Nonparametric Bandits with Covariates

Philippe Rigollet

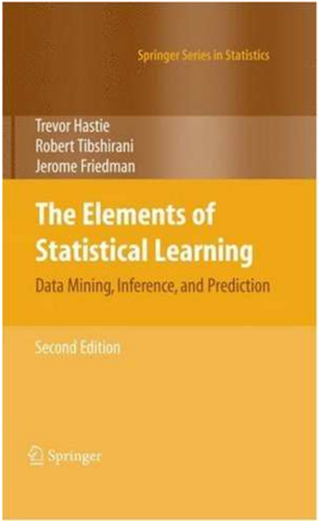
Princeton University

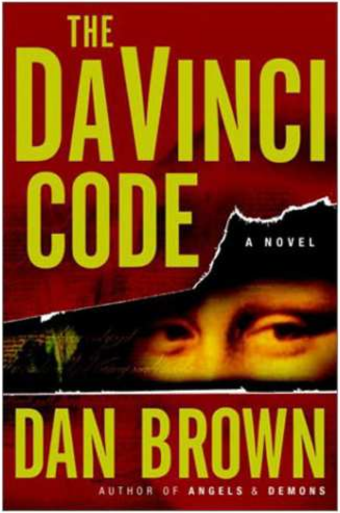
with A. Zeevi (Columbia University)

Support from NSF (DMS-0906424)

Example: Real time web page optimization

Example: Real time web page optimization





Example: Real time web page optimization

Which ad will generate the most \$/clicks ?

Characteristics of the problem

- A choice must be made for each customer.
- Cannot observe the outcome of the alternative choice.
- Try to maximize the rewards.

Exploration vs. Exploitation dilemma

Exploration: which one is the best?
Exploitation: display the best as much as possible.

Two armed bandit problem: setup

- Two **arms** (e.g.: actions, ads): $i \in \{1, 2\}$.
- At time t, random reward $Y_t^{(i)}$ is observed when arm i is pulled.
- A **policy** π is a sequence $\pi_1, \pi_2, \ldots \in \{1, 2\}$, which indicates which arm to pull at each time t.
- Performance: Expected cumulative reward at time n

$$\mathbb{E} \sum_{t=1}^{n} Y_t^{(\pi_t)}$$

- Goal: MAXIMIZE reward.

Two armed bandit problem: regret

- Oracle policy $\pi^\star = (\pi_1^\star, \pi_2^\star, \ldots)$ pulls at each time t the best arm (in expectation)

$$\pi_t^\star = \operatorname*{argmax}_{i=1,2} \mathbb{E}[Y_t^{(i)}]\,.$$

- We measure our performance by the **regret**

$$R_n(\pi) = \mathbb{E} \sum_{t=1}^{n} \big(Y_t^{(\pi_t^\star)} - Y_t^{(\pi_t)}\big)$$

Static Environment

- The problem is not new: Robbins ('52), Lai & Robbins ('85)

Static Environment

- The problem is not new: Robbins ('52), Lai & Robbins ('85)
- Key assumption:

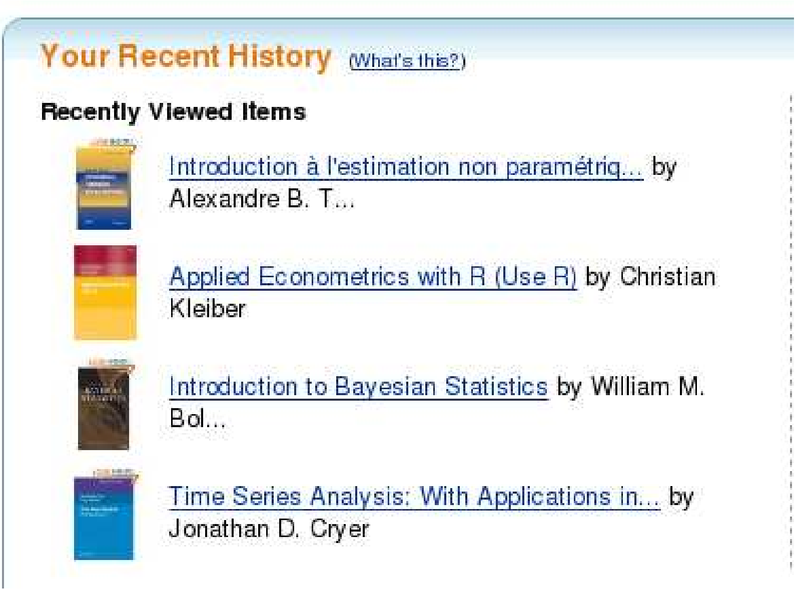
Static environment

- i.e., the (unknown) expected rewards $\mu_i = \mathbb{E}[Y_t^{(i)}]$ are **constant**.
- One way to solve the problem is to use

Upper **C**onfidence **B**ounds policy.

Side information

Side information



Side information

Side information and covariates

- At time t, the reward of each arm $i \in \{1,2\}$ depends on a **covariate** $X_t \in \mathcal{X}(\subset (\mathbb{R}^d))$

$$Y_t^{(i)} = f^{(i)}(X_t) + \varepsilon_t, \quad t = 1, 2, \dots, \qquad i = 1, 2\,.$$

  with standard regression assumptions on $\{\varepsilon_t\}$.
- A policy is now a **sequence of functions**

$$\pi_t \;:\; \mathcal{X} \to \{1, 2\}\,.$$

- Oracle policy

$$\pi^\star(x) = \operatorname*{argmax}_{i=1,2} \mathbb{E}[Y_t^{(i)} | X_t = x] = \operatorname*{argmax}_{i=1,2} f^{(i)}(x)$$

Assumptions on the expected rewards

Assume now that $\mathcal{X} = [0, 1]$.

1. **Constant:** Static model studies by Lai & Robbins:

$$f^{(i)}(x) = \mu_i,\ \ i = 1, 2 \qquad \mu_i \text{ unknown}$$

2. **Linear:** One-armed bandit problem, studied by Goldenshluger & Zeevi (2008)

$$f^{(1)}(x) = x - \theta,\ \ i = 1, 2 \qquad \theta \text{ unknown}$$

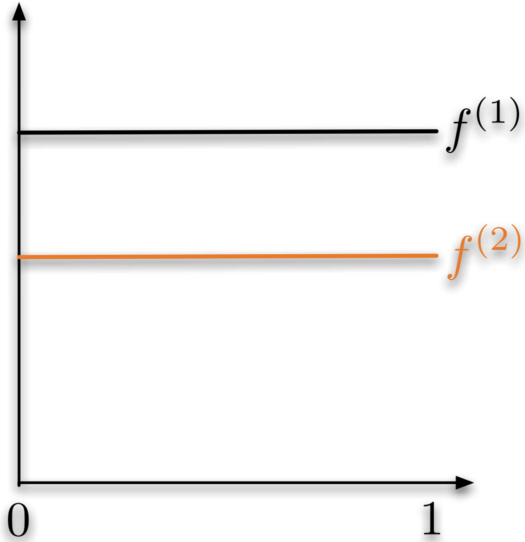
   and $f^{(2)}(x) = 0$ is constant and **known**.

3. **Smooth:** We assume that the functions are Hölder smooth with parameter $\beta \leq 1$:

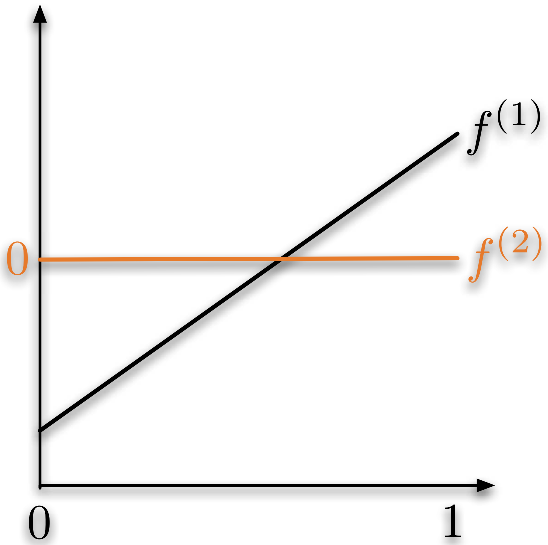
$$|f^{(i)}(x) - f^{(i)}(x')| \leq L|x - x'|^{\beta}\,.$$

   (Consistency studied by Yang & Zhu, 2002)

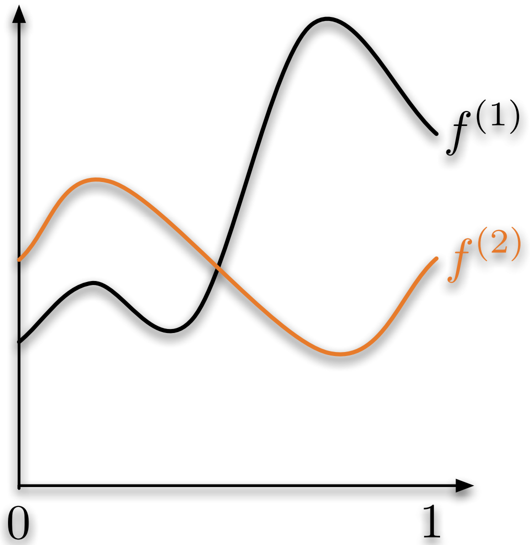
Constant rewards



One-armed linear reward



Smooth rewards



Nonparametric bandit with covariates

Two armed bandit problem with uniform covariates

- Covariates: $\{X_t\}$ i.i.d in $[0,1]$ with uniform distribution
- Rewards: $Y_t^{(i)} \in [0,1]$

$$\mathbb{E}\big[Y_t^{(i)} | X_t\big] = f^{(i)}(X_t) \quad t = 1, 2, \ldots, \; i = 1, 2\,,$$

where $|f^{(i)}(x) - f^{(i)}(x')| \le L|x - x'|^\beta,\; \beta \le 1,\; i = 1, 2$

- Oracle policy pulls at time t

$$\pi^\star(X_t) = \operatorname*{argmax}_{i=1,2} f^{(i)}(X_t)$$

- Regret

$$R_n(\pi) = \mathbb{E} \sum_{t=1}^{n} \big(f^{(\pi^\star(X_t))}(X_t) - f^{(\pi_t(X_t))}(X_t)\big)$$

Margin condition

Margin condition

$$\mathbb{P}\left[\,0 < |f^{(1)}(X) - f^{(2)}(X)| \le \delta\,\right] \le C\delta^{\alpha}\,.$$

- first used by Goldenshluger and Zeevi (2008) in the one-armed bandit setting
- In the one-armed setup, it is an assumption on the distribution of X only
- Here: fixed marginal (e.g. uniform) so it measures how close the functions are

Margin condition

Margin condition

$$\mathbb{P}\left[\, 0 < |f^{(1)}(X) - f^{(2)}(X)| \le \delta \right] \le C\delta^{\alpha}\,.$$

- first used by Goldenshluger and Zeevi (2008) in the one-armed bandit setting
- In the one-armed setup, it is an assumption on the distribution of X only
- Here: fixed marginal (e.g. uniform) so it measures how close the functions are

Proposition: Conflict α vs. β

$$\alpha\beta > 1 \Longrightarrow \pi^\star \text{ is a.s constant on } [0,1]$$

Illustration of the margin condition

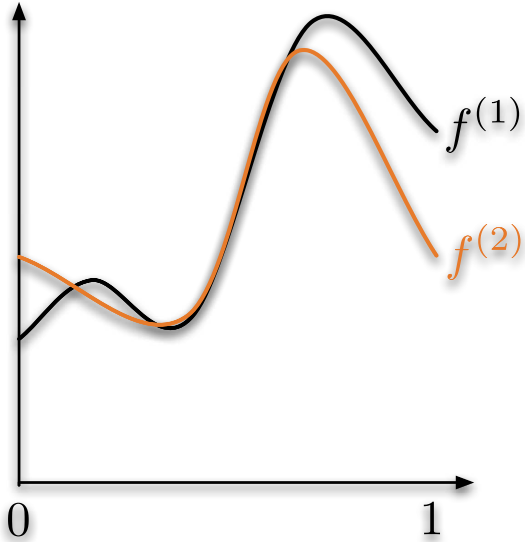


Illustration of the margin condition

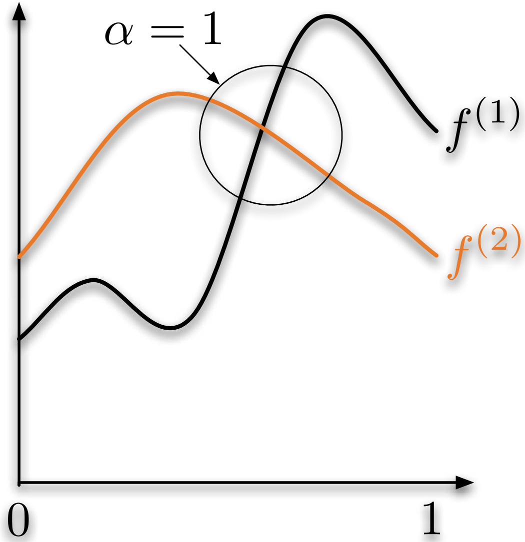
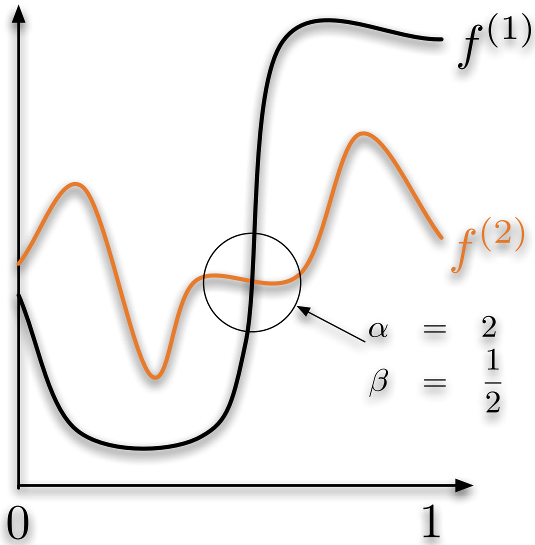


Illustration of the margin condition



Binning (Exploiting smoothness)

- Fix $M > 1$. Consider the bins

$$B_j = [j/M, (j+1)/M)$$

- Consider the **average reward** on each bin

$$\bar{f}_j^{(i)} = \frac{1}{p_j} \int_{B_j} f^{(i)}(x) \mathrm{d}x \,,$$

$$\boxed{Z_t = j \text{ iff } X_t \in B_j}$$

Binned UCB

- For uniformly distributed X_t, we have

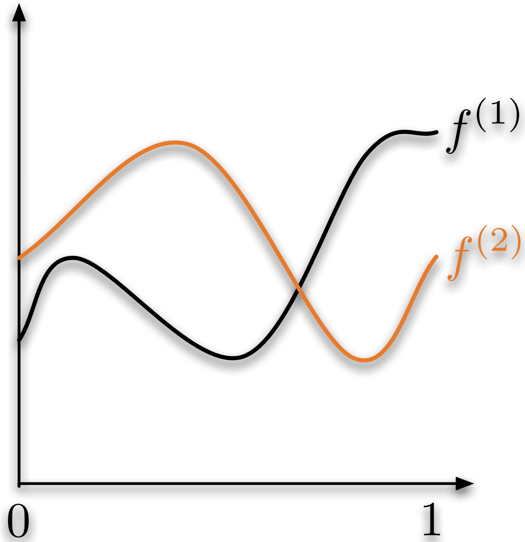
$$p_j = \mathbb{P}(Z_t = j) = \mathbb{P}(X_t \in B_j) = 1/M$$

- The rewards are

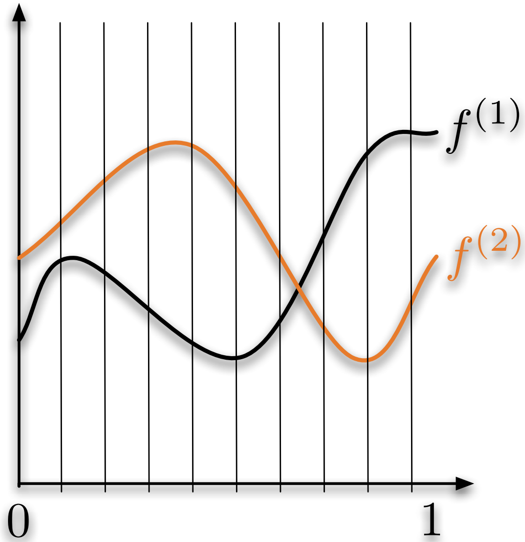
$$\mathbb{E}\left[Y_t^{(i)} | Z_t = j\right] = \bar{f}_j^{(i)} \quad t = 1, 2, \ldots, \; i = 1, 2\,,$$

Play UCB on the $(Z_t, Y_t), t = 1, \ldots, n$

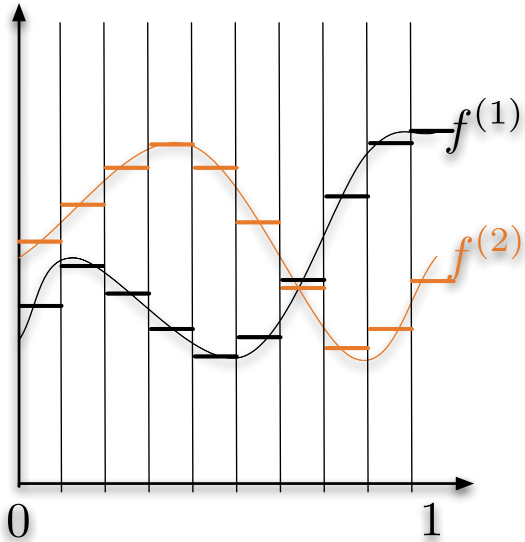
Binned problem



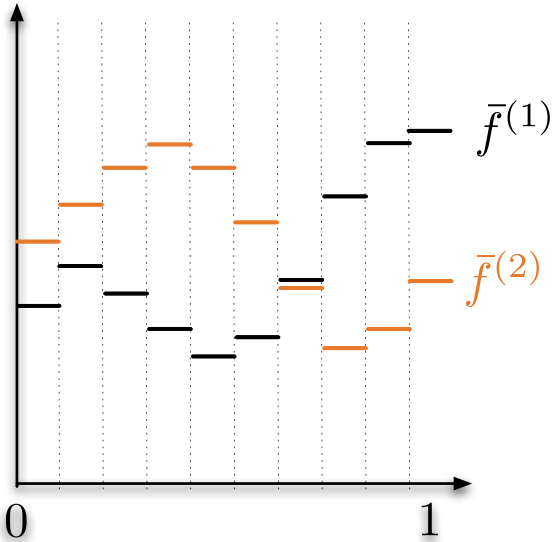
Binned problem



Binned problem



Binned problem



Two armed bandit problem with discrete covariates

- Covariates: $\{Z_t\}$ i.i.d in $\{1, \ldots, M\}$

$$P(Z_t = j) = p_j, \quad t = 1, 2, \ldots$$

- Rewards: $Y_t^{(i)} \in [0, 1]$

$$\mathbb{E}\left[Y_t^{(i)} | Z_t = j\right] = \bar{f}_j^{(i)} \quad t = 1, 2, \ldots, \; i = 1, 2\,,$$

- Oracle policy pulls at time t

$$\pi^\star(Z_t) = \underset{i=1,2}{\operatorname{argmax}}\, \bar{f}_{Z_t}^{(i)}$$

Regret

- Regret given by

$$R_n(\pi) = \mathbb{E} \sum_{j=1}^{M} \sum_{t=1}^{n} \left(\bar{f}_j^{(\pi^\star(j))} - \bar{f}_j^{(\pi_t(j))} \right) \mathbb{I}(Z_t = j)$$

Regret

- Regret given by

$$R_n(\pi) = \mathbb{E} \sum_{j=1}^{M} \sum_{t=1}^{n} \left(\bar{f}_j^{(\pi^\star(j))} - \bar{f}_j^{(\pi_t(j))} \right) \mathbb{I}(Z_t = j)$$

Idea: play independently for each $j = 1, \ldots M$

UCB policy for discrete covariate

- Based **Upper Confidence Bounds** given by concentration inequalities (Hoeffding or Bernstein):

$$B_t(s) := \sqrt{\frac{2 \log t}{s}} .$$

- Define the number of times $\hat{\pi}$ prescribed to pull arm i and $Z_t = j$, before time t

$$N_j^{(i)}(t) = \sum_{s=1}^{t} \mathbb{1}(Z_s = j, \hat{\pi}_s(Z_s) = i) ,$$

- Average reward collected at those times

$$\overline{Y}_j^{(i)}(t) = \frac{1}{N_j^{(i)}(t)} \sum_{s=1}^{t} Y_s^{(i)} \mathbb{1}(Z_s = j, \hat{\pi}_s(Z_s) = i) ,$$

A first bound on the regret

Binned UCB policy: conditionally on $Z_t = j$,

$$\hat{\pi}_t(j) = \operatorname*{argmax}_{i=1,2} \left\{ \overline{Y}_j^{(i)}(t) + B_t(N_j^{(i)}(t)) \right\}$$

Theorem 1. A first bound on the regret

Denote by $\Delta_j = |\bar{f}_j^{(1)} - \bar{f}_j^{(2)}|$.

$$R_n(\hat{\pi}) \leq C \sum_{j=1}^{M} \left(\Delta_j + \frac{\log n}{\Delta_j} \right)$$

Direct consequence of Auer, Cesa-Bianchi & Fischer (2002)

Margin condition

$$\sum_{j=1}^{M} \left(\Delta_j + \frac{\log n}{\Delta_j} \right)$$

- The previous bound can become arbitrary large if one the $\Delta_j, j = 1, \ldots, M$ becomes too small.
- Using the margin condition we can make local conclusions on gaps Δ_j:

Few j's such that Δ_j is small

Upper bound

Theorem 2. A bound on the regret for the binned UCB policy

Fix $\alpha > 0$ and $0 < \beta \leq 1$ and choose $M \sim (n/\log n)^{\frac{1}{2\beta+1}}$. Then

$$R_n(\hat{\pi}) \leq \begin{cases} Cn\left(\frac{n}{\log n}\right)^{-\frac{\beta(1+\alpha)}{2\beta+1}} & \text{if } \alpha < 1 \\ Cn\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} & \text{if } \alpha > 1 \end{cases}$$

Suboptimality for $\alpha > 1$

- If $\alpha > 1$, the bound becomes

$$R_n(\hat{\pi}) \leq C\Big[nM^{-\beta(1+\alpha)} + M \log n\Big]$$

- Minimum for

$$M \sim \left(\frac{n}{\log n}\right)^{\frac{1}{\beta(1+\alpha)+1}}$$

- which yields

$$R_n(\hat{\pi}) \leq Cn\left(\frac{n}{\log n}\right)^{-\frac{\beta(1+\alpha)}{\beta(1+\alpha)+1}}$$

- Problem is: too many bins. Solution: Online/adaptive construction of the bins

Conditional distributions

- The distribution of $Y^{(i)}|X$ belongs to $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, where θ is the mean parameter:

$$\theta = \int x \mathrm{d}P_\theta(x)$$

- Assume that the family $\mathcal{P}$ is such that

$$\mathcal{K}(P_\theta, P_{\theta'}) \leq \frac{(\theta - \theta')^2}{\kappa}, \quad \kappa > 0\,.$$

  For any $\theta, \theta' \in \Theta \subset \mathbb{R}$

- Satisfied in particular for Gaussian (location) and Bernoulli families.

Minimax lower bound

Theorem 3.

Let $\alpha\beta \leq 1$ and the covariates $\{X_t\}$ be uniformly distributed on $[0,1]^d$. Assume also that $\{P_\theta^{(i)},\ \theta \in \mathrm{Im}_{f^{(i)}}(\mathcal{X})\}$ satisfies the condition on Kullback-leibler for any $i = 1, 2$. Then, for any policy π,

$$\sup_{f^{(1)}, f^{(2)} \in \Sigma(\beta, L)} R_n(\pi) \geq Cn \cdot n^{-\frac{\beta(1+\alpha)}{2\beta+1}},$$

for some positive constant C.

Comments

- Same bound as in the full information case (see Audibert & Tsybakov, 07)
- Gap (of logarithmic size) between upper and lower bound.

Extensions

- Higher dimension $d \geq 2$, choose $\|\cdot\|_\infty$

$$R_n(\hat{\pi}) \leq C(d) n \left(\frac{n}{\log n}\right)^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

- The lower bound also holds.
- Unknown n: doubling trick

K-armed bandit

- K-armed bandit problem

$$\mathbb{P}\Big[\, 0 < \min_{i \neq i^\star(X)} |f^{(i)}(X) - f^{(i^\star(X))}(X)| \leq \delta \Big] \leq C\delta^\alpha \,.$$

where $i^\star(x) = \mathrm{argmax}_{1 \leq i \leq K} f^{(i)}(x)$

$$R_n(\hat{\pi}) \leq CKn\left(\frac{n}{\log n}\right)^{-\frac{\beta(1+\alpha)}{2\beta+1}}$$

Conclusion

- We introduced a simple model to handle covariates and proposed a naive policy.
- It has near optimal rates on the regret
- Same rates as full information case but new techniques.
- Current research"
    1. Adaptive partitioning to handle $\alpha > 1$
    2. Use of kernel-type (smooth) regression estimators (fill the gap??)
    3. Time varying rewards