MULTI-ARMED BANDITS WITH COVARIATES: THEORY AND APPLICATIONS

By

Dong Woo Kim
Tze Leung Lai
Huanzhong Xu

Technical Report No. 2020-15
November 2020

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

MULTI-ARMED BANDITS WITH COVARIATES: THEORY AND APPLICATIONS

By

Dong Woo Kim
Microsoft Corporation

Tze Leung Lai
Huanzhong Xu
Stanford University

Technical Report No. 2020-15
November 2020

This research was supported in part by National Science Foundation grant DMS 1811818.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Statistica Sinica

MULTI-ARMED BANDITS WITH COVARIATES: THEORY AND APPLICATIONS

Dong Woo Kim, Tze Leung Lai, and Huanzhong Xu

Microsoft Corporation and Stanford University

Abstract: "Multi-armed bandits" were introduced by Robbins (1952) as a new direction in the then-nascent field of sequential analysis developed during World War II in response to the need for more efficient testing of anti-aircraft gunnery, and subsequently by Bellman (1957) as a concrete application of dynamic programming and optimal control of Markov decision processes. A comprehensive theory that unified both directions emerged in the 1980s and provided important insights and algorithms for diverse applications in many STEM (Science, Technology, Engineering and Mathematics) fields. The turn of the millennium marks the onset of the "personalization revolution" – from personalized medicine to online personalized advertising and recommender systems (such as Netflix's recommendations for movies and TV shows, Amazon's recommendations for products to purchase, and Microsoft's Matchbox recommender) – that calls for the extension of classical bandit theory to nonparametric contextual bandits, in which "contextual" refers to the incorportation of personal information as covariates. Such theory is developed herein, together with illustrative applications, statistical models and computational tools for its implementation.

Key words and phrases: contextual multi-armed bandits, ϵ-greedy randomization, personalized medicine, recommender system, reinforcement learning.

1. Introduction and Background

The k-armed bandit problem was introduced by Robbins (1952) for $k = 2$ in his seminal paper on sequential design of experiments, in which he outlined new directions in sequential statistical methods beyond Wald's sequential probability ratio test (SPRT). Specifically he considered sequential sampling from two populations with unknown means to maximize the total expected reward $\mathbb{E}(y_1 + \cdots + y_n)$, where y_i has mean μ_1 (or μ_2) if it is sampled from population 1 (or 2) and n is the total sample size. Letting $s_n = y_1 + \cdots + y_n$, he applied the law of large numbers to show that $\lim_{n\to\infty} n^{-1}\mathbb{E}s_n = \max(\mu_1, \mu_2)$ is attained by the following rule: Sample from the population with the larger sample mean except at times belonging to a designated sparse set T_n of times, and sample from the population with the smaller sample size at these designated times; T_n is called "sparse" if $\#(T_n) \to \infty$ but $\#(T_n)/n \to 0$ as $n \to \infty$, where $\#(\cdot)$ denotes the cardinality of a set.

Thirty years passed, where Robbins revisited this problem to come up with a definitive solution to the problem of the optimal rule of convergence for $n(\max_{1\leqslant j\leqslant k} \mu_k) - \mathbb{E}\big(\sum_{t=1}^n y_t\big)$, leading to his 1985 paper with Lai who

was working on a general theory of sequential tests of composite hypotheses around that time. The first key step in the approach Lai and Robbins (1985) is the formulation of an *adaptive allocation rule* ϕ as a sequence of random variables $\phi_1, \ldots, \phi_n$ with values in the set $\{1, \ldots, k\}$ and such that the event $\{\phi_i = j\}, j \in \{1, \ldots, k\}$, belongs to the σ-field $\mathcal{F}_{i-1}$ generated by the previous observations $\phi_1, y_1, \ldots, \phi_{i-1}, y_{i-1}$. Letting $\mu(\theta) = \mathbb{E}_\theta y$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, it follows that

$$\mathbb{E}_{\boldsymbol{\theta}}\Big(\sum_{t=1}^{n} y_t\Big) = \sum_{t=1}^{n}\sum_{j=1}^{k} \mathbb{E}_{\boldsymbol{\theta}}\big\{\mathbb{E}_{\boldsymbol{\theta}}\big(y_t I_{\{\phi_t=j\}}|\mathcal{F}_{t-1}\big)\big\} = \sum_{j=1}^{k} \mu(\theta_j)\mathbb{E}_{\boldsymbol{\theta}}\tau_n(j),$$

where $\tau_n(j) = \#\{1 \leqslant t \leqslant n : \phi_t = j\}$ and Π_j is assumed to have density function $f_{\theta_j}(\cdot)$ from a parametric family of distributions. Hence, maximizing $\mathbb{E}_{\boldsymbol{\theta}}\big(\sum_{t=1}^{n} y_t\big)$ is equivalent to minimizing the regret

$$\begin{aligned} R_n(\boldsymbol{\theta}) &= n\mu^*(\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}}\Big(\sum_{t=1}^{n} y_t\Big) \\ &= \sum_{j:\mu(\theta_j)<\mu^*(\boldsymbol{\theta})} (\mu^*(\boldsymbol{\theta}) - \mu(\theta_j))\mathbb{E}_{\boldsymbol{\theta}}\tau_n(j), \end{aligned} \tag{1.1}$$

where $\mu^*(\boldsymbol{\theta}) = \max_{1\leqslant j\leqslant k} \mu(\theta_j)$. This representation enabled Lai to apply sequential testing theory, with which Lai and Robbins (1985) derived the basic lower bound for the regret (1.1) of uniformly good rules:

$$R_n(\boldsymbol{\theta}) \geqslant \left\{ \sum_{j:\mu(\theta_j)<\mu^*(\boldsymbol{\theta})} \frac{\mu(\theta^*) - \mu(\theta_j)}{I(\theta_j, \theta^*)} + o(1) \right\} \log n, \tag{1.2}$$

where $\theta^* = \theta_{j(\boldsymbol{\theta})}, j(\boldsymbol{\theta}) = \arg\max_{1\leqslant j\leqslant k} \mu(\theta_j)$, and an adaptive allocation rule is called "uniformly good" if $R_n(\boldsymbol{\theta}) = o(n^a)$ for every $a > 0$ and $\boldsymbol{\theta} \in$

Θ^k. Making use of the duality between hypothesis testing and confidence intervals, they also developed "upper confidence bound" (UCB) rules to attain the asymptotic lower bound (1.2).

The rest of this section begins with a summary of the dynamic programming approach to multi-armed bandits initialized by Bellman (1957), culminating in the index policy based on the dynamic allocation index (Gittins (1979); Whittle (1980)) of an arm, which Chang and Lai (1987) and Lai (1987) showed to be asymptotically equivalent to the UCB. This unified theory is reviewed in Section 1.1. Section 1.2 and 1.3 give overviews of two areas of subsequent developments – the first extends the parametric setting to nonparametric multi-armed bandits and the second extends it to (parametric) contextual bandits that also incorporate covariate information in the definition of regret.

Section 2 develops the methodology of nonparametric contextual bandits, and gives a synopsis of the "personalized revolution" in personalized medicine and recommender systems in the past two decades, which has called for the development of such methodology and its theory. Section 3 describes an extension of the methodology in Section 2 to high-dimensional covariates in the current Big Data and Multi-cloud era, and concludes with related literature and further discussion.

1.1 UCB rule and Gittins index: Asymptotic theory

Bellman (1957) introduced the dynamic programming approach to the 2-armed adaptive allocation problem considered by Robbins (1952), generalizing it to k arms and calling it a "k-armed bandit problem". The name derives from an imagined slot machine with k arms such that when an arm is pulled the player wins a random reward. For each arm j, there is an unknown probability distribution Π_j of the reward, hence there is a fundamental dilemma between "exploration" (to generate information about $\Pi_1, \ldots, \Pi_k$ by pulling the individual arms) and "exploitation" (of the information so that inferior arms are pulled minimally). Dynamic programming offers a systematic solution of the dilemma in the Bayesian setting but suffers from the "curse of dimensionality" as k and n increase. Gittins and Jones (1974) and Gittins (1979) considered the discounted version of this problem (thereby circumventing the issue of large horizon n) and showed that the k-dimensional stochastic optimization problem has an "index policy" (which does not have the curse of dimensionality) as its solution: At stage t, pull the arm with the largest "dynamic allocation index" (DAI) that depends only on the posterior distribution of the reward given the observed rewards from that arm up to stage t. The DAI is the solution to a non-standard optimal stopping problem that maximizes the quotient

$\mathbb{E}_j\big(\sum_{t=0}^{\tau-1}\beta^t Z_t\big)/\mathbb{E}_j\big(\sum_{t=0}^{\tau-1}\beta^t\big)$, where $\mathbb{E}_j$ denotes expectation under the posterior distribution of Π_j of the reward Z_t from arm j, given the observed rewards from the arm up to the stopping time τ, and $0 < \beta < 1$ is a discount factor. Whittle (1980) provided an alternative formulation of the DAI, which he called the "Gittins index", in terms of a family (indexed by a retirement reward M) of standard optimal stopping problems (involving $\mathbb{E}_j$ but not the quotient) that can be solved by dynamic programming.

We next review Lai (1987) who (a) connects UCB to generalized likelihood ratio (GLR) test statistics and to the Gittins index, and (b) shows that the UCB rule is uniformly good and attains the asymptotic lower bound (1.2) for the regret. He begins by considering the special case of $k = 2$ normal populations with means θ_1, θ_2, and variance 1, in which $\theta_2 = 0$ is known and θ_1 has a prior distribution with mean 0. In this case, the optimal rule is to sample from Π_1 until stage $\tilde{n} = \inf\{m \leqslant n : m^{-1}\sum_{i=1}^{m} y_i + a_{m,n} < 0\}$, and then take the remaining $n - \tilde{n}$ observations from Π_2, where $a_{m,n}$ are positive constants. Writing $t = m/n, w(t) = (y_1 + \cdots + y_m)/n^{1/2}, \delta = \theta n^{1/2}$, and treating $0 < t \leqslant 1$ as a continuous variable for large n, he approximates the Bayes stopping time for this special case by $n\tilde{\tau}(h)$, where $\tilde{\tau}(h) = \inf\{t \in (0,1] : w(t) + h(t) \leqslant 0\}$, and shows that an asymptotically optimal solution is the UCB rule: Sample at stage $t+1$ from Π_1 or Π_2 (with

known mean 0) according to $U_{1,t} > 0$ or $U_{1,t} \leqslant 0$, where $U_{j,t}$ is the upper confidence bound

$$U_{j,t} = \inf\left\{\theta : \theta \geqslant \hat{\theta}_{j,t} \text{ and } I\left(\hat{\theta}_{j,t}, \theta\right) \geqslant t^{-1} g(t/n)\right\}, \tag{1.3}$$

($\inf \emptyset = \infty$), $I(\lambda, \theta)$ is the the Kullback–Leibler information number, and $\hat{\theta}_{j,n}$ is the MLE of θ_j based on the observations from Π_j up to stage n. For the normal case, $\hat{\theta}_{1,n}$ is the sample mean from Π_1, $I(\lambda, \theta) = (\lambda - \theta)^2/2$, and $h(t) = (2tg(t))^{1/2}$. Lai (1987) has also extended the UCB rule to the exponential family in the k-armed bandit problem: Sample at stage $n+1$ from arm Π_j with the largest upper confidence bound (1.3). It is also shown in Lai (1987) that the UCB rule asymptotically minimizes the Bayes regret as $n \to \infty$ for a general class of prior distributions H. Although one can in principle use dynamic programming to minimize the Bayes regret $\int R_n(\boldsymbol{\theta}) dH(\boldsymbol{\theta})$, this approach is analytically and computationally intractable for large n. Instead of the finite-horizon problem that involves a given horizon n, Gittins (1979) considers the discounted infinite-horizon problem of maximizing $\int \cdots \int \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{\infty} \beta^{i-1} y_i\right] d\nu_1(\theta_1) \ldots d\nu_k(\theta_k)$, assuming a discount factor $0 < \beta < 1$ and independent prior distributions ν_j on the parameter space Θ. The optimal rule samples at stage $t+1$ from the arm Π_j with the largest Gittins index $G(\nu_{j,t})$, where $\nu_{j,t}$ is the posterior distribution of θ_j based on all the observations sampled from Π_j up to stage

t; see Whittle (1980). The index $G(\nu)$ of a distribution ν on θ is shown in Chang and Lai (1987) to be asymptotically equivalent to the upper confidence bound (1.2) when $n \sim 1/(1-\beta)$ and $t = o(n)$. Lai (1987) shows that the UCB rule also attains the Bayes regret

$$\int R_n(\boldsymbol{\theta}) dH(\boldsymbol{\theta}) \sim C(\log n)^2, \tag{1.4}$$

where C depends on the prior density function which is assumed to be positive and continuous over $\theta_j \in (\theta_j^* - \rho, \theta_j^* + \rho)$ for $1 \leqslant j \leqslant k$, with $\rho > 0$ and $\theta_j^* = \max_{i \neq j} \theta_i$.

1.2 Nonparametric extensions of classical bandit theory

Making use of large deviation bounds for sums of uniformly recurrent Markov chains or mixing stationary sequences, Lai and Yakowitz (1995) have extended the logarithmic lower bound (1.2) for the regret and the UCB rule attaining the bound to the nonparametric setting, pioneered by Yakowitz and Lowe (1991), in which "the only observables are the cost values and the probability structure and loss function are unknown to the designer" of the "black-box methodology". Assuming independent and bounded observations so that the "Chernoff-Hoeffding" large deviation bounds for their sums can be applied, Auer, Cesa-Bianchi and Fischer (2002) have developed another nonparametric method to attain the logarithmic lower bound

(1.2) for the regret. Instead of an UCB-type rule, they use the ϵ-greedy randomization algorithm in reinforcement learning proposed by Sutton and Barto (1998). Futher theoretical background and implementation details of the algorithm will be given in Section 2.2, where we also generalize it for our development of nonparametric contextual bandit methods.

1.3 Covariate information and parametric contextual bandits

Contextual multi-armed bandit problems, also called multi-armed bandits with side information, refer to the case where the decision maker also observes a covariate vector $\boldsymbol{x}_t$ that contains information on θ_j if y_t is sampled from Π_j at time t. Thus, arm Π_j is characterized by the conditional densities $f_{\theta_j}(\cdot|\boldsymbol{x}_t)$ for the reward y_t when the arm is pulled at time $t \geqslant 1$. Woodroofe (1979) was the first to consider the contextual multi-armed bandit problem for the case of $k = 2$ populations and univariate x_t that has distribution H, so that $f_{\theta_1}(y|x) = f(y - x - \theta_1)$ for some given density function f (i.e., f_{θ_1} is a location family), and $f_{\theta_2}(y|x) = f(y|x)$ does not have unknown parameters. Assuming a prior density function on θ_1 that is positive and continuous over an open interval and is 0 outside the interval, he showed that the myopic rule, which selects Π_1 whenever x_t exceeds the posterior mean of θ_1 given the observations up to time $t - 1$, is asymptotically op-

timal for the Bayesian discounted infinite-horizon problem of minimizing $\mathbb{E}\Big(\sum_{t=1}^{\infty}\beta_{t-1}y_t\Big)$ as $\beta \to 1$. This result was subsequently extended to the exponential family under certain regularity conditions by Sarkar (1991). Goldenshluger and Zeevi (2009) considered the finite-horizon non-Bayesian problem of choosing the n pulls sequentially to minimize $\mathbb{E}\big(y_1 + \cdots + y_n\big)$, assuming Π_2 to be degenerate at 0 and Π_1 to be normal with mean $x_t + \theta$ conditional on x_t. Analogous to (1.1), they define the regret in this simple case by

$$\begin{aligned} R_n(\theta) &= \mathbb{E}_\theta\bigg(\sum_{t=1}^{n}\phi_t^* y_t - \sum_{t=1}^{n}\phi_t y_t\bigg) \\ &= \mathbb{E}_\theta\bigg(\sum_{t=1}^{n} I_{\{\phi_t^* \neq \phi_t\}}|x_t + \theta|\bigg), \end{aligned} \tag{1.5}$$

where $\phi_t^* = I_{\{x_t+\theta \geqslant 0\}}$ is the oracle policy that assumes θ to be known, and show that the minimax regret $\inf_\phi \sup_\theta R_n(\theta)$ can be bounded or grow to ∞ with n at various rates that depend on the behavior of $\nu([-\theta-\delta, -\theta+\delta])$ as $\delta \to 0$. They also point out the paucity of papers on contextual bandit theory "in contrast to the voluminous literature on traditional multi-armed bandit problems."

Wang, Kulkarni and Poor (2005) were the first to generalize the parametric "one-armed" contextual bandit problem to the case $k = 2$, for which they proved two possibilities when the univariate covariate can only assume finitely many values: the "implicitly revealing" parameter configura-

tion with regret of the $O(1)$ order and other configurations for which the regret has the order of $\log n$. The case of more general covariates $\boldsymbol{x} \in \mathbb{R}^p$ in nonlinear regression models led Kim and Lai (2019) to develop the following general theory of parametric contextual bandits as a complete parallel to the classical context-free case. Assume the covariate vectors $\boldsymbol{x}_t$ are i.i.d. with common distribution H. Let supp$H$ denote the support of H, $f_\theta(y|\boldsymbol{x})$ denote the density function, depending on a parameter $\theta \in \Theta$ of the reward Y (with respect to some dominating measure ν on the real line) when the covariate vector has value $\boldsymbol{x}$, $\mu(\theta, \boldsymbol{x}) = \int y f_\theta(y|\boldsymbol{x}) d\nu(y)$,

$$j^*(\boldsymbol{x}) = \arg\max_{1 \leqslant j \leqslant k} \mu(\theta_j, \boldsymbol{x}), \quad \theta^*(\boldsymbol{x}) = \theta_{j^*(\boldsymbol{x})}, \tag{1.6}$$

in which θ_j is the parameter associated with arm j. Letting $\mathcal{F}_{t-1}$ denote the σ-field generated by $\{\boldsymbol{x}_t\} \cup \{(\boldsymbol{x}_s, y_s) : s \leqslant t-1\}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, the problem of choosing an adaptive allocation rule $\phi = (\phi_1, \ldots, \phi_n)$ to maximize $\mathbb{E}_{\boldsymbol{\theta}}\big(\sum_{t=1}^n y_t\big)$ is equivalent to minimizing the regret

$$\begin{aligned} & R_n(\boldsymbol{\theta}, B) \\ &= n \int_B \mu(\theta^*(\boldsymbol{x}), \boldsymbol{x}) dH(\boldsymbol{x}) - \sum_{t=1}^n \sum_{j=1}^k \mathbb{E}_{\boldsymbol{\theta}}\big\{\mathbb{E}_{\boldsymbol{\theta}}\big[y_t I_{\{\phi_t = j, \boldsymbol{x}_t \in B\}} | \mathcal{F}_{t-1}\big]\big\} \\ &= \sum_{j=1}^k \int_B \Big(\mu(\theta^*(\boldsymbol{x}), \boldsymbol{x}) - \mu(\theta_j, \boldsymbol{x})\Big) \mathbb{E}_{\boldsymbol{\theta}} \tau_n(j, \boldsymbol{x}) dH(\boldsymbol{x}), \end{aligned} \tag{1.7}$$

for Borel subsets B of supp$H$, for which $\mathbb{E}_{\boldsymbol{\theta}} \tau_n(j, B) := \sum_{t=1}^n \mathbb{P}_{\boldsymbol{\theta}}\{\phi_t = j, \boldsymbol{x}_t \in B\}$ defines a measure that is absolutely continuous with respect to the

common distribution H of the i.i.d. covariate vectors $\boldsymbol{x}_t$. Hence the term $\mathbb{E}_{\boldsymbol{\theta}}\tau_n(j,\boldsymbol{x})$ in (1.7)is the Radon-Nikodym derivative of the measure $\mathbb{E}_{\boldsymbol{\theta}}\tau_n(j,\cdot)$ with respect to H.

In view of the inclusion of the covariate set B in the definition (1.7) of the regret, Kim and Lai (2019) have extended the asymptotic lower bound (1.2) for the regret to contextual bandits as follows. An adaptive allocation rule ϕ is called "uniformly good" over $B \subset \text{supp}H$ if

$$R_n(\boldsymbol{\theta}, B) = o(n^a) \qquad \text{for every } a > 0 \text{ and } \boldsymbol{\theta} \in \Theta^k. \tag{1.8}$$

Moreover, an analogue of $I(\theta_j, \theta^*)$ in (1.2) for the contextual setting is

$$\begin{aligned} I(\theta, \lambda; \boldsymbol{x}) &= \mathbb{E}_\theta\left\{ \log \frac{f_\theta(y|\boldsymbol{x})}{f_\lambda(y|\boldsymbol{x})} \right\}, \\ I_{\boldsymbol{x}}(\theta, \theta') &= \inf_{\lambda:\mu(\lambda,\boldsymbol{x})=\mu(\theta',\boldsymbol{x})} I(\theta, \lambda; \boldsymbol{x}). \end{aligned} \tag{1.9}$$

Note that $I(\theta, \lambda; \boldsymbol{x})$ is a natural extension of the Kullback–Leibler information number to conditional densities. The quantity $I_{\boldsymbol{x}}(\theta, \theta')$ corresponds to $I(\theta, \lambda; \boldsymbol{x})$ with the least informative λ over the surface $\mu(\lambda, \boldsymbol{x}) = \mu(\theta', \boldsymbol{x})$.

Kim and Lai (2019) have extended (1.2) to contextual bandits under mild regularity conditions:

Theorem 1. *(i) If j^* is constant over B, then*

$$R_n(\boldsymbol{\theta}, B) \geqslant (1 + o(1)) \sum_{j:p_j(\boldsymbol{\theta})=0} (\log n) \int_B \frac{\mu(\theta^*(\boldsymbol{x}), \boldsymbol{x}) - \mu(\theta_j, \boldsymbol{x})}{I_{\boldsymbol{x}}(\theta_j, \theta^*(\boldsymbol{x}))} dH(\boldsymbol{x}), \tag{1.10}$$

where $\sum_j$ *over an empty set is interpreted as* $O(1)$ *and* $p_j(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}\{j^*(\boldsymbol{X}) = j\}$*, in which* $\boldsymbol{X} \in \mathbb{R}^p$ *has distribution* H*.*

(ii) If j^* *is non-constant over* B *(i.e.,* B *contains leading arm transitions), then*

$$R_n(\boldsymbol{\theta}, B) \geqslant C(\boldsymbol{\theta})(\log n)^2. \tag{1.11}$$

(iii) The adaptive allocation rule summarized in the last sentence of Section 2.2 attains the preceding asymptotic lower bounds.

2. Theory of Nonparametric Contextual Bandits

In Section 2.1 we generalize the definition of regret (1.7) for contextual bandits to the nonparametric setting and derive the analog of (1.10) and (1.11) for the asymptotic lower bound of the regret. Section 2.2 develops an adaptive allocation rule ϕ_{opt} whose regret has the same "minimax rate" (which will be defined there) as that of the lower bound. After an overview of statistical and computational tools for its implementation, Section 2.3 provides simulation studies of its performance. A synopsis of personalized medicine and recommender systems in the past two decades is given in Supplementary Material S2, which also describes how the methodology developed in Section 2.2 can be applied to these areas.

2.1 Lower bound of the regret over a covariate set

For the classical (context-free) multi-armed bandit problem, Lai and Yakowitz (1995) define the regret as $\sum_{j=1}^{k}(\mu^* - \mu_j)\mathbb{E}\tau_n(j)$ as a natural extention of (1.1) to the nonparametric setting, where μ_j is the expected reward from arm j and $\mu^* = \max_{1\leqslant j\leqslant k}\mu_j$. Combining this with (1.7) for parametric contextual bandits leads to the definition of the regret

$$R_{n,\phi}(B) = \sum_{j=1}^{k}\int_B (\mu^*(\boldsymbol{x}) - \mu_j(\boldsymbol{x}))\mathbb{E}\tau_n(j,\boldsymbol{x})dH(\boldsymbol{x}) \tag{2.1}$$

of an adaptive allocation rule ϕ over Borel subsets B of supp$H$, where $\mathbb{E}\tau_n(j,\boldsymbol{x})$ is the Radon-Nikodym derivative of the measure $\mathbb{E}\tau_n(j,\cdot)$ with respect to H. Moreover, in analogy with (1.8), we call ϕ "uniformly good" over B if $R_{n,\phi}(B) = o(n^a)$ for every $a > 0$. It will be shown that the nonparametric family $\mathcal{P}$ generating the data contains a least favorable parametric subfamily and that the regret of the adaptive allocation rule ϕ_{opt} defined in the next subsection attains the minimax risk rate for this parametric subfamily under certain regularity conditions on $\mathcal{P}$. Details and the background literature for this approach are given in Supplementary Material S1.

2.2 ϵ-greedy randomization and arm elimination

The UCB rule in Section 1.1, introduced by Lai (1987) to approximate the index policy of Gittins and Whittle in classical (context-free) parametric multi-armed bandits, basically samples from an inferior arm until the sample size from it reaches a threshold define by (1.3) involving the Kullback–Leibler information number. For contextual bandits, an arm that is inferior at $\boldsymbol{x}$ may be best at another $\boldsymbol{x}'$. Hence the index policy that samples at stage t from the arm with the largest upper confidence bound (which modifies the sample mean reward by incorporating its sampling variability at $\boldsymbol{x}_t$) can be improved by deferral to future time t' when it becomes the leading arm (based on the sample mean reward up to time t'), as shown for contextual parametric bandits by Kim and Lai (2019, Section III) who propose to use the ϵ-greedy randomization algorithm in reinforcement learning (Sutton and Barto, 1998), which we generalize to nonparametric contextual bandits as follows. Let K_t denote the set of arms to be sampled from and

$$J_t = \Big\{ j \in K_t : \Big| \hat{\mu}_{j,t-1}\big(\boldsymbol{x}_t\big) - \hat{\mu}^*_{t-1}\big(\boldsymbol{x}_t\big) \Big| \leqslant \delta_t \Big\}, \tag{2.2}$$

where $\hat{\mu}_{j,s}(\cdot)$ is the regression estimate (which will be described in the next paragraph) of $\mu_j(\cdot)$ based on observations up to time s, $\hat{\mu}^*_s(\cdot) = \max_{j \in K_s} \hat{\mu}_{j,s}(\cdot)$, and δ_t lumps treatments with effect sizes close to that of

the apparent leader into a single set J_t. At time t, choose arms randomly with probabilities $\pi_{j,t} = \epsilon/|K_t \setminus J_t|$ for $j \in K_t \setminus J_t$ and $\pi_{j,t} = (1-\epsilon)/|J_t|$ for $j \in J_t$, where $|A|$ denotes the cardinality of a finite set A. The set K_t is related to the arm elimination scheme described later.

Ibragimov and Has'minskii (1981) and Begum et al. (1983) have introduced the theory of information bounds and minimax risk into nonparametric or semiparametric (i.e., parametric for the parameters of interest and infinite-dimensional nonparametric nuisance parameters). This is also closely related to the least favorable parametric subfamily of the nonparametric family $\mathcal{P}$ introduced by Stein (1956) and Bickel (1982). Fan (1993) has shown that local polynomial regression has minimax risk rate for univariate regressors; see also Hastie and Loader (1993) and Fan and Gijbels (1996) for subsequent developments, including Ruppert and Wand (1994) who have extended local linear regression to multivariate regressors.

Arm Elimination. Choose $n_i \sim a^i$ for some interger $a > 1$. For $n_{i-1} < t \leqslant n_i$, eliminate surviving arm j if

$$\hat{\mu}_{j,t-1}(\boldsymbol{x}_t) < \hat{\mu}^*_{t-1}(\boldsymbol{x}_t) \text{ and } \Delta_{j,t-1} > g(n_{j,t-1}/n_i), \tag{2.3}$$

where $n_{j,s} = T_s(j)$, g is given in (1.3) and $\Delta_{j,t-1}$ is the square of the Welch

statistic based on $\{(\boldsymbol{x}_\ell, y_\ell) : 1 \leqslant \ell \leqslant t-1\}$, i.e.,

$$\Delta_{j,t-1} = \sum_{\ell=1}^{t-1} I_{\{\phi_\ell = j\}} \frac{\Big(\hat{\mu}_{j,\ell-1}(\boldsymbol{x}_\ell) - \tilde{\mu}_{j,\ell-1}(\boldsymbol{x}_\ell)\Big)_+^2}{\Big(y_\ell - \hat{\mu}_{j,\ell-1}(\boldsymbol{x}_\ell)\Big)^2 + \Big(y_\ell - \tilde{\mu}_{j,\ell-1}(\boldsymbol{x}_\ell)\Big)^2}\,, \tag{2.4}$$

in which $a_t = \max(a, 0)$ and $\tilde{\mu}_{j,s}(\cdot) = \max_{j' \in K_s} \hat{\mu}_{j'}(\cdot)$ if $j \in K_s \setminus J_s$, which corresponds to the local linear regression estimate of $\mu_j(\cdot)$ under the null hypothesis $H_{j,s}$, under which $\tilde{\mu}_{j,s}(\cdot) = \hat{\mu}_{j,s}(\cdot)$ if $j \in J_s$. This adaptive allocation procedure will be denoted by ϕ_{opt}.

Note that (2.4) is the nonparametric analog of the generalized likelihood rate (GLR) statistic for testing $H_{j,t}$ in parametric models described in this penultimate paragraph of Section 1.3, for which Kim and Lai (2019, Section III) replace (2.2) by

$$J_t = \Big\{ j \in K_t : \Big| \mu(\hat{\theta}_{j,t-1}, \boldsymbol{x}_t) - \mu(\hat{\theta}^*_{t-1}(\boldsymbol{x}_t), \boldsymbol{x}_t) \Big| \leqslant \delta_t \Big\}$$

and (2.4) by

$$\Delta_{j,t-1} = \sum_{\ell=1}^{t-1} I_{\{\phi_t = j\}} \log \Big(f_{\hat{\theta}_{j,\ell-1}}(y_\ell | \boldsymbol{x}_\ell) \Big/ f_{\tilde{\theta}_{j,\ell-1}}(y_\ell | \boldsymbol{x}_\ell) \Big), \tag{2.5}$$

letting $\hat{\theta}_{j,\ell-1}$ (respectively, $\tilde{\theta}_{j,\ell-1}$) be the MLE (respectively, constrained MLE under the constraint $\mu(\theta_j, \boldsymbol{x}_\ell) \geqslant \max_{j' \in K_\ell \setminus \{j\}} \mu(\hat{\theta}_{j',\ell-1}, \boldsymbol{x}_\ell)$ for $1 \leqslant \ell \leqslant t-1$) and using the same notation as in (1.6) and (1.9).

2.3 Asymptotic efficiency and simulation study of performances

The adaptive allocation procedure in the preceding subsection, using (a) nonparametric local linear regression estimate $\hat{\mu}_{j,s}(\cdot)$ of $\mu_s(\cdot)$ in (2.1), (b) ϵ-greedy randomization to sample from the set K_t of surviving arms, and (c) the arm elimination rule define by (2.3) and (2.4), has regret that attains the rate of the minimax risk under certain regularity conditions on $\mathcal{P}$ that are given in the Supplementary Material S1, which also gives the proof of the following.

Theorem 2. *Under the regularity conditions and the choice of bandwidth for* $(\hat{\mu}_{j,s}(\cdot) - \tilde{\mu}_{j,s}(\cdot))_+$ *given in S1,* ϕ_{opt} *attains the asymptotic minimax rate (as* $n \to \infty$*) of the risk functions for adaptive allocation rules.*

The background of minimax risk in asymptotic statistical decision theory will be given in S1 and then applied to the proof of Theorem 2. We also report a simulation study of performances of ϕ_{opt}, using "binned" regression estimate and local linear regression estimate of $\mu_s(\cdot)$ respectively.

We simulate a bandit consisting of six arms with their means being horizontally shifted sine functions. Given x_t, arm i follows a normal distribution with mean $\mu_i(x_t) = \sin\left(x_t + \frac{i}{6}\pi\right)$, where $i = 0, \ldots, 5$, and standard deviation 0.1. The mean reward functions are shown in Figure 1.

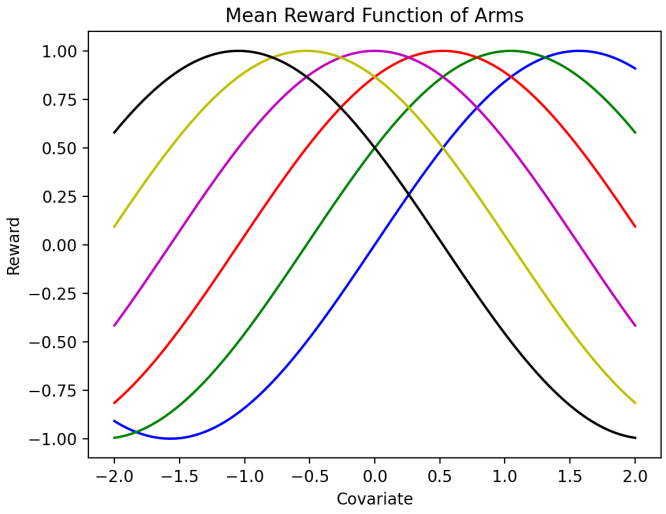

Figure 1: Mean Reward Function of Arms

We test three policies in this simulation study. The first policy we pick is UCBogram by Rigollet and Zeevi (2010), which estimates $\mu_s(\cdot)$ by "binned regression" and adopt UCB for each bin. We denote this policy by ϕ^{bin}. The second policy we simulate is ϕ^{bin}_{opt} where we replace UCB by ϵ-greedy randomization and the arm elimination rule. The process of ϕ^{bin}_{opt} learning reward functions is shown in Figure 2. Finally we replace the "binned" regression by local linear regression, and denote this policy by ϕ_{opt}.

The simulation is run for a horizon $N = 30000$. For each $t = 1, \ldots, 30000$, x_t is simulated from the uniform distribution between $(-2, 2)$. The number of bins for both ϕ^{bin}_{opt} and ϕ^{bin} is 60. The regrets over time of three policies

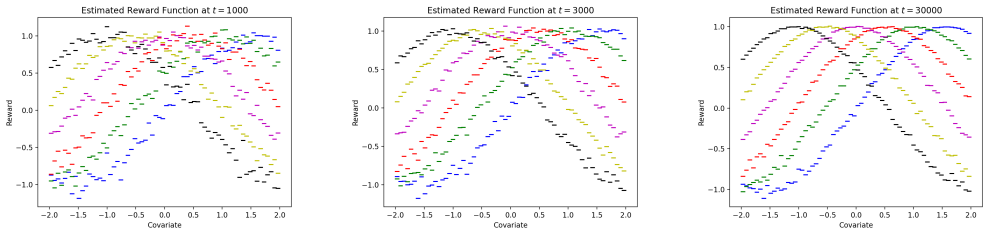

Figure 2: Fitting Reward Functions

are shown in Figure 3.

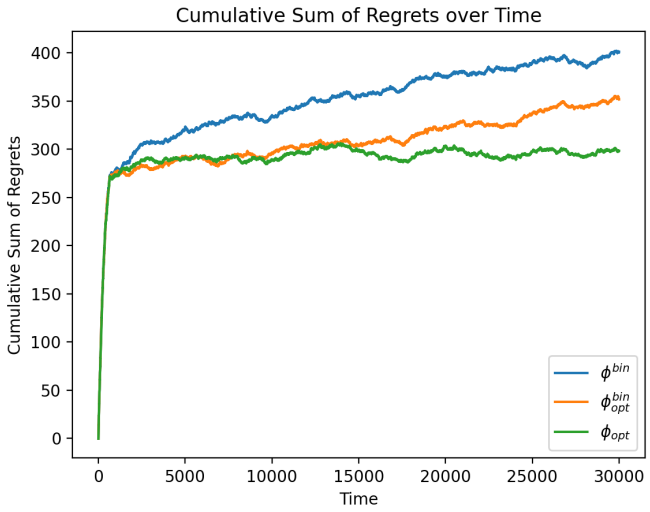

Figure 3: Cumulative Regrets of ϕ^{bin}, ϕ^{bin}_{opt} and ϕ_{opt}

3. High-dimensional Covariates and Concluding Remarks

Supplementary Material S3 gives an overview of machine learning for recommender systems and personalization technologies in the current Big Data

and Multi-cloud era, after extending the nonparametric contextual bandit theory in Section 2 (dealing with the case of fixed p and large n) to high-dimensional covariates for which $p = p_n$ may exceed n. In this connection it also reviews the work of Birgé and Massart (1993), Shen and Wong (1994), Yang and Barron (1999), Yang and Tokdar (2015) on the information-theoretic approach to minimax rates of convergence that provides a powerful method to tackle high-dimensional covariates.

In conclusion, multi-armed bandits with "side information" or covariates, also called contextual multi-armed bandits arise in many fields of applications, in which the development of personalized strategies or recommender systems has its statistical underpinnings in the theory of contextual multi-armed bandits. We have developed herein a comprehensive theory of contextual multi-armed bandits and derived new results on nonparametric contextual bandits and extended them to high-dimensional covariates, which are of particular interest in the current Big Data and Multi-Cloud era.

Acknowledgement

Lai's research is supported by the National Science Foundation under DMS-1811818.

References

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine Learning* **47**, 235–256.

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432–452.

Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647–671.

Birgé, L. and Massart, P., (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory & Related Fields* **97**, 113–150.

Chang, F. and Lai, T. L. (1987). Optimal stopping and dynamic allocation. *Adv. in Appl. Probab.* **19**, 829–853.

Fan, J. (1993). Local linear regression smoothness and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

Fan, J. and Gigbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC, Boca Raton, FL.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41**, 148–177.

Gittins, J. C. and Jones D. M. (1974). *A dynamic allocation index for the sequential design of experiments*. University of Cambridge, Department of Engineering.

Goldenshluger, A. and Zeevi, A. (2009). Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* **19**, 1603–1633.

Hastie, T. and Loader, C. (1993). Local regression: automatic kernel carpentry. *Statist. Sci.* **8**, 120–129.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, Heidelberg-Berlin-New York.

Kim, D. W. and Lai, T. L. (2019). Asymptotically Efficient Randomized Allocation Schemes for the Multi-armed Bandit Problem with Side Information.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15**, 1091–1114.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6**, 4–22.

Lai, T. L. and Yakowitz, S. (1995). Machine learning and nonparametric bandit theory. *IEEE Transactions and Automatic Control* **40**, 1199-1209.

Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630.*

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58**, 527–535.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression.

Ann. Statist. **22**, 1346–1370.

Sarkar, J. (1991). One-armed bandit problems with covariates. *Ann. Statist.* **19**, 1978–2002.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580–615.

Stein, C. (1956). Efficient Nonparametric Testing and Estimation. *Proc. Third Berkeley Symp. on Math. Statist. and Prob.* **1**, 187–195.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Wang, C. C., Kulkarni, S. R. and Poor, H. V. (2005). Bandit problems with side observations. *IEEE Trans. Automat. Control* **50**, 338–355.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B* **42**, 143–149.

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

Yakowitz, S. and Lowe, W. (1991). Nonparametric bandit methods. *Annals of Operations Research* **28**, 291-312.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27**, 1564–1599.

Yang, Y. and Tokdar, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43**, 652–674.

Analysis and Experimentation Team, Microsoft Corporation

E-mail: dongwookim80@gmail.com

Department of Statistics, Stanford University

E-mail: lait@stanford.edu

Institute for Computational and Mathematical Engineering, Stanford University

E-mail: xuhuanvc@stanford.edu

Running Head: CONTEXTUAL MULTI-ARMED BANDITS

MULTI-ARMED BANDITS WITH COVARIATES: THEORY AND APPLICATIONS

Microsoft Corporation and Stanford University

Supplementary Material

S1 Background literature and proof of Theorem 2

We have summarized in the second paragraph of Section 2.2 some background literature on local linear regression estimate of $\mu_j(\cdot)$ in the regret (2.1) and the associated minimax risk. We want to add here the works of Yang and Zhu (2002) and Rigollet and Zeevi (2010) who consider local polynomials of degree 0 (i.e., piecewise constant or "binned" regression estimates), and subsequent work along this line by Perchet and Rigollet (2013). We need to emphasize upfront a major difference between our method (in particular, $\Delta_{j,t-1}$ defined by (2.4)) and these previous approaches to contextual bandits via nonparametric classification and regression (involving minimax estimation of $\mu_j(\cdot)$ for $j = 1, 2, \ldots, k$). As pointed out in the last sentence of that section, $\Delta_{j,t-1}$ originated from the GLR statistic (2.5) in

parametric contextual bandits reviewed in Section 1.3, where Theorem 1 provides a definitive result on the asymptotic lower bound for the regret and attainment of that bound by using ϵ-greedy randomization and arm elimination. Since $(\hat{\mu}_{j,\ell-1}(\boldsymbol{x}_\ell) - \tilde{\mu}_{j,\ell-1}(\boldsymbol{x}_\ell))_+$ is the key ingredienct in (2.4), contextual bandits should consider estimation of $(\mu_j(\cdot) - \max_{j' \neq j} \mu_{j'}(\cdot))_+$, instead of $\mu_j(\cdot), 1 \leqslant j \leqslant k$ in the previous methods. This approach yields that if $\mu_j(\cdot)$ exceeds $\max_{j' \neq j} \mu_{j'}(\cdot)$ by a substantial amount over a covariate set $B \subset \text{supp}H$ as in Theorem 1(i), then the regret over B is of order $O(\log n)$. On the other hand, if B contains leading arm transitions for which it is difficult to distinguish locally two leading arms j and j', then the regret is of $O((\log n)^2)$ under smoothness conditions on $\mu_j(\cdot) - \mu_{j'}(\cdot)$. Perchet and Rigollet (2013, p.695) have actually introduced an "adaptively binned successive elimination (ABSE)" procedure to "partition the space of covariates in a fashion that adapts to the local difficulty of the problem: cells are smaller when different arms are hard to distinguish and bigger when one arm dominates the other", which seems to be similar to our approach. On the other hand, the regret rate of ABSE which is claimed in their Section 5 to be "optimal in a minimax sense" (of nonparametric k-class classification due to Audibert and Tsybakov, 2007) differs from the minimax rate over $B \subset \text{supp}H$ in Theorem 2 on the asymptotic statistical decision problem

associated with nonparametric contextual k-armed bandits.

Choice of bandwidth in Theorem 2. For univariate covariates ($p = 1$), Fan (1993) has shown that the bandwidth choice $b_n \approx n^{-1/5}$ for the local linear regression estimate

$$\hat{m}(x) = \sum_{\ell=1}^{n} w_\ell(x) y_\ell \Big/ \sum_{\ell=1}^{n} \Big(w_\ell(x) + n^{-2} \Big) \tag{S1}$$

of a regression function $m(x) = \int y f(y|x) d\nu(y)$, based on a random sample $(x_\ell, y_\ell), 1 \leqslant \ell \leqslant n$, from a distribution with unknown conditional density function $f(\cdot|x)$ with respect to some measure ν, yields asymptotically minimax rates for mean squared errors, where $\approx$ denotes the same order of magnitude (i.e., $c_1 n^{-1/5} \leqslant b_n \leqslant c_2 n^{-1/5}$ for some constant $c_1 < c_2$). The weights $w_\ell(x)$ in (S1) are given by

$$w_\ell(x) = K\big((x-x_\ell)/b_n\big)\big\{s_{n,2} - (x-x_\ell)s_{n,1}\big\},\ s_{n,j} = \sum_{\ell=1}^{n} K\big((x-x_\ell)/b_n\big)\big(x-x_\ell\big)^j$$

for $j = 0, 1, 2$, in which $K \geqslant 0$ is a kernel function (i.e., $\int_{-\infty}^{\infty} K(u) \mathrm{d}u = 1$). For multivariate covariates $\boldsymbol{x}_\ell$, Ruppert and Wand (1994) define the $n \times (p+1), (p+1) \times 1$, and $n \times 1$ matricies

$$\boldsymbol{A}_n(\boldsymbol{x}) = \begin{bmatrix} 1 & \begin{pmatrix} \boldsymbol{x}_1^T - \boldsymbol{x}^T \end{pmatrix} \\ 1 & \begin{pmatrix} \boldsymbol{x}_2^T - \boldsymbol{x}^T \end{pmatrix} \\ \vdots & \begin{pmatrix} \vdots \end{pmatrix} \\ 1 & \begin{pmatrix} \boldsymbol{x}_n^T - \boldsymbol{x}^T \end{pmatrix} \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \boldsymbol{Y}_n = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \tag{S2}$$

and the $p \times p$ bandwidth matrix $\boldsymbol{B}_n = \mathrm{diag}(b_n^1, \ldots, b_n^p)$ so that

$$\hat{m}(\boldsymbol{x}) := \boldsymbol{e}^T \Big[\boldsymbol{A}_n(\boldsymbol{x}) \boldsymbol{W}_n(\boldsymbol{x}) \boldsymbol{A}_n(\boldsymbol{x}) \Big]^{-1} \boldsymbol{A}_n^T(\boldsymbol{x}) \boldsymbol{W}_n(\boldsymbol{x}) \boldsymbol{Y}_n \tag{S3}$$

is the local linear regression estimate of $m(\boldsymbol{x}) := \mathbb{E}(Y|\boldsymbol{x})$, in which $\boldsymbol{W}_n(\boldsymbol{x}) = \mathrm{diag}\big(K_n(\boldsymbol{x}_1 - \boldsymbol{x}), \ldots, K_n(\boldsymbol{x}_n - \boldsymbol{x})\big)$, where $K_n(\boldsymbol{u}) = |\boldsymbol{B}_n|^{-1/2} K\big(\boldsymbol{B}_n^{1/2} \boldsymbol{u}\big)$ and K is a bounded kernel such that $\int \boldsymbol{u}\boldsymbol{u}^T K(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \propto \boldsymbol{I}_p$ when certain regularity conditions are satisfied; see Ruppert and Wand (1994, p.1349–1350). Hence Fan's argument can be extended to multivariate covariates by choosing $b_n^i \approx n^{-1/5}$ for $i = 1, \ldots, p$.

Choice of δ_t in (2.2) and regularity conditions in Theorem 2. Kim and Lai (2019) choose $\delta_t > 0$ such that $\delta_t^2 = \big(2 \log t\big)/t$, which they use to prove Theorem 1(iii) given in Section 1.3 above. As will be shown in the proof of Theorem 2 in the next paragraph, this choice also works for nonparametric contextual bandits for which it is particularly effective in the vicinity of leading arm transitions. We next state the regularity conditions, which relax somewhat those of Ruppert and Wand (1994, p.1349–1350) and Fan (1993, p.199, in the simpler case $p = 1$), for Theorem 2:

(a) The common distribution H of the i.i.d. covariate vectors $\boldsymbol{x}_t$ has a positive density function f (with respective to Lebesgue measure) which is continuously differentiable on a hyperrectangle in $\mathbb{R}^p$.

(b) m is twice continuously differentiable and $\sigma^2(\boldsymbol{x}) := \mathrm{Var}\big(Y|\boldsymbol{x}\big)$ is positive and continuous on $\mathrm{supp}H$ (i.e., the hyperrectangle in (a)).

(c) The bounded kernel K is continuous and $\int |\boldsymbol{u}|^r K(\boldsymbol{u})\mathrm{d}\boldsymbol{u} < \infty$ for all $r \geqslant 1$, $\int u_i K(\boldsymbol{u})\mathrm{d}\boldsymbol{u} = 0$ for $i = 1, \ldots, p$.

Least favorable parametric subfamily and nonparametric minimax rates in asymptotic decision theory. In Section 2.1 we have mentioned the least favorable parametric subfamily approach to deriving lower bounds for the risk functions in statistical decision problems. This idea dated back to Stein (1956), and Bickel (1982) gave a review of the developments in adaptive estimation during the twenty-five years after Stein's seminal work on the problem of "estimating and testing about a Euclidean parameter θ, or more generally, a function $q(\theta)$ in the presence of an infinite-dimensional nuisance parameter G" so that θ or $q(\theta)$ can be estimated nonparametrically (without knowledge of G) as well asymptotically as knowing G. Begun et al. (1983) develop these lower bounds for semiparametric estimation of a finite-dimensional (multivariate) parameter $\boldsymbol{\theta}$ in the presence of an infinite-dimensional nuisance parameter G via "representation theorems (for regular estimators) and asymptotic minimax bounds". In particular, they apply this approach to prove the efficiency of Cox regression for censored data in the proportional hazards model for survival analysis. Lai and Ying (1992)

consider rank estimators in the usual regression model when the observed responses are subject to left truncation and right censoring, for which they extend the asymptotic minimax bounds of Begun et al. (1983) by making use of (a) the martingale structure of left truncated and right censored data and martingale central limit theorem, (b) quadratic-mean differentiability of the hazard function, and (c) the Hájek convolution theorem for regular estimators in parametric submodels of the nonparametric model for G. To estimate a regression function that satisfies regularity conditions of the type in the preceding paragraph, Fan (1993) shows that the local linear estimator introduced therein attains asymptoticlly minimax rates in the sense that the minimax risk (Bickel, 1982; Pinsker, 1980; Donoho, Liu and MacGibbon, 1990) has order $\approx n^{-4/5}$ whereas the local linear estimator has minimax risk of the order $n^{-4/5+o(1)}$; Fan considers the univariate case $p = 1$ and mean squared error as the risk function.

Exponential bounds for self-normalized statistics. Exponential bounds have been established for the GLR statistics (2.5), which are self-normalized, in parametric models; see de la Peña, Lai and Shao (2009, p.207–210, 216). The Welch statistics (2.4) in the nonparametric setting are generalized Studentized (and therefore self-normalized) statistics, for which exponential bounds hold and play an important role in the proof of Theorem 2.

Minimax theorem and asymptotic decision theory. Whereas the asymptotic minimax rates of the background literature reviewed in the preceding paragraphs are stated in terms of nonparametric regression or classification, the nonparametric contextual k-armed bandit problem is actually about asymptotically minimax statistical decision rules for sequential selection (rather than estimation or classification) from k given arms as described in Section 2.1; see Strasser (1985, p.238–242, 308–327) for an overview of asymptotic statistical decision theory and minimax decision rules. A subtle point is that the minimax bounds and statistical decision theory in this and preceding references are for samples of fixed size n, hence the asymptotic rates associated with $n \to \infty$, whereas adaptive allocation in multi-armed bandits is a sequential decision problem as we have already reviewed in Section 1. A key to bridge the differences between the fixed-sample and sequential theories is provided by Kim and Lai (2019). It is summarized in Section 2.2 that describes the sequential Arm Elimination procedure as follows: Choose $n_i \sim a^i$ for some integer $a \geqslant 1$, let $n_{j,t-1} = T_{t-1}(j)$ and eliminiate surviving arm j at time $t \in \{n_{i-1}+1, \ldots, n_i\}$ if (2.3) holds, in which $\Delta_{j,t-1}$ is the GLR statistic (2.5). This idea actually dates back to Lai (1987, p.1100-1103) in the proof of his theorem that the Bayes risk of UCB rules (with respect to general prior distributions H on $\boldsymbol{\theta}$) satisfies

(1.4). For contextual parametric bandits, H is a distribution on the covariate space (instead of a prior distribution on $\boldsymbol{\theta}$), and Kim and Lai (2019) basically modifies the aforementioned argument of Lai (1987) to derive a similar result.

Proof of Theorem 2. Consider the regret (2.1) over $B \subset \text{supp}H$ as the risk function of the statistical decision problem of sequential selection of k given arms as mentioned in the preceding paragraph, in which it is pointed out that $n_i \sim a^i$ plays the role of the fixed sample size in the asymptotic minimax rates for local linear regression estimates of $\mu_j(\cdot)$. We first explain the choice $\delta_t^2 = (2\log t)/t$ and why it is "particularly effective in the vicinity of leading arm transitions", as mentioned in the paragraph on the regularity conditions for Theorem 2. Note that (2.2) lumps treatments whose effect sizes are close to that of the apparent leader into a single set J_t of leading arms $j \in J_t$ for which $\tilde{\mu}_{j,t-1}(\cdot) = \hat{\mu}_{j,t-1}(\cdot)$ (and therefore $\Delta_{j,t-1} = 0$ in view of (2.4)). Such lumping is particularly important when the covariates are near leading arm transitions at which a leading arm can transition to an inferior one due to transitions in the covariate values. Because of the stated regularity conditions, the transition does not change its states as a member of the set of leading arms so that the ϵ-greedy randomization algorithm still chooses it with probability $(1-\epsilon)/|J_t|$. For parametric contextual bandits,

Kim and Lai (2019) choose $n_i \sim a^i$ for some integer $a > 1$ and consider $n_{i-1} < t \leqslant n_i$. For $j \in K_t$, $\hat{\theta}_{j,t-1}$ and $\tilde{\theta}_{j,t-1}$ are based on samples of size n_i. Combining this with the expcted time for elimination of arm $j \in K_t \setminus J_t$ shows that the parametric version of ϕ_{opt} (with (2.5) replacing (2.4)) attains the asymptotic lower bounds in Theorem 1(i), (ii). As pointed out in the preceding paragraph, the details of the proof basically modifty those of Lai (1987, p.1100–1103).

Nonparametric contextual bandits are much more difficult because the sample size of the local linear regression estimate $(\hat{\mu}_{j,t-1}(\cdot) - \tilde{\mu}_{j,t-1}(\cdot))_+$ for $n_{i-1} < t \leqslant n_i$ and $j \in K_t$ is of the order $n_i^{4/5}$ if the selected bandwidth has order $n_i^{-1/5}$ for univariate covariates as in Fan (1993), or if $b_{n_i}^1 \approx \cdots \approx b_{n_i}^p \approx n_i^{-1/5}$ for multivariate covariates with bandwidth matrix $\boldsymbol{B}_{n_i} = \text{diag}(b_{n_i}^1, \cdots, b_{n_i}^p)$ as in Ruppert and Wand (1994). It is not possible to obtain precise lower bounds of the type in Theorem 1(i) and (ii) and to attain these bounds using ϕ_{opt} (with (2.5) instead of (2.4)). Instead of the p-dimensional parametric family considered by Kim and Lai (2019), we use a cubic spline with evenly spaced knots (with the bandwidth as the spacing) in the univariate case and tensor product of these univariate splines for multivariate covariates. Details are give in the next paragraph. In conjunction with this parametric choice of $m(\boldsymbol{x})$, we also use the true

density function of $(y - m(\boldsymbol{x}))/\sigma(\boldsymbol{x})$ (Ruppert and Wand, 1994, p.1347) to define a parametric subfamily. It will be shown in the next paragraph that the minimax risk, under this parametric subfamily, of sequential selection of k arms up to time horizon n is of order $n^{4/5}$ and that ϕ_{opt} has minimax risk of order $n^{4/5} + o(1)$ under the regularity conditions of Theorem 2. This proves that the parametric subfamily is least favorable and that ϕ_{opt} attains the minimax rate of the risk function for adaptive allocation rules.

Minimax risk is the minimum (over all adaptive allocation rules) of the worse-case (or maximum) risk over Borel subsets B of $\text{supp}H$, which occurs around leading arm transitions. For the parametric subfamily in Theorem 1, the minimax risk is of order $(\log n)^2$ and is attained by ϕ_{opt} with (2.5) replacing (2.4). For the parametric subfamily in the preceding paragraph, because the spacing between the knots of the cubic spline for the regression function is of order $n^{-1/5}$, a straightforward modification of the argument in the proof of Theorem 1(ii) can be used to show that the minimax risk is of order $n^{4/5}$. Moreover, combining this argument with those of Fan (1993) and Ruppert and Wand (1994) shows that ϕ_{opt} has minimax risk of order $n^{4/5+o(1)}$ under the regularity conditions (a), (b), and (c) listed above.

S2 Personalization revolution and nonparametric contextual bandits

S3 Information-theoretic minimax rates and machine learning for applications in Big Data Era

Birgé and Massart (1993) and Shen and Wong (1994) have derived convergence rates of minimum contrast estimators and sieve MLE or other sieve estimators obtained by optimizing some empirical criteria. As noted by Shen and Wong (1994, p.581), the rate derived has not been proved to be optimal "although it coincides with the known optimal rate in several special cases of density estimation and nonparametric regression." Yang and Barron (1999) subsequently proved general results to determine minimax rates for the risk in density estimation using global measures of loss such as integrated squared error, squared Hellinger distance or Kullback–Leibler divergence, by applying information theory such as Fano's inequality; see Yu (1996), Cover and Thomas (2006, p.38–40, 146–153). The problem of minimax rates for the risk in nonparametric regression, however, is much more difficult than density estimation, and was solved by Yang and Tokdar (2015) that we review in the next paragraph.

To estimate the regression function $\mu(\cdot)$ nonparametrically from the regression model

$$y_t = \beta + \mu(\boldsymbol{x}_t) + \epsilon_t,\ 1 \leqslant t \leqslant n, \tag{S4}$$

in which ϵ_t are i.i.d. with mean 0 and variance σ^2 and are independent of the i.i.d. $\boldsymbol{x}_t \in \mathbb{R}^p$ with $p = p_n$ such that $\mathbb{E}\mu(\boldsymbol{x}_t) = 0$, Yang and Tokdar (2015, p.653, 657) make the following assumption M3 on the regression function $\mu(\cdot)$ and assumption Q on the common distribution H of the $\boldsymbol{x}_t$.

Assumption M3: $\mu \in L_2(H)$ depends on $d \approx \min(n^\gamma, p_n)$ variables for some $0 < \gamma < 1$ and is generated from a generalized additive model (Hastie and Tibshirani, 1986) such that the ℓth summand in the additive representation of $\mu(\cdot)$ depends on a small number d_ℓ of these variables, precise details of which will be stated using the notation of the next paragraph.

Assumption Q: H is compactly supported, hence it can be assumed without loss of generality that $\text{supp}H \subset [0,1]^p$. Moreover, H is absolutely continuous with respect to Lebesgue measure on $[0,1]^p$ with density function h such that $\bar{q} := \sup_{\boldsymbol{x}} h(\boldsymbol{x}) < \infty$ and there exist $\underline{q} > 0$ and $\delta > 0$ such that $\inf_{\boldsymbol{x}:|x_i - 1/2| \leqslant \delta, \forall i} h(\boldsymbol{x}) \geqslant \underline{q}$.

To state their main result under these assumptions, they have introduced the following notation in their Section 2. Let $C^{\alpha,d}$ denote the Banach

space of Hőlder α-smooth functions f on $[0,1]^d$ with the norm

$$||f||_\alpha = \sum_{a \leqslant \alpha} ||D^a f||_\infty + \max_{\boldsymbol{x} \neq \boldsymbol{y} \in [0,1]^d} \Big| D^{\lfloor \alpha \rfloor} f(\boldsymbol{x}) - D^{\lfloor \alpha \rfloor} f(\boldsymbol{y}) \Big| \Big/ ||\boldsymbol{x} - \boldsymbol{y}||^{\alpha - \lfloor \alpha \rfloor},$$

where $D^a = \partial^a / \partial x_1^{a_1} \dots \partial x_p^{a_p}$ for $a = a_1 + \dots a_p$ such that each a_i is a nonnegative integer. Let $C_1^{\alpha,d}$ denote the unit ball of $C^{\alpha,d}$. For $b = b_1 + \dots + b_p$ such that $b_i \in \{0,1\}$ for $1 \leqslant i \leqslant p$, define $T^b : C(\mathbb{R}^b) \to C(\mathbb{R}^p)$ by $(f(x_i), b_i = 1) \mapsto (T^b f)(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^p$, and let

$$\Sigma_p(\lambda, \alpha, d) = \Bigg(\bigcup_{b_i \in \{0,1\} : b_1 + \dots + b_p = d} T^b \Big(\lambda C_1^{\alpha,d} \Big) \Bigg) \bigcap \left\{ f \in C([0,1])^p) : \int f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0 \right\}$$

be the space of centered elements of $C([0,1]^p)$ that are α-smooth functions with sparsity d and bound λ. With this notation, Yang and Tokdar (2015, p.655) define the sparse additive representation of μ in Assumption M3 as $\mu = \sum_{\ell=1}^L \lambda_\ell T^{b^\ell} f_\ell$, where $f_\ell \in C_1^{\alpha_\ell, d_\ell}$ and $b^1, \dots, b^L \in \{0,1\}$ such that $b^1 + \dots + b^L \leqslant \bar{d}$. Their Theorem 3.1 states that there exist $0 < c_1 < 1 < c_2$ and positive integer n_0, all depending on $\bar{d}, \max_{1 \leqslant \ell \leqslant L} \lambda_\ell, \min_{1 \leqslant \ell \leqslant L} \lambda_\ell$, $\max_\ell \alpha_\ell, \min_\ell \alpha_\ell, \max_\ell d_\ell$ such that

$$c_1 \underline{\epsilon}_n^2 \leqslant \inf_{\hat{\mu} \in A_n} \sup_{\mu \in \Sigma_{p,L}^{\bar{d}}(\lambda, \alpha, d)} \mathbb{E}_{\beta,\sigma,H} ||\hat{\mu} - \mu|| \leqslant c_2 \bar{\epsilon}_n^2, \text{ where}$$

$$\underline{\epsilon}_n^2 = \sum_{\ell=1}^L \lambda_\ell^2 \big(\sqrt{n} \lambda_\ell / \sigma \big)^{-4\alpha_\ell/(2\alpha_\ell + d_\ell)} + \frac{\sigma^2}{n} \Bigg(\sum_{\ell=1}^L d_\ell \Bigg) \log \Bigg(p \Big/ \sum_{\ell=1}^L d_\ell \Bigg), \tag{S5}$$

$$\bar{\epsilon}_n^2 = \sum_{\ell=1}^L \lambda_\ell^2 \big(\sqrt{n} \lambda_\ell / \sigma \big)^{-4\alpha_\ell/(2\alpha_\ell + d_\ell)} + \frac{\sigma^2}{n} \Bigg(\sum_{\ell=1}^L d_\ell \Bigg) \log \Bigg(p \Big/ \min_{1 \leqslant \ell \leqslant L} d_\ell \Bigg).$$

In (S5) A_n is "the space of all measurable mappings of data to $L_2(H)$",

$\mathbb{E}_{\beta,\sigma,H}$ denotes expectation under the model $\mathbb{E}\big(y_t|\boldsymbol{x}_t\big) = \beta, \mathrm{Var}\big(y_t|\boldsymbol{x}_t\big) = \sigma^2$ and $\boldsymbol{x}_t \sim H$, and $\Sigma^{\bar{d}}_{p,L}\big(\lambda, \alpha, d\big)$ consists of $\mu \in \Sigma_p(\lambda, \alpha, d)$ that satisfies the aforementioned sparse additive representation $\mu = \sum_{\ell=1}^{L} \lambda_\ell T^{b^\ell} f_\ell$.

Assumption M3 with the sparse additive representation "offers a platform to break away from (previously assumed and overly restrictive) sparsity conditions" in the literature, as have been assumed by Raskutti, Wainwright, Yu (2012) and others who are inspired by variable selection such as the Lasso and the Dantzig selector for high-dimensional sparse regression to assume that μ depends on a small subset of d predictors with $d \leqslant \min(n, p)$. This corresponds to the special case $L = 1 = \bar{d}$ in (S5), in which the second summand in $\underline{\epsilon}_n^2$ or $\bar{\epsilon}_n^2$ is "the typical risk associated with variable selection uncertainty" and the first summand is the "minimax risk of estimating a d-variate, α-smooth regression function when there is no parameter uncertainty"; see Remark 3.3 of Yang and Tokdar (2015, p.658) who point out the implication of (S5) that in this case "meaningful statistical learning is possible only when the true number of important predictors is much smaller than the total predictor count".

For the application to contextual nonparametric k-armed bandits with high-dimensional covariates, we choose $n_i \sim a^i$ for some integer $a > 1$ and use Yang and Tokdar's minimax-optimal nonparametric regression estimate

$\hat{\mu}_{j,t-1}(\cdot)$ (or the constrained estimate $\tilde{\mu}_{j,t-1}(\cdot)$) of $\mu_j(\cdot)$ for $n_{i-1} < t \leqslant n_i$ and $j = 1, \ldots, k$. Under assumptions Q on H and M3 on μ_j for $j = 1, \ldots, k$, with the sparse additive representation $\mu_j = \sum_{\ell=1}^{L} \lambda_\ell^j T^{b_j^\ell} f_\ell$, in which $b_j^1, \ldots, b_j^L \in \{0, 1\}$, λ_ℓ^j and β_j depend on j (where α, L and $\bar{d}$ can be assumed to be applicable to all k arms), it follows from (S5) that we still have the ingredients of the proof of Theorem 2 given in the last part of S1. Hence the argument used there for fixed p can be modified via (S5) to extend it to the case of high-dimensional covariates under assumptions M3 and Q.

Additional References

Audibert, J-Y and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35**, 608–633.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd edition. Wiley, Hoboken, NJ.

de la Peña, V. H., Lai, T. L. and Shao, Q-M (2009). *Self-normalized Processes: Limit Theory and Statistical Applications.* Springer-Verlag, Heidelberg-Berlin-New York.

Donoho, D., Liu, R. C. and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18**, 1416–1437.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statist. Sci.* **1**, 297–310.

Lai, T. L. and Ying, Z. (1992). Asymptotically efficient estimation in censored and truncated regression models. *Statistica Sinica* **2**, 17–46.

Perchet, V. and Rigollet, P. (2013). The multi-armed bandit problem with covariates. *Ann. Statist.* **41**, 693–721.

Pinsker, M. S. (1980). Optimal filtering of square-integrable signals in gaussian noise. *Probl. Peredachi Inf.* **16**, 52–68; *Problems Inform. Transmission* **16**, 120–133.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Machine Learning Res.* **13**, 389–427.

Rigollet, P. and Zeevi, A. (2010). Nonparametric bandits with covariates. In *Conference on Learning Theory Proceedings*, 54–66.

Strasser, H. (1985). *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. De Gruyter, Berlin-New York.

Yang, Y. and Zhu, D. (2002). Randomized Allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.* **30**, 100–121.

Yu, B. (1996). Assoud, Fano, and Le Cam. In *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam* (D. Pollard, E. Turgensen and G. Yang, eds.) 423–435. Sprnger, New York.