MULTI-ARMED BANDITS WITH COVARIATES:

THEORY AND APPLICATIONS

Dong Woo Kim, Tze Leung Lai, and Huanzhong Xu

Microsoft Corporation and Stanford University

Supplementary Material

S1 Background literature, proof of Theorem 2, and simulation study

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 ingredient 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

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

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

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

and the $p \times p$ bandwidth matrix $\boldsymbol{B}_n = \operatorname{diag}(b_n^1, \ldots, b_n^p)$ 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 $\boldsymbol{W}_n(\boldsymbol{x}) = \operatorname{diag}\big(K_n(\boldsymbol{x}_1 - \boldsymbol{x}), \ldots, K_n(\boldsymbol{x}_n - \boldsymbol{x})\big)$, 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 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 eliminate 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 $B \subset \mathrm{supp} H$ 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 status 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 expected 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 $b^1_{n_i} \approx \cdots \approx b^p_{n_i} \approx n_i^{-1/5}$ for multivariate covariates with bandwidth matrix $\boldsymbol{B}_{n_i} = \mathrm{diag}(b^1_{n_i}, \cdots, b^p_{n_i})$ 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 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.

We next report a simulation study of the performance of ϕ_{opt} in the setting of $k = 6$ arms and univariate covariates x_t that are i.i.d. with common uniform distribution Unif$(-2, 2)$. Given x_t, the reward of arm j follows a normal distribution with mean $\mu_j(x_t) = \sin\left(x_t + \frac{j}{6}\pi\right)$ and standard deviation 0.1; Figure 1 plots the six mean reward functions and shows the locations of leading arm transitions. Figure 2 plots $\mathbb{E}\hat{\mu}_{j,n_i}(\cdot)$ of the local linear regression estimate $\hat{\mu}_{j,n_i}(\cdot)$ (details of which are given in the last paragraph of S1) at times $n_1 = 1000, n_2 = 3000$, and $n_3 = 30000$.

Figure 2 shows the limitation of minimax-rate results for nonparametric contextual bandits. As already noted in Section 1.3, and in particular Theorem 1 (see also S1), the statistical problem of sequential selection from k arms can have risks $O(\log n)$ over certain subsets B of supp$H$, while still attaining the minimax rate of the risk in the vicinity of leading arm transitions. This was first pointed out by Robbins for “asymptotically subminimax” decision rules in the context of compound statistical decision problems. Subsequently Hannan and Robbins (1955) related this to an empirical Bayes approach, which Robbins, Stein, and Efron later developed

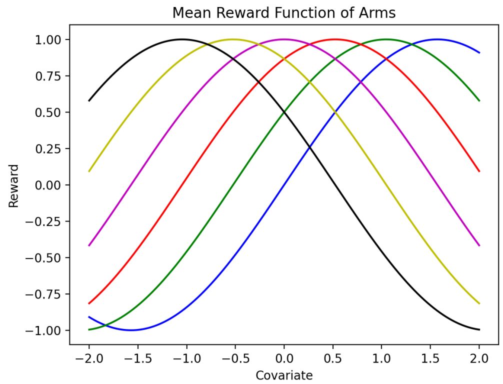

Figure 1: Mean Reward Function of Six Arms

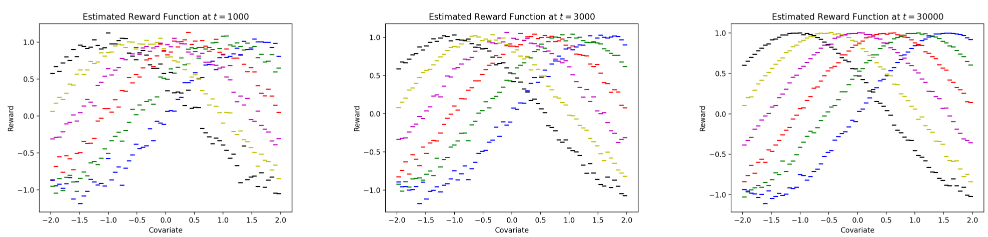

Figure 2: Means of Estimated Reward Functions

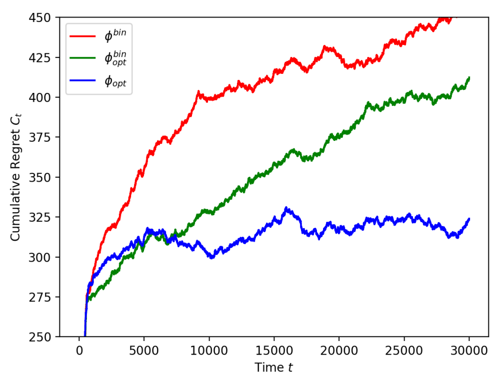

Figure 3: Cumulative Regret for $\phi_{opt}, \phi_{opt}^{bin}$, and ϕ^{bin}

into a foundational methodology for statistical analysis and modeling. Figure 3 compares the cumulative regret $C_{t,\phi} := \int R_{t,\phi}(x)dH(x)$ over time for $\phi = \phi_{opt}$ with that of $\phi = \phi^{bin}$ or ϕ^{bin}_{opt} defined below, where $R_{t,\phi}(x)$ is the Radon–Nikodym derivative of the measure (2.1) (where n is replaced with t) with respect to H. Yang and Zhu (2002) and Rigollet and Zeevi (2010) used "binned" regression estimates (i.e., piecewise constant functions or local polynomials of degree zero) to estimate $\mu_j(\cdot)$, and their procedure does not involve arm elimination. Moreover, although Rigollet and Zeevi still used the UCB rule, Yang and Zhu used ϵ-greedy randomization, which is what we refer to as ϕ^{bin}. Replacing the local linear regression in ϕ_{opt} with a binned (piecewise constant) regression leads to the procedure ϕ^{bin}_{opt}. Figure 3 , which plots the cumulative regret $C_{t,\phi}$ for $\phi = \phi_{opt}$ (blue), ϕ^{bin}_{opt} (green), and ϕ^{bin} (red), shows great improvement of ϕ_{opt} over ϕ^{bin}_{opt}, which is, in turn, a marked improvement over ϕ^{bin}.

S2 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

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

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 $b^1 + \cdots + b^L \leqslant \bar{d}$. 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

$$
\begin{aligned}
&c_1 \underline{\epsilon}_n^2 \leqslant \inf_{\hat{\mu} \in A_n} \sup_{\mu \in \Sigma_{p,L}^{\bar{d}}(\lambda,\alpha,d)} \mathbb{E}_{\beta,\sigma,H} ||\hat{\mu} - \mu|| \leqslant c_2 \bar{\epsilon}_n^2, \text{ where} \\
&\underline{\epsilon}_n^2 = \sum_{\ell=1}^{L} \lambda_\ell^2 \big(\sqrt{n}\lambda_\ell/\sigma\big)^{-4\alpha_\ell/(2\alpha_\ell+d_\ell)} + \frac{\sigma^2}{n}\left(\sum_{\ell=1}^{L} d_\ell\right) \log\left(p \Big/ \sum_{\ell=1}^{L} d_\ell\right), \qquad \text{(S5)} \\
&\bar{\epsilon}_n^2 = \sum_{\ell=1}^{L} \lambda_\ell^2 \big(\sqrt{n}\lambda_\ell/\sigma\big)^{-4\alpha_\ell/(2\alpha_\ell+d_\ell)} + \frac{\sigma^2}{n}\left(\sum_{\ell=1}^{L} d_\ell\right) \log\left(p \Big/ \min_{1\leqslant \ell \leqslant L} d_\ell\right).
\end{aligned}
$$

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 $\Sigma_{p,L}^{\bar{d}}\big(\lambda,\alpha,d\big)$ 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 (whereas 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.

The past five years have witnessed major advances in machine learning methods that facilitate the implementation of personalized prediction and recommender systems which make use of high-dimensional covariate information. In particular, personalized information filtering developed by Zhu, Shen and Ye (2016) uses a "likelihood method to seek a sparsest latent factorization (of a user-over-item preference matrix into two matrices, each representing a user's preference and an item preference by users) from a class of overcomplete factorizations, possibly with a high percentage of missing values", thereby providing "additional sparsity beyond rank reduction." Computationally, because the method involves a "decomposition and combination strategy" that breaks large-scale optimization "into many small subproblems to solve in a recursive and parallel manner", it can be implemented "through multi-platform shared-memory parallel programming, and through Mahout, a library for scalable machine learning and data mining, for mapReduce computation." The method is shown through theoretical and numerical investigations to be a "significant improvement over state-of-the-art methods" such as collaborative filtering and content-based filtering. An alternative method, subsequently developed by Bi, Qu and Shen (2018), uses a multilayer tensor to integrate information from multiple sources as in "context-aware recommender systems" (CARS) that incorporate the effect of contextual variables (such as time, location, users' companions, stores' promotion strategies in business marketing), with "an additional layer of nested latent factors to accommodate between-subjects dependency", thereby addressing the "cold-start issue in the absence of information from new customers, new products or new contexts" through subgroup information. A scalable algorithm is also developed to carry out the computations by "incorporating a maximum block improvement strategy into a cyclic blockwise-coordinate-descent procedure." Subsequent modifications and enhancements were developed by Dai et al. (2019, 2020).

Additional References

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

Bi, X., Qu, A. and Shen, X. (2018). Multilayer tensor factorization with applications to recommender systems. *Ann. Statist.* **46**, 3308–3333.

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

Dai, B., Shen, X., Wang, J. and Qu, A. (2020). Scalable collaborative ranking for personalized prediction. *J. Amer. Statist. Assoc.* **114**, 1–9.

Dai, B., Wang, J., Shen, X. and Qu, A. (2019). Smooth neighborhood recommender systems.

J. Machine Learning Res. **20**, 589–612.

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.

Hannan, J. F. and Robbins, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Statist.* **26**, 37–51.

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. Springer, New York.

Zhu, Y., Shen, X., and Ye, C. (2016). Personalized prediction and sparsity pursuit in latent factor models. *J. Amer. Statist. Assoc.* **109**, 1683–1696.