Statistica Sinica **31** (2021), 2275-2287
doi:https://doi.org/10.5705/ss.202020.0454

MULTI-ARMED BANDITS WITH COVARIATES: THEORY AND APPLICATIONS

Dong Woo Kim[1], Tze Leung Lai[2] and Huanzhong Xu[2]

[1]*Microsoft Corporation and* [2]*Stanford University*

Abstract: "Multi-armed bandits" were introduced 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 later 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, providing important insights and algorithms for diverse applications in many science, technology, engineering and mathematics fields. The turn of the millennium marked the onset of a "personalization revolution," from personalized medicine and online personalized advertising and recommender systems (e.g. Netflix's recommendations for movies and TV shows, Amazon's recommendations for products to purchase, and Microsoft's Matchbox recommender). This has required an extension of classical bandit theory to nonparametric contextual bandits, where "contextual" refers to the incorporation 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 the 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

Corresponding author: Huanzhong Xu, Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA. E-mail: xuhuanvc@stanford.edu.

the population with the smaller sample size at these designated times. Here 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 later Robbins developed 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 idea of 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 $\{1, \ldots, k\}$, such that the event $\{\phi_i = j\}$, for $j \in \{1, \ldots, k\}$, 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

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

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

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

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 $\theta^* = \theta_{j(\boldsymbol{\theta})}, j(\boldsymbol{\theta}) = \operatorname{argmax}_{1\leqslant j\leqslant k}\mu(\theta_j)$, and an adaptive allocation rule is called "uniformly good" if $R_n(\boldsymbol{\theta}) = o(n^a)$ for every $a > 0$ and $\boldsymbol{\theta} \in \Theta^k$. Using the duality between hypothesis testing and confidence intervals, they also developed "upper confidence bound" (UCB) rules to attain the asymptotic lower bound (1.2).

In the remainder of this section, we first summarize the dynamic programming approach to multi-armed bandits introduced by Bellman (1957), before pre-

senting an 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. Sections 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 the parametric setting to (parametric) contextual bandits that also incorporate covariate information in the definition of regret. Section 2 develops the methodology of nonparametric contextual bandits, and Section 3 describes its extension to high-dimensional covariates in the current big-data and multi-cloud era.

1.1. UCB rule and Gittins index: asymptotic theory

Bellman (1957) introduced the dynamic programming approach for the two-armed adaptive allocation problem considered by Robbins (1952), generalizing it to k arms, and calling it a "k-armed bandit problem." The name is derived from an imagined slot machine with k arms (levers), 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 to 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 a discounted version of this problem (thereby circumventing large horizon n), and showed that the k-dimensional stochastic optimization problem has an "index policy" (which does not suffer from the curse of dimensionality) as its solution. At stage t, pull the arm with the largest "dynamic allocation index" (DAI), which 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 nonstandard 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 using dynamic programming.

We next review Lai (1987), who (a) connects the 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 one, in which $\theta_2 = 0$ is known and θ_1 has a prior distribution with mean zero. 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 as $n\tilde{\tau}(h)$, where $\tilde{\tau}(h) = \inf\{t \in (0,1] : w(t) + h(t) \leqslant 0\}$. He then shows that the following UCB rule is asymptotically optimal solution: 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 UCB

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

($\inf \emptyset = \infty$), $I(\lambda, \theta)$ 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) also extends 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 UCB (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 $\int \cdots \int \mathbb{E}_{\boldsymbol{\theta}}\Big[\sum_{i=1}^{\infty} \beta^{i-1} y_i\Big] d\nu_1(\theta_1) \cdots d\nu_k(\theta_k)$, 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 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 UCB (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 $1 \leqslant j \leqslant k$, $\rho > 0$, and $\theta_j^* = \max_{i \neq j} \theta_i$.

1.2. Nonparametric extensions of classical bandit theory

Using large deviation bounds for sums of uniformly recurrent Markov chains or mixing stationary sequences, Lai and Yakowitz (1995) extended the logarithmic lower bound (1.2) for the regret and the UCB rule. As such, they attained the bound in 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) developed another nonparametric method to attain the logarithmic lower bound (1.2) for the regret. Instead of a UCB-type rule, they use the ϵ-greedy randomization algorithm in reinforcement learning, as proposed by Sutton and Barto (1998). Further theoretical background and implementation details of the algorithm are 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 with distribution H, such 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 zero 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 optimal for the Bayesian discounted infinite-horizon problem of minimizing $\mathbb{E}(\sum_{t=1}^{\infty} \beta_{t-1} y_t)$ 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 nonBayesian 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 zero and Π_1 to be normal with mean $x_t + \theta$ conditional on x_t. Analogously to (1.1),

they define the regret in this simple case as

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

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 can 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 studies 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 configuration with regret $O(1)$, and other configurations for which the regret is of order $\log n$. 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 independent and identically distributed with common distribution H. Let $\mathrm{supp}H$ 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 $\boldsymbol{x}$, $\mu(\theta, \boldsymbol{x}) = \int y f_\theta(y|\boldsymbol{x}) d\nu(y)$, and

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

where 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.

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

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

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 $\mu(\lambda, \boldsymbol{x}) = \mu(\theta', \boldsymbol{x})$.

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 nonconstant 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, which will be summarized in the last paragraph 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}, the regret of which has the same "minimax rate" (defined there) as that of the lower bound in Section 2.1. After providing an overview of statistical and computational tools for its implementation, Section 2.3 studies its performance and gives further discussion and recent literature on applications of nonparametric contextual bandits to personalized medicine and recommender systems.

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 extension 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 $\mathrm{supp}H$, 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, analogously to (1.8), we call ϕ "uniformly good" over B if $R_{n,\phi}(B) = o(n^a)$, for every $a > 0$. We show 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 the 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 reaches a threshold defined 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 UCB (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'). This is shown for contextual parametric bandits by Kim and Lai (2019, Sec. III), who propose using 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

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

where $\hat{\mu}_{j,s}(\cdot)$ is the regression estimate (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 is used to lump 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) introduced the theory of information bounds and minimax risk into nonparametric or semiparametric cases (which is parametric for the parameters of interest and contains 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) shows that a local polynomial regression has the 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 extended the local linear regression to multivariate regressors.

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

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

where $n_{j,s} = T_s(j)$, g is given in (1.3), and $\Delta_{j,t-1}$ is the square of Welch's Studentized t-statistic based on $\{(\boldsymbol{x}_\ell, y_\ell) : 1 \leqslant \ell \leqslant t-1\}$; that is,

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

where $a_t = \max(a, 0)$ 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 is denoted by ϕ_{opt}.

Note that (2.4) is the nonparametric analog of the GLR statistic for testing the null hypothesis $H_{j,\ell}$ that $\mu_j(\boldsymbol{x}_\ell)$ is not significantly below $\max_{i \in K_\ell} \mu_i(\boldsymbol{x}_\ell)$ if $j \in J_\ell$ for $\ell \leqslant t-1$, in parametric models described in Section 1.3, for which Kim and Lai (2019, Section III) replace (2.2) with

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

and (2.4) with

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

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, simulation study, and discussion

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

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 is given in S1 which also reports a simulation study of the performance of ϕ_{opt}. The importance of nonparametric contextual bandit methodology to precision medicine and drug development is discussed in the recent works of Sklar, Shih and Lavori (2020) and Lai, Sklar and Weissmueller (2020). Earlier, Lai, Choi and Tsang (2019) described its important role in recommender systems, online experimentation and precision health.

3. High-Dimensional Covariates and Concluding Remarks

The Supplementary Material S2 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 context, it also reviews the works of Birgé and Massart (1993), Shen and Wong (1994), Yang and Barron (1999), and Yang and Tokdar (2015) on the information-theoretic approach to minimax rates of convergence, which provides a powerful method for tackling high-dimensional covariates.

In conclusion, multi-armed bandits with "side information" or covariates, also called contextual multi-armed bandits, arise in many fields, 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 a comprehensive theory and derived new results on nonparametric contextual bandits. These results are also generalized to high-dimensional covariates, which are of particular interest in the current big-data and multi-cloud era.

Supplementary Material

The online Supplementary Material contains a simulation study of the performance of ϕ_{opt} in the setting of $k = 6$ arms, the proof of Theorem 2 and related background literature (S1), information-theoretic minimax rates and machine learning for applications in the era of big data (S2), and additional references.

Acknowledgments

Lai's research was 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, Princeton.

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. *Progress in Statistics* (Edited by J. Gani), 241–266 . Elsevier, Amsterdam.

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. Technical Report. Department of Statistics, Stanford University.

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

Lai, T. L., Choi, A. and Tsang, K. W. (2019). Statistical science in information technology and precision medicine. *Ann. Math. Sci. & Appl.* **4**, 413–438.

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

Lai, T. L., Sklar, M. B. and Weissmueller, N. T. (2020). Novel clinical trial designs and statistical methods in the era of precision medicine. *Statist. Biopharm. Res.*. DOI: 10.1080/19466315.2020.1814403.

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

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.

Sklar, M. B., Shih, M. C. and Lavori, P. W. (2020). Bandit theory: Applications to learning healthcare systems and clinical trials. Technical Report. Department of Statistics, Stanford University.

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, Cambridge.

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.

Dong Woo Kim

Analysis and Experimentation Team, Microsoft Corporation, Redmond, WA 98052, USA.

E-mail: dongwookim80@gmail.com

Tze Leung Lai

Department of Statistics, Stanford University, Stanford, CA 94305, USA.

E-mail: lait@stanford.edu

Huanzhong Xu

Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA.

E-mail: xuhuanvc@stanford.edu

(Received November 2020; accepted November 2020)