D})$, it is also influenced by the transfer exponent γ. As γ goes to zero, the upper bound on the number of pulls $\tau_k^\star(B; \mathcal{D})$ decreases. Therefore, when the covariate shift is slight, Algorithm 1 selects suboptimal arms less frequently, leading to a decrease in the cumulative regret.

Finally, we note that Algorithm 1 takes the horizon length n_Q as input that may be unknown in practice. Fortunately, this issue can be circumvented by the well-known doubling trick (see Auer et al. (1995)).

3.2. *Minimax rate of convergence.* We proceed to discuss the theoretical guarantees of Algorithm 1. To begin with, Theorem 1 gives an upper bound on the regret of the Q-bandit. The proof is postponed to Appendix B in the Supplementary Material (Cai, Cai and Li (2024)).

THEOREM 1 (Upper bound). *Assume that $\alpha\beta \le d$. Then the expected regret of the policy π given by Algorithm 1 satisfies*

$$\sup_{\Pi(K,\beta,\alpha,\gamma,\kappa)} \mathbb{E}\big[R_{n_Q}(\pi)\big] \le C n_Q \big(n_Q + (\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}}\big)^{-\frac{\beta(1+\alpha)}{d+2\beta}}, \tag{11}$$

where $C > 0$ is a constant independent of n_Q and n_P.

In addition, Theorem 2 below shows that the regret of Algorithm 1 matches the minimax lower bound, thereby justifying Algorithm 1 is rate-optimal. The proof can be found in Appendix C in the Supplementary Material (Cai, Cai and Li (2024)).

THEOREM 2 (Lower bound). *Assume that $\alpha\beta \le d$. Then one has*

$$\inf_{\pi} \sup_{\Pi(K,\beta,\alpha,\gamma,\kappa)} \mathbb{E}\big[R_{n_Q}(\pi)\big] \ge c n_Q \big(n_Q + (\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}}\big)^{-\frac{\beta(1+\alpha)}{d+2\beta}}, \tag{12}$$

where $c > 0$ is a constant independent of n_Q and n_P.

Here, the infimum is taken over the class of admissible policies obeying the selected arm π_t at time t depends only on observations prior to time t, that is, $\{(X_s^Q, \pi_s, Y^{Q,(\pi_s)})\}_{s<t} \cup \{X_t^Q\}$.

We now discuss several important implications.

- Theorems 1 and 2 together establish the minimax regret for transfer learning under the covariate shift model when the number of arms $K \asymp 1$:

$$\inf_{\pi} \sup_{\Pi(K,\beta,\alpha,\gamma,\kappa)} \mathbb{E}\big[R_{n_Q}(\pi)\big] \asymp n_Q \big(n_Q + (\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}}\big)^{-\frac{\beta(1+\alpha)}{d+2\beta}}. \tag{13}$$

- In the classical setting where one has no access to the source P-data, the minimax regret is established in Perchet and Rigollet (2013), given by

$$\inf_{\pi} \sup_{\Pi(K,\beta,\alpha)} \mathbb{E}\big[R_{n_Q}(\pi)\big] \asymp n_Q^{1-\frac{\beta(1+\alpha)}{d+2\beta}}. \tag{14}$$

  We can see that (14) is a special case of (13) by setting $n_P = 0$.
- Comparing the minimax regret (13) in the transfer learning setting with the minimax regret (14) in the standard setting, it becomes evident that the incorporation of source data leads to a faster convergence rate of the regret. The reduced regret quantifies the information allowed to be transferred from the source data to the target bandit, with a precise characterization of the dependence on the smoothness β, dimension d, transfer exponent γ and exploration coefficient κ.

- Due to the difference between the source and target distributions, it is reasonable to anticipate that the value of the source data differs from that of the target data. This intuition is elucidated in (13) where we can interpret $(\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}}$ as the effective sample size provided by the source data. Moreover, given that $\kappa \in [0, 1]$ and $\frac{d+2\beta}{d+2\beta+\gamma} \leq 1$ for $\gamma \geq 0$, the minimax regret (13) further suggests that the samples from the source data set are always inferior unless $P = Q$ in the context of nonparametric contextual multi-armed bandits.
- We proceed to examine the roles of the parameters introduced for transfer learning. As we intuitively discussed in Section 2, the challenge of transfer learning becomes more formidable as the transfer exponent γ increases. This observation is validated theoretically in (13), demonstrating that a smaller value of γ results in a faster convergence rate of the regret.
- Shifting our focus to the exploration coefficient κ, we can see that the dependence of the minimax rate (13) on κ underscores the importance of extensive exploration in the source data in order to achieve reliable transfer learning. When $\kappa = 0$, the effective source sample size is zero. To see this, let us consider a three-armed bandit problem where the policy in the source domain exclusively selects arm 3. In such a case, even if the source sample size n_P were to go to infinity, which provides us with precise knowledge of the reward function $f_3(x)$ of arm 3, we would still be confronted with a two-armed bandit problem, with minimax rate of convergence of the regret given by $n_Q^{1-\frac{\beta(1+\alpha)}{d+2\beta}}$. This indicates that a substantial reduction in regret cannot be expected unless the source data set is highly exploratory.
- We would like to remark that the minimax lower bound (12) in Theorem 2 is the same as the one established for the classification setting in Kpotufe and Martinet (2021). This means that even if we can observe the rewards generated by all the arms in each round, it remains impossible to design a policy that can improve the regret upper bound (12) in terms of n_P and n_Q. Intuitively, this implies that although we are faced with the challenge of sequential decision-making, the hardness of nonparametric estimation ultimately dictates the complexity of transfer learning for nonparametric contextual multi-armed bandits.

We conclude this section by comparing our work with some intimately connected prior research in the literature on transfer learning. As mentioned earlier, the transfer exponent used in the current paper was originally proposed in Kpotufe and Martinet (2021). Along with its variants (see, e.g., Cai and Wei (2021), Pathak, Ma and Wainwright (2022)), they have been broadly deployed in transfer learning for various supervised learning problems, where the primary focus is on leveraging the source data set to develop optimal estimators for the regression functions of interest. In contrast, the sequential decision-making nature of bandit problems introduces new algorithmic and technical challenges in data integration compared to these prior works. For example, in order to address the trade-off between exploration and exploitation, it is essential not only to construct suitable reward function estimators but also to utilize the source data intelligently to develop confidence intervals for their uncertainties. Moreover, the samples for each arm are collected adaptively in the bandit setting. This results in more complicated distributions of samples and reward function estimators, necessitating more careful statistical analysis.

4. Adaptivity. The primary drawback of Algorithm 1 is its reliance on prior knowledge of the smoothness parameter β and transfer parameters γ, κ, which are often unknown in practical applications. Therefore, it is natural to ask if one can develop a data-driven algorithm that achieves the minimax optimal rate of convergence while adapting to a wide range of parameter spaces $\Pi(K, \beta, \alpha, \gamma, \kappa)$. However, it is widely acknowledged in the bandit literature that one cannot hope to develop a smoothness-adaptive algorithm that attains the

minimax regret simultaneously over different classes of multi-armed bandits with varying smoothness (Locatelli and Carpentier (2018), Gur, Momeni and Wager (2022), Cai and Pu (2022b)). Additional structural assumptions on the reward functions are needed to achieve smoothness adaptivity.

Given the general impossibility of adaptation, we focus on a setting where the reward functions satisfy the self-similarity condition. This condition has been used for adaptive confidence intervals in the nonparametric regression literature (Picard and Tribouley (2000), Giné and Nickl (2010)) and the adaptive multi-armed bandit literature (Gur, Momeni and Wager (2022), Cai and Pu (2022b)). In the remainder of this section, we first introduce the self-similarity condition in Section 4.1 and subsequently present a data-driven algorithm with theoretical guarantees in Section 4.2. As a side note, we believe it is also possible to develop adaptive strategies for the bandits with shape-constrained reward functions (e.g., concavity), and we defer the discussion to Section 5.

4.1. *The self-similarity condition.* For any function $f(\cdot)$ on $\mathcal{X}$, bin B in $\mathcal{X}$, and probability distribution λ over $\mathcal{X}$, let $\Gamma_B f(\cdot;\lambda)$ be the $L_2(\lambda)$-projection of f onto the class of piecewise-constant functions over B, namely

$$\Gamma_B f(x;\lambda) := \frac{1}{\lambda(B)} \int_B f(u)\,\mathrm{d}\lambda(u) \tag{15}$$

if $\lambda(B) > 0$, and $\Gamma_B f(x;\lambda) := 0$ otherwise.

Recall that $\mathcal{B}_l$ denotes the partition of $\mathcal{X}$ that consists of bins with side length 2^{-l}. We now present the self-similarity condition as follows.

DEFINITION 3 (Self-similarity). Let $f : \mathcal{X} \to \mathbb{R}$ be a Hölder continuous function in $\mathcal{H}(\beta, C_\beta)$ with $0 < \beta \le 1$. For any probability distribution λ and constants $l_0 \ge 0$, $b > 0$, we say f is self-similar under λ with parameters l_0 and b, if the following holds for any integer $l \ge l_0$:

$$\sup_{B\in\mathcal{B}_l} \sup_{x\in B} \left|\Gamma_B f(x;\lambda) - f(x)\right| \ge b2^{-\beta l}.$$

In a nutshell, the self-similarity condition imposes a global lower bound on the approximation error of a function using piecewise-constant functions. Therefore, it can be viewed as a complement to the Hölder smoothness condition, which implies an upper bound on the error. We note that the difficulty of smoothness adaptation arises from the possibility of functions being highly irregular on small scales, a scenario precluded by the self-similarity condition. Interested readers are referred to Bull (2012), Nickl and Szabó (2016) for more in-depth discussions on the self-similarity conditions. Below, we present an example of the self-similar functions.

EXAMPLE. Fix some constants $a, c \in \mathbb{R}$. The function $f(x) = ax^\beta + c$ is self-similar under the uniform distribution over $[0, 1]$ with parameters $l_0 = 0$ and $b = \frac{a}{\beta+1}$.

To verify this, let us fix an arbitrary $l \ge 0$ and denote $B_0 = [0, 2^{-l}]$. It is straightforward to compute

$$\begin{aligned}\sup_{B\in\mathcal{B}_l} \sup_{x\in B} \left|\Gamma_B f\big(x; \mathsf{Unif}[0,1]\big) - f(x)\right| &\ge \left|\Gamma_{B_0} f\big(0; \mathsf{Unif}[0,1]\big) - f(0)\right| \\ &= \left|\int_{B_0} f(x)2^l\,\mathrm{d}x - c\right| = a\int_0^{2^{-l}} x^\beta 2^l\,\mathrm{d}x = \frac{a}{\beta+1}2^{-\beta l}.\end{aligned}$$

As a consequence, this shows that $f(x) = ax^\beta + c$ is self-similar by Definition 3.

We now introduce the self-similarity assumption regarding the reward functions $\{f_k\}_{k=1}^K$ as follows.

ASSUMPTION 4 (Self-similarity). Assume that there exists some $k \in [K]$ such that the reward function f_k is self-similar under both Q_X and $P_{X\,|\,\pi=k}$ with some parameters $l_0 \geq 0$ and $b > 0$.

We denote by $\Pi(K, \beta, C_\beta, \alpha, C_\alpha, \underline{q}, \overline{q}, \gamma, c_\gamma, \kappa, l_0, b)$ the class of nonparametric contextual K-armed bandits that satisfy Assumptions 1–4 and Definitions 1–2. Whenever clear from the context, the shorthand $\Pi(K, \beta, \alpha, \gamma, \kappa, l_0, b)$ and Π are used.

4.2. *Adaptive algorithm.* In this section, we present a two-stage adaptive transfer learning algorithm designed for contextual multi-armed bandits with self-similar reward functions. In essence, this adaptive algorithm begins by computing a reasonably precise estimate $\widehat{\beta}$ for the Hölder smoothness parameter β (see Procedure 2). It then uses $\widehat{\beta}$ as an input to a variant of Algorithm 1, which guarantees a near-optimal statistical performance if given an accurate smoothness estimate. A comprehensive description of this algorithm is summarized in Algorithm 2.

In the first phase, our primary objective is to estimate the Hölder smoothness parameter β under the self-similarity condition. The procedure is rooted in the critical insight that the local piecewise-constant regression estimator of a function f over a bin B is sufficiently close to its piecewise-constant projection $\Gamma_B f$ with high probability. Hence, when applying the local piecewise-constant regression method to estimate the reward function, the self-similarity condition ensures that the estimation bias does not decay too rapidly. Combined with the Hölder smooth condition, this yields a tight bound on the estimation bias, which depends on the smoothness parameter β. Even though we lack direct access to the estimation bias, we can adapt Lepski's method (Lepskii (1991, 1992, 1993), Lepski, Mammen and Spokoiny (1997)) suitably to obtain a reliable estimate by comparing the difference between estimators with different bin side lengths. To this end, Procedure 2 first creates two partitions of the covariate space by using bins of different sizes. Next, based on these two partitions, it collects independent samples $\mathcal{D}_{\mathsf{se}}$ to construct two separate local piecewise-constant regression estimators for each reward function in each bin. Clearly, the estimation bias will be larger in the bin with a larger side length. As long as we collect adequate samples such that the estimation bias dominates the standard deviation, the maximal difference between the two regression estimates is approximately of the same order as the larger estimation bias. This allows us to infer the smoothness of the reward functions. In particular, with high probability, we can obtain a smoothness estimate $\widehat{\beta}$ with statistical guarantee $\beta - O(\log_2(\log(n))/\log_2(n)) \leq \widehat{\beta} \leq \beta$, which suffices to attain a near-optimal regret. The detailed smoothness estimation procedure is presented in Procedure 2. It is worth noting that depending on the relationship between n_P and n_Q, the samples for the smoothness estimation are collected from either the Q-bandit or P-bandit. In the former case, the smoothness estimation phase only takes a vanishingly small portion of the time horizon length n_Q in the Q-bandit. Thus, the regret incurred during this stage is negligible compared to the minimax regret. On the other hand, when $n_Q < n_P$, we split the source samples for the smoothness estimation and for the decision-making in the target bandit separately. Similarly, the sample size used for the smoothness estimation is considerably smaller than the total source sample size n_P and has no impact on the minimax regret.

With the smoothness estimate $\widehat{\beta}$ in hand, the second stage of Algorithm 2 takes it as an input and operates in a manner similar to Algorithm 1. We would like to highlight several pivotal differences. To begin with, as a portion of the source data may have been used in the

Algorithm 2 Adaptive transfer learning algorithm for contextual multi-armed bandits

Input: arm set $\mathcal{I}$, horizon length n_Q, lower and upper bounds on smoothness $\underline{\beta}$ and $\overline{\beta}$, upper bound on Lipschitz constant $\overline{C}_\beta$, upper bound on transfer exponent $\overline{\gamma}$, P-data $\mathcal{D}^P$.
Run Procedure 2 (K, n_Q, $\underline{\beta}$, $\overline{\beta}$, $\overline{\gamma}$, $\mathcal{D}^P$) to get the smoothness estimate $\widehat{\beta}$, time steps s_P and s_Q, and policy $\{\pi_t^{\mathsf{a}}\}_{1\le t<s_Q}$.
Split the source data $\mathcal{D}^P_{\mathsf{dm}} \leftarrow \{(X_i^P, \pi_i^P, Y^{P,(\pi_i^P)})\}_{i=s_P}^{n_P}$ to aid in decision-making in the Q-bandit.
Set $\mathcal{L}_{s_Q} \leftarrow \{\mathcal{X}\}$, and $\mathcal{I}(\mathcal{X}) \leftarrow \mathcal{I}$. ▷ initialize partition and arm set
Initialize the policy $\widetilde{\pi}(\mathcal{X})$ by Procedure 1 ($\mathcal{X}, \mathcal{I}(\mathcal{X}), \widehat{U}, \mathcal{D}^P_{\mathsf{dm}}$).
Initialize $N(\mathcal{X}) \leftarrow 0$. ▷ set time to 0 for policy $\widetilde{\pi}(\mathcal{X})$
Initialize $\tau_k(\mathcal{X}) \leftarrow 0$, and $\widehat{\tau}_k^\star(\mathcal{X}; \mathcal{D}^P_{\mathsf{dm}})$ as in (18), $\forall k \in \mathcal{I}(\mathcal{X})$. ▷ initialize rounds and round upper bounds
for $t = s_Q, \ldots, n_Q$ **do**
  Draw a sample $X_t^Q \sim Q_X$.
  Find the bin $B \in \mathcal{L}_t$ such that $X_t^Q \in B$.
  while $|\mathcal{I}(B)| > 1$ and $\tau_k(B) \ge \widehat{\tau}_k^\star(B; \mathcal{D}^P_{\mathsf{dm}})$, $\forall k \in \mathcal{I}(B)$ **do** ▷ keep partitioning B until reaching suitable scale
    if $\widehat{\tau}_k^\star(B; \mathcal{D}^P_{\mathsf{dm}}) = 0$, $\forall k \in \mathcal{I}(B)$ **then** ▷ no exploration needed in B: discard suboptimal arms
      Set $\underline{Y}^\star(B; \mathcal{D}^P_{\mathsf{dm}}) \leftarrow \max_{k\in\mathcal{I}(B)}\{\overline{Y}_k^P(B; \mathcal{D}^P_{\mathsf{dm}}) - \widehat{U}_k(0, B; \mathcal{D}^P_{\mathsf{dm}})\}$ ▷ set largest reward lower bound
      Set $\mathcal{I}(B) \leftarrow \{k \in \mathcal{I}(B) : \overline{Y}_k^P(B; \mathcal{D}^P_{\mathsf{dm}}) + \widehat{U}_k(0, B; \mathcal{D}^P_{\mathsf{dm}}) \ge \underline{Y}^\star(B; \mathcal{D}^P_{\mathsf{dm}})\}$. ▷ update arm set
    end if
    for $B' \in \mathsf{child}(B)$ **do**
      Set $\mathcal{I}(B') \leftarrow \mathcal{I}(B)$. ▷ assign remaining arms in B as initial arms in its children
      Initialize the policy $\widetilde{\pi}(B')$ by Procedure 1 ($B', \mathcal{I}(B'), \widehat{U}, \mathcal{D}^P_{\mathsf{dm}}$).
      Set $N(B') \leftarrow 0$. ▷ initialize time for policy $\widetilde{\pi}(B')$
      Set $\tau_k(B') \leftarrow 0$, and $\widehat{\tau}_k^\star(B'; \mathcal{D}^P_{\mathsf{dm}})$ as in (18), $\forall k \in \mathcal{I}(B')$. ▷ initialize rounds and round upper bounds
    end for
    Set $\mathcal{L}_t \leftarrow (\mathcal{L}_t \setminus B) \cup \mathsf{child}(B)$. ▷ replace B with its children in partition
    Find the bin $B \in \mathcal{L}_t$ such that $X_t^Q \in B$.
  end while
  Set $N(B) \leftarrow N(B) + 1$. ▷ update times $X_t^Q \in B$
  Set $\pi_t^{\mathsf{a}} \leftarrow \widetilde{\pi}_{N(B)}(B)$. ▷ choose arm by policy $\widetilde{\pi}(B)$
  Set $\mathcal{I}(B) \leftarrow \widetilde{\mathcal{I}}_{N(B)}(B)$. ▷ update arm set by policy $\widetilde{\pi}(B)$
  Set $\tau_k(B) \leftarrow \widetilde{\tau}_{N(B),k}(B)$, $\forall k \in \mathcal{I}(B)$. ▷ update numbers of rounds by policy $\widetilde{\pi}(B)$
end for
Output: policy $\pi^{\mathsf{a}} = \{\pi_t^{\mathsf{a}}\}_{t\ge1}$.

smoothness estimation process, Algorithm 2 utilizes a subset of the source data set $\mathcal{D}^P_{\mathsf{dm}} \subset \mathcal{D}^P$ to assist in distinguishing the optimal arm in the target bandit. Additionally, note that the confidence bound $U_k(\tau, B; \mathcal{D})$ (cf. (9)) used in Algorithm 1 requires the knowledge of the unknown parameters γ and κ. To construct an adaptive procedure in Algorithm 2, we substitute it with the confidence bound $\widehat{U}_k(\tau, B; \mathcal{D})$ defined as follows. Given a subset $\mathcal{D}$ of the source data $\mathcal{D}^P$ and the smoothness estimate $\widehat{\beta}$ returned by Procedure 2, for any nonnegative

Procedure 2 Smoothness estimation procedure

Input: arm number K, horizon length n_Q, lower and upper bounds on smoothness $\underline{\beta}$ and $\overline{\beta}$, upper bound on transfer exponent $\overline{\gamma}$, P-data $\mathcal{D}^P$, tuning parameters C_1, C_2.
Set $n \leftarrow n_P \vee n_Q$.
if $n_P > n_Q$ **then**
  Set $l_1 \leftarrow \lceil \frac{\underline{\beta}\log_2(n)}{(d+2\overline{\beta}+\overline{\gamma})^2} \rceil$, $l_2 \leftarrow l_1 + \lceil \frac{1}{d}\log_2(\log(n)) \rceil$, $l_3 \leftarrow \lceil \frac{\overline{\beta}}{\underline{\beta}} l_1 + \frac{1}{\underline{\beta}}\log_2(\log(n)) \rceil$. ▷ bandwidths of reward function estimators
  Set $T \leftarrow \lceil C_1 K n^{\frac{\underline{\beta}}{d+2\overline{\beta}}} \log^{\frac{d+\overline{\gamma}}{d}}(Kn) \rceil$. ▷ sample size for smoothness estimation
  Set $\mathcal{D}_{\mathsf{se}} \leftarrow \{(X_i^P, \pi_i^P, Y_i^{P,(\pi_i^P)})\}_{i=1}^T$, $s_P \leftarrow T+1$, $s_Q \leftarrow 1$. ▷ samples for smoothness estimation
else
  Set $l_1 \leftarrow \lceil \frac{\underline{\beta}\log_2(n)}{(d+2\overline{\beta})^2} \rceil$, $l_2 \leftarrow l_1 + \lceil \frac{1}{d}\log_2(\log(n)) \rceil$, $l_3 \leftarrow \lceil \frac{\overline{\beta}}{\underline{\beta}} l_1 + \frac{1}{\underline{\beta}}\log_2(\log(n)) \rceil$. ▷ bandwidths of reward function estimators
  Set $T \leftarrow \lceil C_1 n^{\frac{\underline{\beta}}{d+2\overline{\beta}}} \log(Kn) \rceil$. ▷ sample size for smoothness estimation
  for $t = 1, \ldots, T$ **do**
    Draw a sample X_t^Q, pull arm π_t uniformly at random from $[K]$, and get the reward $Y_t^{Q,(\pi_t)}$.
  end for
  Set $\mathcal{D}_{\mathsf{se}} \leftarrow \{(X_t^Q, \pi_t, Y_t^{Q,(\pi_t)})\}_{t=1}^T$, $s_P \leftarrow 1$, $s_Q \leftarrow T+1$. ▷ samples for smoothness estimation
end if
Define the grid $\mathcal{M}$

$$\mathcal{M} := \{x \in \mathcal{X} : x_i = (k_i - 1/2)2^{-l_3},\ k_i = [2^{l_3}], i = [d]\}.$$

for $i \in \{1,2\}$, $k \in [K]$ **do**
  Set $\mathcal{D}_{\mathsf{se}}^{(k)} \leftarrow \{(X_t, Y_t) : (X_t, \pi_t, Y_t) \in \mathcal{D}_{\mathsf{se}} \text{ such that } \pi_t = k\}$.
  for $x \in \mathcal{M}$ **do**
    Find the bin $B_i(x) \in \mathcal{B}_{l_i}$ such that $x \in B_i(x)$.
    Define the reward function estimator $\widehat{f}_k(x; 2^{-l_i}) := \widehat{\eta}_k(x; B_i(x))$, where

$$\widehat{\eta}_k(x; B) := \frac{\sum_{(X_t,Y_t)\in\mathcal{D}_{\mathsf{se}}^{(k)}} Y_t \mathbb{1}\{X_t \in B\}}{1 \vee \sum_{(X_t,Y_t)\in\mathcal{D}_{\mathsf{se}}^{(k)}} \mathbb{1}\{X_t \in B\}} \quad \forall B \in \mathcal{B}_{l_i}. \tag{16}$$

  end for
end for
Set $\mathsf{b} \leftarrow \max_{k\in[K]} \max_{x\in\mathcal{M}} |\widehat{f}_k(x; 2^{-l_1}) - \widehat{f}_k(x; 2^{-l_2})|$.
Set $\widehat{\beta} \leftarrow -\frac{1}{l_1}\log(\mathsf{b}) - C_2 \frac{\log_2(\log(n))}{\log_2(n)}$.
Output: smoothness estimate $(\underline{\beta} \vee \widehat{\beta}) \wedge \overline{\beta}$, time steps s_P and s_Q, policy $\{\pi_t\}_{1\le t<s_Q}$.

integer $\tau \ge 0$, bin B and arm k, we define

$$\widehat{U}_k(\tau, B; \mathcal{D}) := \begin{cases} 2\sqrt{\dfrac{2}{\tau + n_k^P(B;\mathcal{D})} \log^+\left(\dfrac{n_Q|B|^d}{\tau}\right)} \vee 2\overline{C}_\beta |B|^{\widehat{\beta}} & \text{if } \tau > 0, \\ 2\sqrt{\dfrac{2}{n_k^P(B;\mathcal{D})} \log\big(n|B|^{d+2\widehat{\beta}}\big)} \vee 2\overline{C}_\beta |B|^{\widehat{\beta}} & \text{if } \tau = 0, \end{cases} \tag{17}$$

where we recall the notation $\log^+(x) := \log(x) \vee 1$, $n := n_P \vee n_Q$ and $1/0 = \infty$. The associated nonnegative upper bound on play rounds is defined by

$$\widehat{\tau}_k^\star(B; \mathcal{D}) := \min_{\tau \in \{0\} \cup \mathbb{N}} \big\{ \tau : \widehat{U}_k(\tau, B; \mathcal{D}) \le 2\overline{C}_\beta |B|^{\widehat{\beta}} \big\}. \tag{18}$$

It is noteworthy that the source sample size within each bin in the partition tree is highly concentrated around its expectation, and that our choice of $\widehat{\tau}_k^\star(B; \mathcal{D})$ achieves a balanced bias-variance trade-off for estimating the reward functions in each bin. This allows Algorithm 2 to automatically adapt to the unknown parameters γ and κ, leading to a near-optimal regret.

Finally, it is worth mentioning that while Algorithm 2 requires an upper bound on the transfer exponent $\overline{\gamma}$, this information is solely utilized in the smoothness estimation stage (i.e., Procedure 2). As previously discussed, our smoothness estimation procedure uses T independent samples to evaluate the estimation bias of the local regression estimator under the self-similarity condition. To guarantee a good statistical performance of the local estimator in a bin with side length h, it is crucial to gather enough samples to balance the standard deviation (of order $(1/\sqrt{Th^{d+\gamma}})$ with the estimation bias (of order h^β). On the other hand, we also need to ensure that the regret incurred during the smoothness estimation phase is relatively small compared to the minimax regret. Therefore, achieving these two objectives necessitates the knowledge of an upper bound on the transfer exponent $\overline{\gamma}$. As an important remark, once the smoothness parameter estimate $\widehat{\beta}$ is generated by Procedure 2, $\overline{\gamma}$ is no longer used in the second phase of Algorithm 2.

We now present the theoretical guarantees of Algorithm 2. The proof is postponed to Appendix D in the Supplementary Material (Cai, Cai and Li (2024)).

THEOREM 3 (Upper bound). *Let $0 < \underline{\beta} < \overline{\beta} \le 1$ and $\overline{\gamma} \ge 0$. Suppose that $\kappa \asymp 1$ and $\alpha\beta \le d$. Then the policy π^{a} given by Algorithm 2 satisfies that for all $\beta \in [\underline{\beta}, \overline{\beta}]$ and $\gamma \in [0, \overline{\gamma}]$,*

$$\sup_{\Pi(K,\beta,\alpha,\gamma,\kappa,l_0,b)} \mathbb{E}\big[R_{n_Q}(\pi^{\mathsf{a}})\big] \le C_1 n_Q \big(n_Q + (\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}}\big)^{-\frac{\beta(1+\alpha)}{d+2\beta}} \log^{C_2}(n_P + n_Q), \tag{19}$$

for some constants $C_1 > 0$ and $C_2 > 0$ independent of n_Q and n_P.

All in all, Theorem 3 demonstrates that Algorithm 2 achieves the near-optimal minimax regret simultaneously for all $\beta \in [\underline{\beta}, \overline{\beta}]$ and $\gamma \in [0, \overline{\gamma}]$ when $K, \kappa \asymp 1$. In comparison with Theorem 1, the regret upper bound (19) contains an additional logarithmic factor, which can be viewed as the cost paid for smoothness adaptation. The condition $\kappa \asymp 1$ essentially assumes that the sample sizes corresponding to each arm in the source data are roughly of the same order, where one can achieve the most effective transfer learning.

Moreover, Theorem 4 below shows that the self-similarity assumption does not reduce the minimax complexity of the problem. The proof is deferred to Appendix C in the Supplementary Material (Cai, Cai and Li (2024)).

THEOREM 4 (Lower bound). *Assume that $\alpha\beta \le d$. For any constant $\beta \in (0, 1]$ and $l_0 \ge 0$, there exists a constant $b > 0$ that only depends on β, C_β, $\underline{q}$, $\overline{q}$ and d such that*

$$\inf_{\pi} \sup_{\Pi(K,\beta,\alpha,\gamma,\kappa,l_0,b)} \mathbb{E}\big[R_{n_Q}(\pi)\big] \ge c n_Q \big(n_Q + (\kappa n_P)^{\frac{d+2\beta}{d+2\beta+\gamma}}\big)^{-\frac{\beta(1+\alpha)}{d+2\beta}}, \tag{20}$$

for some constant $c > 0$ independent of n_Q and n_P.

Similar to Theorem 2, the infimum is taken over the class of admissible policies. Recognizing that the self-similar function space $\Pi(K, \beta, \alpha, \gamma, \kappa, l_0, b)$ is a subset of the general space $\Pi(K, \beta, \alpha, \gamma, \kappa)$, Theorem 4 and 2 together demonstrates that the minimax regret under the self-similar condition is the same as that in the general case. As a result, the self-similar condition does not reduce the complexity of the problem.

REMARK 5. We would like to remark that Suk and Kpotufe (2021) has also studied nonparametric contextual multi-armed bandits under the covariate shift model, with a particular focus on Lipschitz reward functions ($\beta = 1$). However, several significant distinctions exist between the analysis in our current work and that in Suk and Kpotufe (2021). For instance, a central challenge in our work is achieving smoothness adaptivity. Integrating the source data set to attain the minimax regret in the target bandit while at the same time adapting to the unknown smoothness parameter requires a substantially more complicated algorithmic design and technical analysis. Also, Suk and Kpotufe (2021) assumed the permission to collect data from the source bandit, allowing for active exploration. In contrast, our work deals with a fixed, precollected source data set. This limitation means that the source data might have been generated by a certain behavior policy, which might not provide sufficiently many data samples for important context-arm pairs. Effectively handling this limited data coverage becomes a critical challenge that governs the statistical efficiency of transfer learning.

5. Discussion. In this paper, we have studied transfer learning for nonparametric contextual multi-armed bandits under the covariate shift model. We establish the minimax regret that captures the amount of information transferred from the source domains to the target domain. A novel transfer learning algorithm is proposed to attain the minimax regret. Moreover, we also develop a data-driven algorithm that achieves within a logarithmic factor of the minimax regret while adapting to the unknown smoothness over a large class of parameter spaces under the self-similarity assumption.

There are several possible extensions worth pursuing. To begin with, while the current paper focuses on "rough" reward functions with smoothness parameter $\beta \in (0, 1]$, it is conceivable to generalize the algorithmic ideas to the case $\beta > 1$—one can adaptively partition the covariate space coupled with static multi-armed bandit procedures. However, we would like to remark on several critical differences. First, in contrast to the case $\beta \in (0, 1]$ where local piecewise constant estimators suffice to estimate the reward functions, one needs to use more complicated local polynomial estimators in the case $\beta > 1$. In addition, in our case $\beta \in (0, 1]$, as reward functions may be nondifferentiable, only samples close to the observed covariate are informative about their corresponding reward functions. Consequently, a fully localized learning strategy—running static multi-armed bandit procedures separately within each bin—guarantees to attain the minimax regret. Unfortunately, this rate-optimality no longer holds in the case $\beta > 1$. As the reward functions become smoother, we need to leverage global information and utilize observations from neighboring bins to extrapolate the reward functions efficiently, as introduced in Hu, Kallus and Mao (2022). Furthermore, since samples are used across bins, this also leads to the statistical dependence between decision-making in different bins. Therefore, the resulting policy is rather complicated and requires more careful statistical analysis. We leave it to future investigation due to the space limit.

Additionally, the regret upper bound of our adaptive policy mismatches the minimax rate by a logarithmic term. Whether this logarithmic factor is an inherent consequence of not knowing smoothness or an artifact of the proof remains unclear. It has been widely recognized in the literature on nonparametric function estimation that sharp adaptation is often achievable under global integrated squared error and that a logarithmic penalty arises under pointwise squared error. In contrast, when constructing confidence intervals, adaptation to

unknown smoothness without additional structural assumptions is typically impossible unless self-similarity or shape constraints are present. Notably, it was shown in Cai, Low and Xia (2013) that adaptive confidence intervals can be constructed for regression functions under shape constraints, such as concavity, and are near optimal for every individual function. Therefore, it is interesting to investigate further the cost of adaptation in the bandit problems, especially in the context of transfer learning. For example, developing adaptive transfer learning procedures for nonparametric contextual multi-armed bandits with concave reward functions presents an appealing direction.

Lastly, it is intriguing to study transfer learning for nonparametric contextual multi-armed bandits under other models. For instance, one avenue for future study is the posterior drift model, where the marginal distributions of the target and source bandits are identical whereas the conditional reward distributions differ. In this framework, it is also of interest to characterize the similarity between the reward distributions, establish the minimax rate of convergence that quantifies transferable information and develop data-driven adaptive algorithms.

SUPPLEMENTARY MATERIAL

Supplement to "Transfer learning for contextual multi-armed bandits" (DOI: 10.1214/23-AOS2341SUPP; .pdf). In this supplementary material, we provide proofs of Theorems 1–4, auxiliary lemmas and numerical experiments.

REFERENCES

ABE, N. and LONG, P. M. (1999). Associative reinforcement learning using linear probabilistic concepts. In *ICML* 3–11. Citeseer.

AGRAWAL, S., AVADHANULA, V., GOYAL, V. and ZEEVI, A. (2019). MNL-Bandit: A dynamic learning approach to assortment selection. *Oper. Res.* **67** 1453–1485. MR4014580 https://doi.org/10.1287/opre.2018.1832

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

AUER, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3** 397–422. MR1984023 https://doi.org/10.1162/153244303321897663

AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In 36*th Annual Symposium on Foundations of Computer Science* (*Milwaukee*, *WI*, 1995) 322–331. IEEE Comput. Soc. Press, Los Alamitos, CA. MR1619094 https://doi.org/10.1109/SFCS.1995.492488

AUER, P. and ORTNER, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Period. Math. Hungar.* **61** 55–65. MR2728432 https://doi.org/10.1007/s10998-010-3055-6

BASTANI, H. and BAYATI, M. (2020). Online decision making with high-dimensional covariates. *Oper. Res.* **68** 276–294. MR4059503 https://doi.org/10.1287/opre.2019.1902

BASTANI, H., BAYATI, M. and KHOSRAVI, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Manage. Sci.* **67** 1329–1349.

BEN-DAVID, S., BLITZER, J., CRAMMER, K. and PEREIRA, F. (2006). Analysis of representations for domain adaptation. *Adv. Neural Inf. Process. Syst.* **19**.

BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F. and WORTMAN, J. (2007). Learning bounds for domain adaptation. *Adv. Neural Inf. Process. Syst.* **20**.

BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456 https://doi.org/10.1214/12-EJS720

CAI, C., CAI, T. T. and LI, H. (2024). Supplement to "Transfer learning for contextual multi-armed bandits." https://doi.org/10.1214/23-AOS2341SUPP

CAI, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. MR2953494 https://doi.org/10.1214/11-STS355

CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494 https://doi.org/10.1214/009053604000000049

CAI, T. T., LOW, M. G. and XIA, Y. (2013). Adaptive confidence intervals for regression functions under shape constraints. *Ann. Statist.* **41** 722–750. MR3099119 https://doi.org/10.1214/12-AOS1068

CAI, T. T. and PU, H. (2022a). Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. Preprint. Available at arXiv:2401.12272.

CAI, T. T. and PU, H. (2022b). Stochastic continuum-armed bandits with additive models: Minimax regrets and adaptive algorithm. *Ann. Statist.* **50** 2179–2204. MR4474487 https://doi.org/10.1214/22-aos2182

CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128. MR4206671 https://doi.org/10.1214/20-AOS1949

CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning* 1042–1051. PMLR.

DEMIREL, I., CELIK, A. A. and TEKIN, C. (2022). Escada: Efficient safety and context aware dose allocation for precision medicine. *Adv. Neural Inf. Process. Syst.* **35** 27441–27454.

DING, K., LI, J. and LIU, H. (2019). Interactive anomaly detection on attributed networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* 357–365.

DÜMBGEN, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.* **26** 288–314. MR1611768 https://doi.org/10.1214/aos/1030563987

FARAHMAND, A.-M., SZEPESVÁRI, C. and MUNOS, R. (2010). Error propagation for approximate policy and value iteration. *Adv. Neural Inf. Process. Syst.* **23**.

GENOVESE, C. R. and WASSERMAN, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist.* **33** 698–729. MR2163157 https://doi.org/10.1214/009053605000000011

GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707 https://doi.org/10.1214/09-AOS738

GOLDENSHLUGER, A. and ZEEVI, A. (2009). Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* **19** 1603–1633. MR2538082 https://doi.org/10.1214/08-AAP589

GOLDENSHLUGER, A. and ZEEVI, A. (2013). A linear response bandit problem. *Stoch. Syst.* **3** 230–261. MR3353472 https://doi.org/10.1214/11-SSY032

GUR, Y., MOMENI, A. and WAGER, S. (2022). Smoothness-adaptive contextual bandits. *Oper. Res.* **70** 3198–3216. MR4538513 https://doi.org/10.1287/opre.2021.2215

HANNEKE, S. and KPOTUFE, S. (2019). On the value of target data in transfer learning. *Adv. Neural Inf. Process. Syst.* **32**.

HENGARTNER, N. W. and STARK, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23** 525–550. MR1332580 https://doi.org/10.1214/aos/1176324534

HU, Y., KALLUS, N. and MAO, X. (2022). Smooth contextual bandits: Bridging the parametric and nondifferentiable regret regimes. *Oper. Res.* **70** 3261–3281. MR4538516

KALLUS, N. and UDELL, M. (2020). Dynamic assortment personalization in high dimensions. *Oper. Res.* **68** 1020–1037. MR4166283 https://doi.org/10.1287/opre.2019.1948

KLEINBERG, R. and LEIGHTON, T. (2003). The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In 44*th Annual IEEE Symposium on Foundations of Computer Science*, 2003. *Proceedings*. 594–605. IEEE, New York.

KPOTUFE, S. and MARTINET, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *Ann. Statist.* **49** 3299–3323. MR4352531 https://doi.org/10.1214/21-aos2084

KULIS, B., SAENKO, K. and DARRELL, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR* 2011 1785–1792. IEEE, New York.

LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947. MR1447734 https://doi.org/10.1214/aos/1069362731

LEPSKII, O. V. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

LEPSKII, O. V. (1992). Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

LEPSKII, O. V. (1993). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation: Adaptive estimators. *Theory Probab. Appl.* **37** 433–448. MR1214353 https://doi.org/10.1137/1137095

LI, G., ZHAN, W., LEE, J. D., CHI, Y. and CHEN, Y. (2023). Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. Preprint. Available at arXiv:2305.10282.

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

LI, S., CAI, T. T. and LI, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 149–173. MR4400393

LI, S., CAI, T. T. and LI, H. (2023). Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *J. Amer. Statist. Assoc.* **118** 2171–2183. MR4646634 https://doi.org/10.1080/01621459.2022.2044333

LI, W., DUAN, L., XU, D. and TSANG, I. W. (2013). Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** 1134–1148.

LOCATELLI, A. and CARPENTIER, A. (2018). Adaptivity to smoothness in X-armed bandits. In *Conference on Learning Theory* 1463–1492. PMLR.

LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412 https://doi.org/10.1214/aos/1030741084

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

MA, C., PATHAK, R. and WAINWRIGHT, M. J. (2023). Optimally tackling covariate shift in RKHS-based nonparametric regression. *Ann. Statist.* **51** 738–761. MR4601000 https://doi.org/10.1214/23-aos2268

MAITY, S., SUN, Y. and BANERJEE, M. (2020). Minimax optimal approaches to the label shift problem. Preprint. Available at arXiv:2003.10443.

MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Domain adaptation: Learning bounds and algorithms. Preprint. Available at arXiv:0902.3430.

MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529–533.

MUNOS, R. (2007). Performance bounds in L_p-norm for approximate value iteration. *SIAM J. Control Optim.* **46** 541–561. MR2309039 https://doi.org/10.1137/040614384

NAKAMOTO, M., ZHAI, Y., SINGH, A., MARK, M. S., MA, Y., FINN, C., KUMAR, A. and LEVINE, S. (2023). Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. Preprint. Available at arXiv:2303.05479.

NICKL, R. and SZABÓ, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Process. Appl.* **126** 3913–3934. MR3565485 https://doi.org/10.1016/j.spa.2016.04.017

NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170

PAN, S. J. and YANG, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

PATHAK, R., MA, C. and WAINWRIGHT, M. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning* 17517–17530. PMLR.

PENG, M., LI, Y., WAMSLEY, B., WEI, Y. and ROEDER, K. (2021). Integration and transfer learning of single-cell transcriptomes via cFIT. *Proc. Natl. Acad. Sci. USA* **118** e2024383118.

PERCHET, V. and RIGOLLET, P. (2013). The multi-armed bandit problem with covariates. *Ann. Statist.* **41** 693–721. MR3099118 https://doi.org/10.1214/13-AOS1101

PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913 https://doi.org/10.1214/aos/1016120374

QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864

QIAN, W. and YANG, Y. (2016). Randomized allocation with arm elimination in a bandit problem with covariates. *Electron. J. Stat.* **10** 242–270. MR3466182 https://doi.org/10.1214/15-EJS1104

QUATTONI, A., COLLINS, M. and DARRELL, T. (2008). Transfer learning for image classification with sparse prototype representations. In 2008 *IEEE Conference on Computer Vision and Pattern Recognition* 1–8. IEEE, New York.

RABBI, M., AUNG, M. S., GAY, G., REID, M. C. and CHOUDHURY, T. (2018). Feasibility and acceptability of mobile phone-based auto-personalized physical activity recommendations for chronic pain self-management: Pilot study on adults. *J. Med. Internet Res.* **20** e10147. https://doi.org/10.2196/10147

RAGHU, M., ZHANG, C., KLEINBER, J. and BENGIO, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

RASHIDINEJAD, P., ZHU, B., MA, C., JIAO, J. and RUSSELL, S. (2022). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Trans. Inf. Theory* **68** 8156–8196. MR4544936 https://doi.org/10.1109/tit.2022.3185139

REEVE, H. W. J., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Adaptive transfer learning. *Ann. Statist.* **49** 3618–3649. MR4352543 https://doi.org/10.1214/21-aos2102

REEVE, H. W. J., MELLOR, J. and BROWN, G. (2018). The k-nearest neighbour UCB algorithm for multi-armed bandits with covariates. In *Algorithmic Learning Theory* 725–752. MR3857327

RIGOLLET, P. and ZEEVI, A. (2010). Nonparametric bandits with covariates. Preprint. Available at arXiv:1003.1630.

RINDTORFF, N. T., LU, M., PATEL, N. A., ZHENG, H. and D'AMOUR, A. (2019). A biologically plausible benchmark for contextual bandit algorithms in precision oncology using in vitro data. Preprint. Available at arXiv:1911.04389.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535. MR0050246 https://doi.org/10.1090/S0002-9904-1952-09620-8

ROSS, S. and BAGNELL, J. A. (2012). Agnostic system identification for model-based reinforcement learning. In *Proceedings of the* 29*th International Conference on Machine Learning*.

ROTHSCHILD, M. (1974). A two-armed bandit theory of market pricing. *J. Econom. Theory* **9** 185–202. MR0496544 https://doi.org/10.1016/0022-0531(74)90066-0

SHI, C., LU, W. and SONG, R. (2020). Breaking the curse of nonregularity with subagging—Inference of the mean outcome under optimal treatment regimes. *J. Mach. Learn. Res.* **21** Paper No. 176, 67 pp. MR4209462

SHRESTHA, S. and JAIN, S. (2021). A Bayesian-bandit adaptive design for N-of-1 clinical trials. *Stat. Med.* **40** 1825–1844. MR4229804 https://doi.org/10.1002/sim.8873

SOEMERS, D., BRYS, T., DRIESSENS, K., WINANDS, M. and NOWÉ, A. (2018). Adapting to concept drift in credit card transaction data streams using contextual bandits and decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32**.

SONG, Y., ZHOU, Y., SEKHARI, A., BAGNELL, J. A., KRISHNAMURTHY, A. and SUN, W. (2022). Hybrid RL: Using both offline and online data can make RL efficient. Preprint. Available at arXiv:2210.06718.

SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J. et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12** e1001779.

SUK, J. and KPOTUFE, S. (2021). Self-tuning bandits over unknown covariate-shifts. In *Algorithmic Learning Theory* 1114–1156. MR4227355

TEWARI, A. and MURPHY, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health* 495–517. Springer, Berlin.

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

WAGENMAKER, A. and PACCHIANO, A. (2023). Leveraging offline data in online reinforcement learning. In *International Conference on Machine Learning* 35300–35338. PMLR.

WANG, J., AGARWAL, D., HUANG, M., HU, G., ZHOU, Z., YE, C. and ZHANG, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16** 875–878.

WANG, Y., CHEN, B. and SIMCHI-LEVI, D. (2021). Multimodal dynamic pricing. *Manage. Sci.* **67** 6136–6152.

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** 1–40.

WOODROOFE, M. (1979). A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.* **74** 799–806. MR0556471

XIE, T. and JIANG, N. (2021). Batch value-function approximation with only realizability. In *International Conference on Machine Learning* 11404–11413. PMLR.

XIE, T., JIANG, N., WANG, H., XIONG, C. and BAI, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Adv. Neural Inf. Process. Syst.* **34** 27395–27407.

YANG, Y. and ZHU, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.* **30** 100–121. MR1892657 https://doi.org/10.1214/aos/1015362186

YU, X., WANG, J., HONG, Q.-Q., TEKU, R., WANG, S.-H. and ZHANG, Y.-D. (2022). Transfer learning for medical images analyses: A survey. *Neurocomputing* **489** 230–254.

ZHOU, Z., WANG, Y., MAMANI, H. and COFFEY, D. G. (2019). How do tumor cytogenetics inform cancer treatments? Dynamic risk stratification and precision medicine using multi-armed bandits. *Dynamic Risk Stratification and Precision Medicine Using Multi-armed Bandits (June 17, 2019)*.