Multi-armed Bandits with Covariates: Theory and Applications

Tze Leung Lai

Stanford University

Outline

Overview of context-free multi-armed bandit theory

Contextual bandits in personalized medicine and marketing

- Web-based recommender systems and marketing
- Biomarker-guided strategies

Theory of multi-armed bandits with covariates

- Parametric contextual bandit theory
- Nonparametric theory

Conclusion: Data science, statistical/machine learning

Classical Multi-armed Bandit Theory

The k-arm bandit problem, introduced by Robbins (1952) for $k = 2$, is prototypical in the area of stochastic adaptive control that addresses the dilemma between "exploration" and "exploitation"

- Exploration / Information: to generate information about unknown system parameters
- Exploitation / Control: to set system inputs that attempt to maximize expected rewards from the outputs

Lai & Robbins (1985) , Lai (1987), and Chang & Lai (1987) introduced the regret to measure performance

Maximizing the reward is equivalent to minimizing the regret

- Asymptotic lower bound for regret
- Upper confidence bound (UCB) rule to attain asymptotic lower bound: UCB asymptotically equivalent to Gittins index

Regret and Asymptotic Lower Bound

- $S_n = y_1 + \cdots + y_n$; $y_i \mid \phi_i = j \sim f(\cdot;\theta_j)$; $\mu(\theta) = E_\theta y$
- Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$ and $\mathcal{F}_t$ =information set (σ-algebra) up to time t. Then for any adaptive allocation rule $\phi = (\phi_1, \cdots \phi_N)$ ($\{\phi_i = j\} \in \mathcal{F}_{i-1}$),

$$E_{\boldsymbol{\theta}} S_N = \sum_{i=1}^{N} \sum_{j=1}^{k} E_{\boldsymbol{\theta}} \left\{ E_{\boldsymbol{\theta}} \left(y_i I_{\{\phi_i = j\}} \mid \mathcal{F}_{i-1} \right) \right\} = \sum_{j=1}^{k} \mu(\theta_j) E_{\boldsymbol{\theta}} T_N(j),$$

where $T_N(j) = \sum_{i=1}^{N} I_{\{\phi_i = j\}}$. Hence maximizing $E_{\boldsymbol{\theta}} S_N$ is equivalent to minimizing

$$\begin{aligned} R_N(\boldsymbol{\theta}) &= N\mu(\theta^*) - E_{\boldsymbol{\theta}} S_N = \sum_{j:\mu(\theta_j) < \mu^*(\boldsymbol{\theta})} \{\mu(\theta^*) - \mu(\theta_j)\} E_{\boldsymbol{\theta}} T_N(j) \\ &\geq (1 + o(1)) \sum_{j:\mu(\theta_j) < \mu^*(\boldsymbol{\theta})} \{\mu(\theta^*) - \mu(\theta_j)\} \frac{\log N}{I(\theta_j, \theta^*)}, \end{aligned}$$

for uniformly good rules ($R_N(\boldsymbol{\theta}) = o(N^a)$ for every $\boldsymbol{\theta} \in \Theta^k$ and $a > 0$), where $\theta^* = \theta_{j(\boldsymbol{\theta})}$, $j(\boldsymbol{\theta}) = \arg\max_j \mu(\theta_j)$ and $I(\theta, \lambda)$ is the KL information #.

Bayesian Approach, Dynamic Programming, and Gittins Index

- Independent prior G_j on θ_j; infinite-horizon problem of maximizing

$$\int \cdots \int E_{\boldsymbol{\theta}} \left(\sum_{i=1}^{\infty} \beta^{i-1} y_i \right) dG_1(\theta_1) \cdots dG_k(\theta_k) = E \left(\sum_{i=1}^{\infty} \beta^{i-1} y_i \right).$$

- Gittins (1979) and Whittle (1981) used Markovian DP to derive the Bayes rule for this problem, which is the index rule ϕ^* that samples at stage $n+1$ from the population Π_{j^*} that has the largest (Gittins) index $M\left(G_{j^*|T_N(j^*)}\right)$ over the posterior distributions $G_{j|T_n(j)}$.
- The Gittins index $M(G)$ of a distribution G is the inf of solutions M of

$$\sup_{\tau \geq 0} E \left\{ \sum_{i=0}^{\tau-1} \beta^i E\left\{\mu(\theta) \mid y_1, \cdots, y_i\right\} + M \sum_{i=\tau}^{\infty} \beta^i \right\} = M \sum_{i=0}^{\infty} \beta^i.$$

- Although $M(G)$ may be difficult to compute, the index rule represents a major advance as it reduces a k-dimensional stochastic control problem to k optimal stopping problems.

Upper Confidence Bounds as Approximations to Gittins Indices

- Chernoff & Ray (1965) considered the finite-horizon one-armed bandit problem that chooses at each stage $n (\leq N)$ between sampling from a normal population Π_1 with unknown mean θ and known variance 1 and another population Π_2 with 0 reward. Assuming a normal prior on θ, the Bayes procedure samples from Π_1 until $T^* = \{n \leq N : \sum_{i=1}^{n} y_i + a_{n,N} \leq 0\}$. This is tantamount to sampling from Π_1 or Π_2 according as $U_n > 0$ or ≤ 0, where $U_n = \bar{y}_n + n^{-1} a_{n,N}$ (upper confidence bound for θ).
- Asymptotic expansion of the boundary h in $a_{n,N} \approx h(n/N)$ as $n/N \to 0$ agrees with that for the Gittins index $M_c(u, v) \approx u + \sqrt{c}\tilde{h}(v/c)$ as $\beta = e^{-c} \to 1$ and $v/c = t^{-1} \to \infty$ (Chang & Lai, 1987).

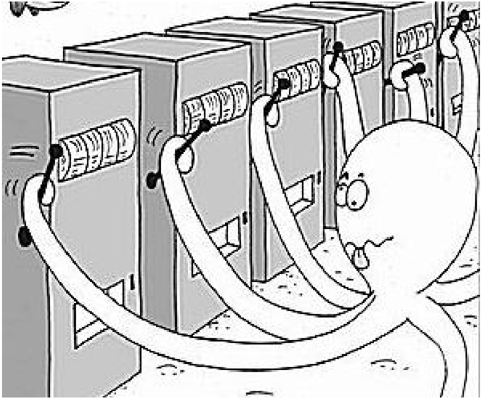
Asymptotically Minimal Regret of UCB Rules

- Lai (1987): Extension to exponential family $f(y;\theta) = e^{\theta y - \phi(\theta)}$: UCB for θ_j based on n observations from Π_j is

$$U_{j,n} = \inf\left\{\theta \geq \hat{\theta}_{j,n} : 2nI\left(\hat{\theta}_{j,n}, \theta\right) \geq g(n/N)\right\},$$

  where $g = h^2$, $\hat{\theta}_{j,n}$ is MLE and $I(\theta, \lambda)$ is KL information $\#$. This suggests the UCB rule: Sample as stage $n+1$ from the population that has the largest $U_{j,T_n(j)}$.
- The UCB rule attains the asymptotic lower bound for $R_N(\boldsymbol{\theta})$ for uniformly good rules, at every fixed $\boldsymbol{\theta}$, as $N \to \infty$.
- The UCB rule also attains asymptotically (as $N \to \infty$) the Bayes regret, which is of order $C(\log N)^2$, when the prior distribution for $\boldsymbol{\theta}$ has positive continuous density over $\theta_j \in \left(\theta_j^* - \rho, \theta_j^* + \rho\right)$ for $1 \leq j \leq k$, where $\theta_j^* = \max_{i \neq j} \theta_i$.

Multi-armed Bandits with Side (Covariate) Information



Web Source: Microsoft Research (Silicon Valley), MAB Team

Analysis & Experimentation Team (Bellevue, WA): Dong Woo Kim, Tong Xia, Alex Deng

Covariate Information and Personalization

- Let $\mu_j = \mu(\theta_j)$. The classical multi-armed bandit (MAB) problem aims at choosing ϕ_i sequentially so that $E_{\boldsymbol{\theta}}\left(\sum_{i=1}^{N} y_i\right)$ is as close as possible to $N \max_{1 \le j \le k} \mu_j$.
- Since the arms now also have covariate information $\mathbf{x}_i$ and $E_{\boldsymbol{\theta}}(y_i) = \sum_{j=1}^{k} E_{\boldsymbol{\theta}}\left\{E_{\boldsymbol{\theta}}\left(y_i I_{\{\phi_i = j\}} \mid \mathbf{x}_i\right)\right\} = \sum_{j=1}^{k} \mu_j(\mathbf{x}_i)$, where $\mu_j(\mathbf{x}) = \mu(\theta_j; \mathbf{x})$, the covariate (contextual) bandit problem replaces $N\mu_j$ by $\sum_{i=1}^{N} \mu_j(\mathbf{x}_i)$ in the classical MAB.
- The covariate information $\mathbf{x}_i$ for the ith subject is therefore used to "personalize" the treatment selection for the subject, as in personalized marketing or web-based recommender systems, or biomarker-guided therapies in personalized medicine.

Web-based Personalization in Marketing and Recommender Systems

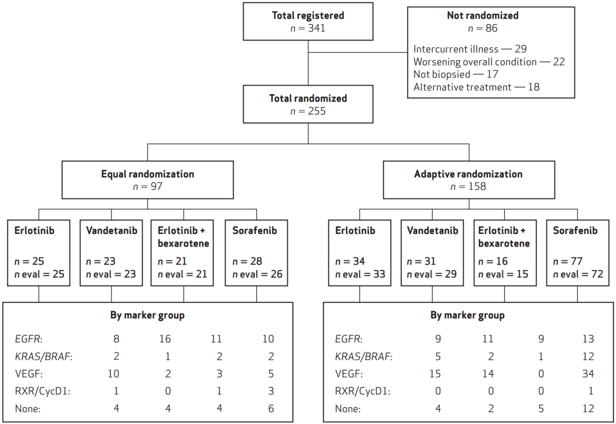
- Personalized marketing (also called one-to-one marketing) uses web sites to track a customer's interests and purchasing records and thereby to market products individualized for the customer, e.g., Amazon. Recommender systems select items such as movies (e.g. Netflix) and news (e.g. Yahoo) for users based on the users' and the items' features (covariates).
- Li, Chu, Langford & Schapire (2010) model the click probability of a news article as a function, estimated by machine learning methods, of the user's and article's features. They apply a UCB-type policy targeted towards maximizing the click probability, but no theoretical analysis or simulation study of the performance of the policy is given.
- Tang, Rosales, Singh & Agarwal (2013) consider web-based personalization in showing online ads for each user, with the goal of maximizing "its effectiveness, measured in terms of click-through rate or total revenue." They formulate the optimization problem as a contextual multi-armed bandit problem with the page request of each user as side (covariate) information and layouts of ads available for the requested page as arms.

Biomarker-guided Therapies in Personalized Medicine

- The development of imatinib (Gleevec), the first drug to target the genetic effects of chronic myeloid leukemia (CML) while leaving healthy cells unharmed, has revolutionized the treatment of cancer, leading to hundreds of kinase inhibitors and other targeted drugs that are in various stages of development in the anticancer drug pipeline. However, most new targeted treatments have resulted in only modest clinical benefit, with less than 50% remission rates and less than one year of progression-free survival.
- Trastuzumab (Herceptin), which treats only patients with HER-2 positive metastatic breast cancer, has better remission rate and longer progression-free survival because it targets the "right" patient population.
- Genome-guided targeted therapies like Herceptin are expected to substantially improve the effectiveness of cancer treatments, hence recent interest in their use for drug development and for comparative effectiveness research (CER) of approved treatments following the health care reform legislation in 2010.

The BATTLE Trial

Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE)



Total registered
$n = 341$

Not randomized
$n = 86$

Intercurrent illness — 29
Worsening overall condition — 22
Not biopsied — 17
Alternative treatment — 18

Total randomized
$n = 255$

Equal randomization
$n = 97$

Erlotinib $n = 25$, n eval = 25

Vandetanib $n = 23$, n eval = 23

Erlotinib + bexarotene $n = 21$, n eval = 21

Sorafenib $n = 28$, n eval = 26

By marker group

EGFR:	8	16	11	10
KRAS/BRAF:	2	1	2	2
VEGF:	10	2	3	5
RXR/CycD1:	1	0	1	3
None:	4	4	4	6

Adaptive randomization
$n = 158$

Erlotinib $n = 34$, n eval = 33

Vandetanib $n = 31$, n eval = 29

Erlotinib + bexarotene $n = 16$, n eval = 15

Sorafenib $n = 77$, n eval = 72

By marker group

EGFR:	9	11	9	13
KRAS/BRAF:	5	2	1	12
VEGF:	15	14	0	34
RXR/CycD1:	0	0	0	1
None:	4	2	5	12

Reference: Kim et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov 2011;1(1):44-53

An Alternative Group Sequential Design

- The BATTLE trial used (i) a hierarchical Bayesian probit model for the response and (ii) a corresponding Bayesian design that uses an adaptive randomization scheme to assign treatments, with randomization probabilities proportional to the posterior probabilities of disease control for the treatments and biomarker classes. The final analysis reported in the 2011 paper, however, uses conventional frequentist inference and ignores the random sample sizes due to adaptive randomization.
- Lai, Liao & Kim (2013) proposed an alternative group sequential design for clinical trials to develop and test biomarker-guided strategies. It uses an adaptive randomization method, with randomization probabilities determined at each interim analysis, and arm elimination based on generalized likelihood ratio statistics, with valid type I error probability to take account of early stopping and adaptive randomization.

Multiple Objectives & Frequentist Inference in New Design

The group sequential design addresses multiple objectives

1. Treat accrued patients with the best available treatment
2. Develop a treatment strategy for future patients
3. Demonstrate that the strategy developed indeed has better treatment effect than the Standard-Of-Care (SOC)

Lai, Liao & Kim showed that it has higher overall disease control rates (DCR) than the design used in the BATTLE study for patients in the trial, and maintains the prescribed type I error probability that the strategy falsely claims better DCR than SOC, and an overall probability guarantee that the best treatment is included in the recommended set of treatments for future patients in each biomarker class.

Works on Parametric Contextual Bandit Theory ($k = 1$)

- Woodroofe (1979) considered the one-armed covariate problem where $\mu_2 = 0$ is known and $\mu_1(x) = \theta + x$, with θ being normally distributed. Under some regularity conditions, he showed that the myopic policy is asymptotically optimal for maximizing $\sum_{t=1}^{\infty} \beta^{t-1} E_\theta y_t$.
- Sarkar (1991) extended Woodroofe's result to $y_{j,t} \sim F_{\theta_j}(\cdot \mid x_t)$ for $j \in \{1, 2\}$, where F_θ belongs to a one-parameter exponential family.
- Clayton (1989) considered the finite-horizon case and used dynamic programming to derive some properties of the optimal rule when $y_{j,t}$ is Bernoulli for $j \in \{1, 2\}$.
- Goldenshulger & Zeevi (2009) also considered the finite-horizon case, but in the minimax setting of minimizing the maximum regret $\sum_{t=1}^{N} E_\theta |x_t + \theta| I_{\{\phi_t \neq \phi_t^*\}}$ over θ, for Woodroofe's problem with $\mu_1(x) = \theta + x$, where $\phi_t^* = I_{\{x_t + \theta \geq 0\}}$ is the optimal (oracle) policy that assumes θ to be known. They showed that the minimax regret can be bounded on grow to ∞ at various rates with N, depending on the behavior of $G([-\theta - \delta, -\theta + \delta])$ as $\delta \to 0$.

A Paper on Parametric Contextual Bandit Theory ($k = 2$)

- Wang, Kulkarni & Poor (2005) considered a two-armed covariate bandit problem where, for $j \in \{1, 2\}$, $y_{j,t}(x_t) \sim F_{\theta_j}(\cdot \mid x_t)$, $x_t \in \mathcal{X}$, and $\theta_j \in \Theta$, with finite $\mathcal{X}$ and Θ being a subset of $\mathbb{R}$.
- A parameter configuration $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is said to be implicitly revealing if $\exists x^1, x^2 \in \mathcal{X}$ such that arms 1 and 2 are the best arms given $x_t = x^1$ and x^2, respectively.
- The paper developed an asymptotically optimal procedure in this very restrictive case, that does not even cover simple linear regression: Finite $\mathcal{X}$, $\Theta \subset \mathbb{R}$ univariate means that either the slope or the intercept is known.

General Parametric Contextual Bandit Theory: Regret

- $y_t \mid \{\phi_t = j, \mathbf{x}_t\} \sim f(\cdot; \theta_j, \mathbf{x}_t)$; $\mathbf{x}_t \overset{i.i.d.}{\sim} G$; $\mu(\theta, \mathbf{x}) = \int y f(y; \theta, \mathbf{x})\, dv(y)$.
- Define $j^*(\mathbf{x}) = \arg\max_{1 \le j \le k} \mu(\theta_j, \mathbf{x})$, $\theta^*(\mathbf{x}) = \theta_{j^*(\mathbf{x})}$, and for $B \subset \operatorname{supp}(G)$, define $R_N(\boldsymbol{\theta}, B)$ by

$$N \int_B \mu(\theta^*(\mathbf{x}), \mathbf{x})\, dG(\mathbf{x}) - \sum_{t=1}^{N} \sum_{j=1}^{k} E_{\boldsymbol{\theta}} \left\{ E_{\boldsymbol{\theta}} \left(y_t I_{\{\phi_{t=j}, \mathbf{x}_t \in B\}} \mid \mathcal{F}_{t-1} \right) \right\}$$

$$= \sum_{j=1}^{k} \int_B \{\mu(\theta^*(\mathbf{x}), \mathbf{x}) - \mu(\theta_j, \mathbf{x})\} \{E_{\boldsymbol{\theta}} T_N(j, \mathbf{x})\}\, dG(\mathbf{x}),$$

where $T_N(j, B) = \sum_{t=1}^{N} I_{\{\phi_{t=j}, \mathbf{x}_t \in B\}}$ and $\mathcal{F}_{t-1}$ is the σ-algebra generated by $\mathbf{x}_t$ and $(\mathbf{x}_s, y_s)$ for $s \le t-1$. Note that the measure $E_{\boldsymbol{\theta}} T_N(j, \cdot)$ is absolutely continuous with respect to G, and therefore we can define its Radon-Nikodym derivative d/dG, which we denote by $E_{\boldsymbol{\theta}} T_N(j, \mathbf{x})$.

Asymptotic Lower Bound for Regret

Extension of classical multi-armed bandit theory involves KL information numbers and an asymptotic lower bound that is attainable by making use of generalized likelihood ratio (GLR) statistics for testing a composite hypothesis. Let ϕ be uniformly good over $B \subset \mathcal{X}$. For $\theta, \theta^{'} \in \Theta$ and $\mathbf{x} \in \operatorname{supp} G$, define

$$I_{\mathbf{x}}\left(\theta, \theta^{'}\right) = \inf_{\lambda : \mu(\lambda, \mathbf{x}) = \mu\left(\theta', \mathbf{x}\right)} I\left(\theta, \lambda; \mathbf{x}\right); \; I\left(\theta, \lambda; \mathbf{x}\right) = E_{\theta}\left\{\log \frac{f\left(Y; \theta, \mathbf{x}\right)}{f\left(Y; \lambda, \mathbf{x}\right)}\right\}.$$

(a) If $j^*\left(\mathbf{x}\right) = \arg\max_{1 \leq j \leq k} \mu\left(\theta_j, \mathbf{x}\right)$ is non-constant over B (with leading arm transitions), then $R_N\left(\boldsymbol{\theta}, B\right) \geq C\left(\boldsymbol{\theta}\right)\left(\log N\right)^2$, assuming $\mu\left(\theta, \mathbf{x}\right)$ and $I\left(\theta, \lambda; \mathbf{x}\right)$ to be continuously differentiable in $\mathbf{x}$ belonging to neighborhoods of leading arm transitions.

(b) If j^* is constant over B, then $R_N\left(\boldsymbol{\theta}, B\right)$

$$\geq \left(1 + o\left(1\right)\right) \sum_{j : P\{\theta_j = \theta^*(\mathbf{X})\} = 0} \left(\log N\right) \int_B \frac{\mu\left(\theta^*\left(\mathbf{x}\right), \mathbf{x}\right) - \mu\left(\theta_j, \mathbf{x}\right)}{I_{\mathbf{x}}\left(\theta_j, \theta^*\left(\mathbf{x}\right)\right)} dG\left(\mathbf{x}\right),$$

taking $\sum_j$ over the empty set as $O\left(1\right)$.

Deferred Sampling from Inferior Arms: Adaptive Randomization

- The UCB rule (index policy) in classical bandit theory basically samples from an inferior arm until the sample size satisfies the information bound (asymptotic lower bound for $E_{\boldsymbol{\theta}} T_N(j)$). For covariate bandits, an arm that is inferior at $\mathbf{x}$ may be the best at another $\mathbf{x}^{'}$. Therefore the uncertainty in the sample mean at $\mathbf{x}_t$ does not need to be immediately reduced. A better way is to use adaptive randomization.
- Let K_t denote the set of arms to be tried at time t (rationale explained in next slide). Let $J_t = \left\{ j \in K_t : \left| u\left(\hat{\theta}_{j,t-1}, \mathbf{x}_t\right) - \mu\left(\hat{\theta}^*_{t-1}\left(\mathbf{x}_t\right), \mathbf{x}_t\right) \right| \leq \delta_t \right\}$, where $\hat{\cdot}_s$ denotes the MLE based on observations up to time s. At time t, choose treatments randomly with probabilities $\pi_{j,t} = \epsilon$ for $j \in K_t \backslash J_t$ and $\pi_{j,t} = \left(1 - \left|K_t \backslash J_t\right| \epsilon\right) / \left|J_t\right|$ for $j \in J_t$.

GLR Tests and Arm Elimination

- The UCB in Lai (1987) basically inverts a one-sided likelihood ratio test of $\theta_j = \theta^{'}$ based on observations from Π_j. We can likewise consider GLR tests of the composite null hypothesis $H_{j,t} : \mu\left(\theta_j, \mathbf{x}_t\right) \geq \max_{j' \neq j} \mu\left(\theta_j, \mathbf{x}_t\right)$. Rejection of the null hypothesis suggests that Π_j is significantly inferior to some other arms for the covariate $\mathbf{x}_t$. The GLR statistic for testing $H_{j,t}$ is

$$L_{j,t-1} = \sum_{i=1}^{t-1} I_{\{\phi_i = j\}} \log\left[f\left(y_i; \hat{\theta}_{j,t-1}, \mathbf{x}_t\right) / f\left(y_i; \tilde{\theta}_{j,t-1}, \mathbf{x}_t\right)\right],$$

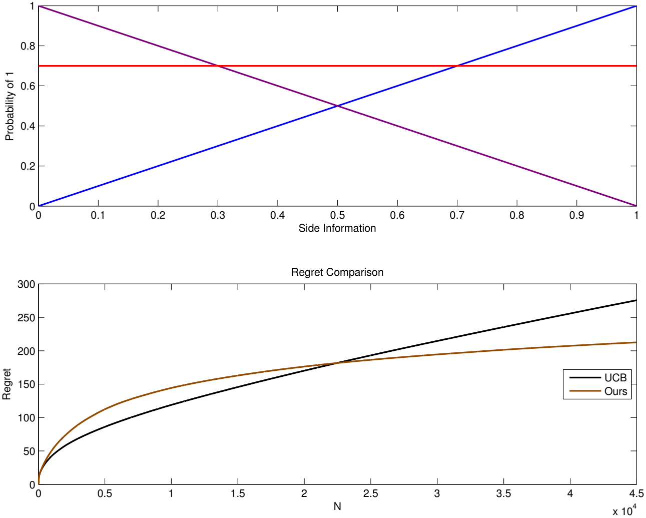
  where $\tilde{\theta}_{j,t-1}$ is the constrained MLE under the constraint $\mu\left(\theta_j, \mathbf{x}_t\right) = \max_{j' \neq j} \mu\left(\hat{\theta}_{j',t-1}, \mathbf{x}_t\right)$.
- Choose $N_i \sim a^i$ for some integer $a > 1$. For $N_{i-1} < t \leq N_i$, eliminate arm j if $\hat{\theta}_{j,t-1} < \hat{\theta}^*_{t-1}\left(\mathbf{x}_t\right)$ and $L_{j,t-1} > g\left(n_{j,t-1}/N_i\right)$, where $n_{j,s} = T_s\left(j\right)$. Thus K_t is the set of surviving arms at the beginning of stage t.

Asymptotic Efficiency of Covariate Bandit Policy

- The preceding bandit policy that uses adaptive randomization (AR) and GLR-based arm elimination attains the asymptotic lower bound for the regret of contextual bandits as $N \to \infty$.
- $I_{\mathbf{x}}\left(\theta, \theta'\right)$ in the asymptotic lower bound is related to the GLR test statistic for the composite hypothesis $H_{j,t}$ that tests non-inferiority of arm j for covariate $\mathbf{x}_t$ based on all observations up to time t.
- The adaptive randomization scheme that assigns most probability to the leading arm and some probability to the apparently inferior ones has been proposed in classical bandit theory under the name "ϵ-greedy algorithm".

A Simulation Example

Three Gaussian arms whose expected reward functions are shown on the top panel



Nonparametric Regression and Covariate Bandits

- The functions $\mu_j(\mathbf{x}) = \mu(\theta_j, \mathbf{x})$ in the preceding parametric theory are regression functions of y on $\mathbf{x}$, one for each arm Π_j. Instead of using a parametric model that involves the regression parameters θ_j, Yang & Zhu (2002) and Rigollet & Zeevi (2010) have used nonparametric regression to estimate μ_j.
- Yang & Zhu use an ϵ_t-greedy algorithm that samples from the leading arm having the largest estimated reward $\hat{\mu}_{j,t-1}(\mathbf{x}_t)$ with probability $1-(k-1)\epsilon_t$ and all other arms with probability ϵ_t. They use some nonparametric regression method to estimate μ_j, which they do not specify but require $\|\hat{\mu}_{j,n} - \mu_j\|_\infty$ to converge to 0 a.s. for every j, as $n \to \infty$.

Nonparametric UCB Rules for Covariate Bandits

Yang & Zhu(2002) have only shown that for their allocation rule ϕ,

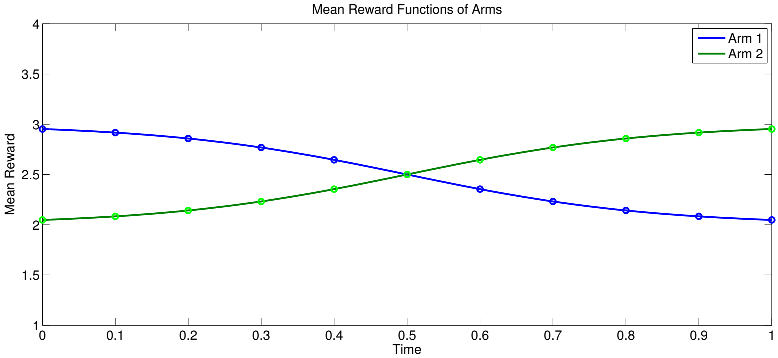
$$N^{-1}\sum_{t=1}^{N}\mu_{\phi_t}\left(\mathbf{x}_t\right)\to\int\mu^*\left(\mathbf{x}\right)dG\left(\mathbf{x}\right)\ \text{a.s., where }\mu^*\left(\mathbf{x}\right)=\max_{1\le j\le k}\mu_j\left(\mathbf{x}\right).$$

but do not have any result on the regret $R_N=\sum_{t=1}^{N}E\left\{\mu^*\left(\mathbf{x}_t\right)-\mu_{\phi_t}\left(\mathbf{x}_t\right)\right\}$. For the case $k=2$, Rigollet & Zeevi (2010) partition the covariate space, which they assume to be $[0,1]^d$, into small bins. The nonparametric regression method they use is the histogram method (also called binning or regressogram). Basically they apply the UCB $\bar{y}_{j,b(\mathbf{x}_t);t-1}+\left(\left(2\log t\right)/n_{j,b(\mathbf{x}_t);t-1}\right)^{1/2}$, where $b\left(\mathbf{x}\right)$ denotes the bin in which $\mathbf{x}$ falls. They basically reduce a covariate bandit problem to B classical bandit problems, where $B=B_n$ is the total number of bins, and thereby obtain bounds of the order $n^{1-\gamma}$ for some $0<\gamma<1$ under certain regularity conditions.

A New Approach to Nonparametric Covariate Bandits

- The bias-variance tradeoff in nonparametric bandit theory is different from that in nonparametric regression. We allow the bin size to decrease with time t but would also combine bins to use a linear approximation of the regression function instead of a step function (histogram) approximation in regions near the intersections of the mean reward functions of different leading arms.
- We basically follow the parametric approach and modify it with quasi-likelihood that formally assumes Gaussian noise so that sample means are the QML estimates. Local regression (first by binning and later by locally linear regression over combined bins) is used to model the unknown reward functions. Adaptive randomization is used in lieu of UCB, and arm elimination using quasi-likelihood is also used. It can be shown to be asymptotically efficient.

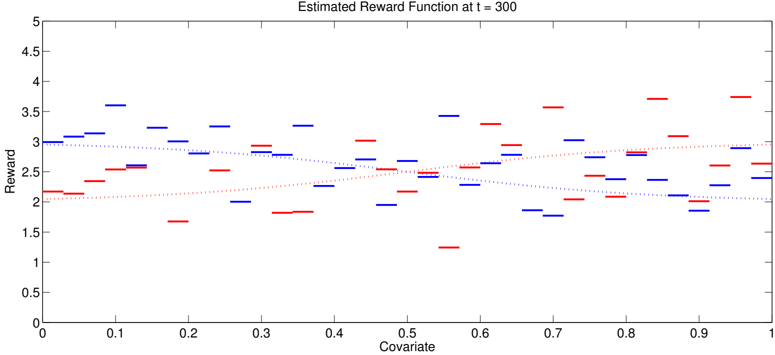
Example ($k = 2$ Mean Reward Functions)



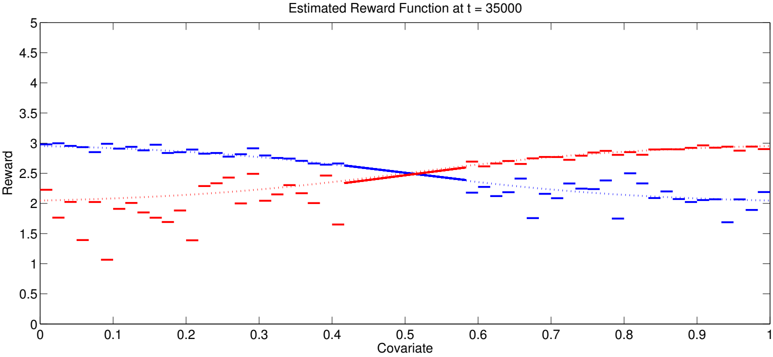
$X_t \sim U[0,1]$: standard normal errors in regression model

Horizon $N = 35,000$

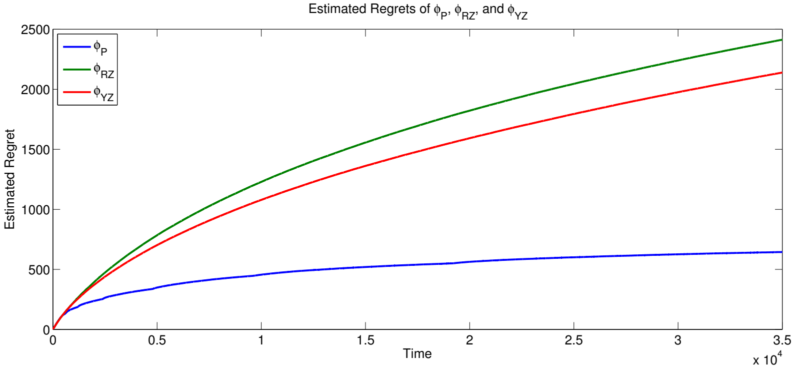
Estimated Mean Reward Functions at $t = 300$



Estimated Mean Reward Functions at $t = 35000$



Regret of Proposed Policy (ϕ_P), ϕ_{RZ} and ϕ_{YZ}



Conclusion

- MAB with side information (covariate/contextual bandits) arises in many fields of application, in which the development of personalized strategies or recommender systems requires both exploration and exploitation
- New definitive theory of ($K \geq 2$)-armed bandits with covariates can
  - provide theoretical support for previous experimental studies in personalized strategies and recommender systems
  - be used to develop new "learn-as-we-go" strategies
- By incorporating statistical/machine learning approaches, the covariate bandit theory can advance "Big Data" analytics for these applications. Implementation of the theory involves modern developments in statistical/machine learning and in data science.