Multi Armed Bandits I: Background

Assaf Zeevi*

Graduate School of Business

Columbia University

Set up: What are multi-armed bandits?

- simple models for sequential decision making under uncertainty
- m slot machines with **random rewards** that are machine-dependent
  - one machine is "best" (has highest average reward)
- gambler plays to maximize profits

  either over infinite horizon (discounted) or finite horizon
- gambler **does not know** identity of "best" machine...
  - needs to "test" and figure out which one is best...
  - ... but each wrong pull gives suboptimal reward

Set up: What are multi-armed bandits?

- simple models for sequential decision making under uncertainty
- m slot machines with **random rewards** that are machine-dependent
  - one machine is "best" (has highest average reward)
- gambler plays to maximize profits

  either over infinite horizon (discounted) or finite horizon
- gambler **does not know** identity of "best" machine...
  - needs to "test" and figure out which one is best...
  - ... but each wrong pull gives suboptimal reward

Q. What strategy maximizes cumulative profits?

Set up: What are multi-armed bandits?

- simple models for sequential decision making under uncertainty
- m slot machines with **random rewards** that are machine-dependent
  - one machine is “best” (has highest average reward)
- gambler plays to maximize profits

  either over infinite horizon (discounted) or finite horizon
- gambler **does not know** identity of “best” machine...
  - needs to “test” and figure out which one is best...
  - ... but each wrong pull gives suboptimal reward

Q. What strategy maximizes cumulative profits?

> *classical tradeoff between* **exploration** *and* **exploitation**

Application areas

applications in adaptive control, economics, statistics, machine learning...

- clinical trials (original motivation, and focus of many papers)
- economics (pricing with unknown demand curve)
- auctions (posted price auctions, ad-word auctions)
- operations management (dynamic assortment planning problems)
- marketing (customized advertising)
- **online advertising and behavioral targeting**
- wireless communications and cognitive radio

Two armed bandits: Problem formulation

- **setup:**
  - two statistical populations (arms) with densities $f(x;\theta_i)$, $i = 1, 2$
  - parameters θ_1, θ_2 are unknown to decision maker...
  - each time t, sample $Y_t^{(i)}$ from one of the populations
- **strategy** $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$
  - $\pi_t \in \{1, 2\}$
  - determines next "pull" based on past actions and observations

    [adapted to history]
- **objective:** maximize expected cumulative returns

Problem formulation (cont'd)

Formulation I: Bayesian, infinite horizon setup

- have **prior** dist'n λ_i over parameter θ_i $i = 1, 2$
- objective: maximize infinite horizon discounted cumulative rewards

$$\max \int \left\{ \mathbb{E}_\theta \sum_{t=1}^{\infty} Y_t^{(\boldsymbol{\pi_t})} \beta^t \right\} \lambda(d\theta)$$

over admissible policies $\boldsymbol{\pi}$

- $\beta \in (0, 1)$ is discount factor

Problem formulation (cont'd)

Formulation II: non-Bayesian, finite horizon

- total rewards up to time n under strategy $\boldsymbol{\pi}$

$$r_n(\boldsymbol{\pi}, \theta) = \mathbb{E}_\theta \sum_{t=1}^{n} Y_t^{(\boldsymbol{\pi_t})},$$

- benchmark: oracle rule $\boldsymbol{\pi}^*$ that **knows** the parameters

$$r_n^*(\theta) = n \cdot \max\{\mu_1, \mu_2\} \quad \text{where } \mu_i := \mathbb{E} Y_t^{(i)}$$

Problem formulation (cont'd)

Formulation II: non-Bayesian, finite horizon

- total rewards up to time n under strategy $\boldsymbol{\pi}$

$$r_n(\boldsymbol{\pi}, \theta) = \mathbb{E}_\theta \sum_{t=1}^{n} Y_t^{(\boldsymbol{\pi_t})},$$

- benchmark: oracle rule $\boldsymbol{\pi}^*$ that **knows** the parameters

$$r_n^*(\theta) = n \cdot \max\{\mu_1, \mu_2\} \quad \text{where } \mu_i := \mathbb{E}Y_t^{(i)}$$

- regret: loss due to not having "full information"

$$\mathcal{R}_n(\boldsymbol{\pi}, \theta) = r_n^*(\theta) - r_n(\boldsymbol{\pi}, \theta)$$

- objective: minimize regret over all admissible policies $\boldsymbol{\pi}$

An abridged history of the subject

- **first formulated** by Thompson (1933) / Robbins (1952)
  - inspired by applications in clinical trials
  - focus was on finite horizon non-Bayesian problem

An abridged history of the subject

- **first formulated** by Thompson (1933) / Robbins (1952)
  - inspired by applications in clinical trials
  - focus was on finite horizon non-Bayesian problem
- voluminous literature with many entries across numerous fields/disciplines
- can classify roughly into three categories (formulation and analysis-wise)
  - Bayesian, dynamic programming (DP), seek optimal solutions
  - Frequentist, non-DP, seek asymptotically optimal solutions
  - Adversarial (mostly CS lit.), non-DP, seek approximate solutions

An abridged history (cont'd)

The first major breakthrough: **Gittins and Jones (1974)**

An abridged history (cont'd)

The first major breakthrough: **Gittins and Jones (1974)**

- Bayesian infinite horizon discounted formulation
- characterized **optimal policy** [index rule / Gittins index]
  - takes simple form: at each stage solve

$$\nu(\lambda_i) = \sup_{\tau \geq 1} \frac{\int \left\{ \mathbb{E}_\theta \sum_{t=1}^{\tau} Y_t^{(i)} \beta^t \right\} \lambda_i(d\theta)}{\int \left\{ \mathbb{E}_\theta \sum_{t=1}^{\tau} \beta^t \right\} \lambda_i(d\theta)}$$

  - use current **posterior** update of λ_i at each stage...
  - optimize over all **stopping times** τ

An abridged history (cont'd)

The first major breakthrough: **Gittins and Jones (1974)**

- Bayesian infinite horizon discounted formulation
- characterized **optimal policy** [index rule / Gittins index]
  - takes simple form: at each stage solve

$$\nu(\lambda_i) = \sup_{\tau \geq 1} \frac{\int \left\{ \mathbb{E}_\theta \sum_{t=1}^{\tau} Y_t^{(i)} \beta^t \right\} \lambda_i(d\theta)}{\int \left\{ \mathbb{E}_\theta \sum_{t=1}^{\tau} \beta^t \right\} \lambda_i(d\theta)}$$

  - use current **posterior** update of λ_i at each stage...
  - optimize over all **stopping times** τ
- requires solving an optimal stopping problem
  - outside of special cases [simple dist'ns, conjugate priors]

  can be quite hard...

Pitfalls of optimal policies

Key observation: **Rothschild (1974)**, McLennan (1984), Brezzi and Lai (2000)

- optimal policy (just discussed) has some interesting “features”

Pitfalls of optimal policies

Key observation: **Rothschild (1974)**, McLennan (1984), Brezzi and Lai (2000)

- optimal policy (just discussed) has some interesting "features"

– it **almost surely** samples from

all but one population only **finitely many times**

Pitfalls of optimal policies

Key observation: **Rothschild (1974)**, McLennan (1984), Brezzi and Lai (2000)

- optimal policy (just discussed) has some interesting "features"

– it **almost surely** samples from

all but one population only **finitely many times**

– **with positive probability** it samples from

the **best** population only **finitely many times**

Pitfalls of optimal policies

Key observation: **Rothschild (1974)**, McLennan (1984), Brezzi and Lai (2000)

- optimal policy (just discussed) has some interesting "features"

– it **almost surely** samples from

all but one population only **finitely many times**

– **with positive probability** it samples from

the **best** population only **finitely many times**

incomplete learning

- intuition: consider a "one armed bandit problem...
  - whenever stop sampling the "unknown arm" never go back to it...
  - can find set of realizations so that with pos. probab. that happens

Pitfalls of optimal policies (cont'd)

- simple [mean-variance] approximation/ bounds for Gittins index:

$$\int \mu_i(\theta)\lambda_i(d\theta) \;\le\; \nu(\lambda_i) \;\le\; \int \mu_i(\theta)\lambda_i(d\theta) + \int \sigma_i(\theta)\lambda_i(d\theta) \cdot \frac{\beta}{1-\beta}$$

  - lower bound achieved by taking $\tau \equiv 1$
  - upper bound uses C-S inequality and $\beta^\tau \le \beta$ for all $\tau \ge 1$
- suggests connection to **myopic** rules [to be revisited shortly...]
  - these may also suffer from incomplete learning...

    [Harrison, Keskin and Z (2011]

Pitfalls of optimal policies (cont'd)

- simple [mean-variance] approximation/ bounds for Gittins index:

$$\int \mu_i(\theta)\lambda_i(d\theta) \;\leq\; \nu(\lambda_i) \;\leq\; \int \mu_i(\theta)\lambda_i(d\theta) + \int \sigma_i(\theta)\lambda_i(d\theta) \cdot \frac{\beta}{1-\beta}$$

  - lower bound achieved by taking $\tau \equiv 1$
  - upper bound uses C-S inequality and $\beta^\tau \leq \beta$ for all $\tau \geq 1$
- suggests connection to **myopic** rules [to be revisited shortly...]
  - these may also suffer from incomplete learning...

    [Harrison, Keskin and Z (2011]

    many interesting economic / game theoretic interpretations...

An abridged history of the subject (cont'd)

The second major breakthrough: **Lai and Robbins (1985)**

- non-Bayesian, finite horizon formulation
- characterized **asymptotically optimal** policies

  "reasonable" policies should satisfy

$$\mathcal{R}_n(\boldsymbol{\pi}, \theta) = o(n)$$

  - are long run average optimal...

An abridged history of the subject (cont'd)

The second major breakthrough: **Lai and Robbins (1985)**

- non-Bayesian, finite horizon formulation
- characterized **asymptotically optimal** policies

  "reasonable" policies should satisfy

$$\mathcal{R}_n(\boldsymbol{\pi}, \theta) = o(n)$$

  - are long run average optimal...

- proposed a strategy $\hat{\boldsymbol{\pi}}$ such that

$$\mathcal{R}_n(\hat{\boldsymbol{\pi}}, \theta) \le [C(\theta) + o(1)] \ \mathbf{\log} \boldsymbol{n}, \quad n \to \infty$$

  - $C(\theta)$ depends on "how far apart" are the two populations...

An abridged history of the subject (cont'd)

The second major breakthrough: **Lai and Robbins (1985)**

- non-Bayesian, finite horizon formulation
- characterized **asymptotically optimal** policies

  "reasonable" policies should satisfy

$$\mathcal{R}_n(\boldsymbol{\pi}, \theta) = o(n)$$

  - ☐ are long run average optimal...
- proposed a strategy $\hat{\boldsymbol{\pi}}$ such that

$$\mathcal{R}_n(\hat{\boldsymbol{\pi}}, \theta) \leq [C(\theta) + o(1)] \;\, \mathbf{\log \boldsymbol{n}}, \quad n \to \infty$$

  - ☐ $C(\theta)$ depends on "how far apart" are the two populations...
- showed that no "reasonable" policy can have better regret than $\hat{\boldsymbol{\pi}}$.
  - ☐ policy is asymptotically optimal

An abridged history of the subject (cont'd)

A major contribution in CS literature: **Auer et al (2002)**

- adversarial setting
  - policy has to do well regardless of possible sequence of rewards
  - the rewards are non-random...
- see recent book by Cesa-Bianchi and Lugosi (2006)
- allows to incorporate non-stationarities
  - identity of "best" arm changes over time
  - related to **restless bandits** line of work [Whittle (1988)]
  - "too hard" unless opponent is restricted in various ways

Other strands of work

continuum armed bandit problems: Agarwal (1995), several recent papers...

- uncountable number of arms $\mathcal{X}$
- essentially a sequential (continuous) stochastic optimization problem...

- pulling an arm $x \in \mathcal{X}$ at time t
  - observe $Y_t = f(x; \varepsilon_t)$ [typically $f(x) + \varepsilon_t$...]
- policy seeks to find $x^* \in \arg\max\{f(x)\}$

Other strands of work

continuum armed bandit problems: Agarwal (1995), several recent papers...

- uncountable number of arms $\mathcal{X}$
- essentially a sequential (continuous) stochastic optimization problem...

- pulling an arm $x \in \mathcal{X}$ at time t
  - observe $Y_t = f(x; \varepsilon_t)$ [typically $f(x) + \varepsilon_t \ldots$]
- policy seeks to find $x^* \in \arg\max\{f(x)\}$
- if function is *strongly concave* then
  - can use standard stochastic approximation type algorithms
- if function is *weakly concave* then
  - need to use search (partition) based sampling...
- if *neither* [can have multiple maxima]
  - can use discretization of standard bandit algorithms

Other strands of work (cont'd)

correlated multi-armed bandits: Mersereau et al (2009)

- arms are not independent
  - a common random variable affects all outcomes

$$Y_t^{(i)} = \theta_0^{(i)} + \theta_1^{(i)} Z + \varepsilon_t$$

  - Z is common to all arms, its distribution is **not known**
  - the other parameters are **known**
- useful when number of arms very large
  - performance typically degrades **linearly** with number of arms
  - above structure can be exploited to control for that...

Discussion of Lai-Robbins results

- simple manipulation shows that

$$\mathcal{R}_n(\boldsymbol{\pi}, \boldsymbol{\theta}) = \textbf{number of ``pulls'' of inferior arm} \cdot (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$$
$$= \mathbb{E}_\theta \sum_{t=1}^{n} \mathbb{I}\{\boldsymbol{\pi_t} \neq \pi^*\} \cdot (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$$

  - $\boldsymbol{\mu_1} > \boldsymbol{\mu_2}$ are the means of the two arms...

- L-R prove that only need about $\log n$ wrong pulls...
  - price of exploration is very small (relative to n)

Discussion of Lai-Robbins (cont'd)

- LR proposed (roughly) the following index

$$\nu_t(i) = \frac{\sum_{\tau=1}^{t} Y_\tau^{(i)} \mathbb{I}\{\boldsymbol{\pi_\tau} = i\}}{T_i(t)} + \sqrt{\frac{C \log t}{T_i(t)}}$$

where $T_i(t)$ = number of pulls in arms i up until time t

$$T_i(t) = \sum_{\tau=1}^{t} \mathbb{I}\{\boldsymbol{\pi_\tau} = i\}$$

- at each time t pull arm with highest index value

Discussion of Lai-Robbins (cont'd)

- LR proposed (roughly) the following index

$$\nu_t(i) = \frac{\sum_{\tau=1}^{t} Y_\tau^{(i)} \mathbb{I}\{\boldsymbol{\pi_\tau} = i\}}{T_i(t)} + \sqrt{\frac{C \log t}{T_i(t)}}$$

where $T_i(t)$ = number of pulls in arms i up until time t

$$T_i(t) = \sum_{\tau=1}^{t} \mathbb{I}\{\boldsymbol{\pi_\tau} = i\}$$

- at each time t pull arm with highest index value
- almost myopic
  - maximize { mean reward to date + "fudge" factor }
  - fudge factor can be interpreted as **upper confidence bound**
- recall connection to upper bound on optimal index rule...

Discussion of Lai-Robbins (cont'd)

- **simpler variation on L-R policy:** given horizon length n
  - pull each arm initially $\log n$ times
  - look at average reward obtained in each arm
  - pick arm with highest mean and pull it exclusively until time n

Discussion of Lai-Robbins (cont'd)

- **simpler variation on L-R policy:** given horizon length n
  - pull each arm initially $\log n$ times
  - look at average reward obtained in each arm
  - pick arm with highest mean and pull it exclusively until time n
- **Intuition:** hypothesis testing problem...

$$\begin{aligned}\mathcal{R}_n(\boldsymbol{\pi}, \theta) &= \mathbb{E}_\theta \sum_{t=1}^{n} \mathbb{I}\{\boldsymbol{\pi_t} \neq \pi^*\} \cdot (\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \\ &= \sum_{t=1}^{n} \mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\} \cdot (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\end{aligned}$$

  - Pr{error} decays exponentially if hypotheses "well separated"
  - $\log n$ pulls in each arm => Pr{error} decays polynomially
  - contribution to regret can be made "small"

Discussion of Lai-Robbins (cont'd)

illustrative example: arm distributions are Gaussian $\mathcal{N}(\mu_i, \sigma^2)$

- using the "forced sampling" startegy:

$$\begin{aligned} \mathcal{R}_n(\boldsymbol{\pi}, \theta) &= (\mu_1 - \mu_2) \cdot \sum_{t=1}^{n} \mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\} \\ &\leq (\mu_1 - \mu_2) \cdot \boldsymbol{\kappa \log n} \; + \; (\mu_1 - \mu_2) \cdot \sum_{t=2\boldsymbol{\kappa \log n}}^{n} \mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\} \end{aligned}$$

Discussion of Lai-Robbins (cont'd)

illustrative example: arm distributions are Gaussian $\mathcal{N}(\mu_i, \sigma^2)$

- using the "forced sampling" startegy:

$$\begin{aligned}\mathcal{R}_n(\boldsymbol{\pi}, \theta) &= (\mu_1 - \mu_2) \cdot \sum_{t=1}^{n} \mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\} \\ &\leq (\mu_1 - \mu_2) \cdot \boldsymbol{\kappa} \log \boldsymbol{n} \; + \; (\mu_1 - \mu_2) \cdot \sum_{t=2\boldsymbol{\kappa}\log \boldsymbol{n}}^{n} \mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\}\end{aligned}$$

- bounding Pr$\{$error$\}$ for $t \geq 2\boldsymbol{\kappa} \log \boldsymbol{n}$

$$\begin{aligned}\mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\} &= \mathbb{P}_\theta \left\{ \sum_{\tau=1}^{\boldsymbol{\kappa}\log \boldsymbol{n}} Y_\tau^{(1)} < \sum_{\tau=1}^{\boldsymbol{\kappa}\log \boldsymbol{n}} Y_\tau^{(2)} \right\} \\ &\leq \exp\left\{ -\boldsymbol{\kappa} \log \boldsymbol{n} \frac{(\mu_1 - \mu_2)^2}{2\sigma^2} \right\}\end{aligned}$$

  - □ quantity in exponent is Kullback-Leibler divergence $\mathcal{K}(\theta_1 \| \theta_2)$

Discussion of Lai-Robbins (cont'd)

illustrative example: arm distributions are Gaussian $\mathcal{N}(\mu_i, \sigma^2)$

- using the "forced sampling" startegy:

$$\mathcal{R}_n(\boldsymbol{\pi}, \theta) = (\mu_1 - \mu_2) \cdot \sum_{t=1}^{n} \mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\}$$

$$\leq (\mu_1 - \mu_2) \cdot \boldsymbol{\kappa \log n} + (\mu_1 - \mu_2) \cdot \sum_{t=2\boldsymbol{\kappa \log n}}^{n} \mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\}$$

- bounding Pr$\{$error$\}$ for $t \geq 2\boldsymbol{\kappa \log n}$

$$\mathbb{P}_\theta\{\boldsymbol{\pi_t} \neq \pi^*\} = \mathbb{P}_\theta \left\{ \sum_{\tau=1}^{\boldsymbol{\kappa \log n}} Y_\tau^{(1)} < \sum_{\tau=1}^{\boldsymbol{\kappa \log n}} Y_\tau^{(2)} \right\}$$

$$\leq \exp\left\{ -\boldsymbol{\kappa \log n} \frac{(\mu_1 - \mu_2)^2}{2\sigma^2} \right\}$$

  - □ quantity in exponent is Kullback-Leibler divergence $\mathcal{K}(\theta_1 \| \theta_2)$

- choose $\boldsymbol{\kappa} = 1/\mathcal{K}$ to balance the regret contributions...

Why is this rate of regret best possible?

- for **any** strategy $\boldsymbol{\pi}$

$$\mathcal{R}_n(\boldsymbol{\pi}, \theta) = (\mu_1 - \mu_2) \cdot \mathbb{E}_\theta \sum_{t=1}^{n} \mathbb{I}\{\boldsymbol{\pi_t} \neq \pi^*\}$$

$$=: (\mu_1 - \mu_2) \cdot \mathbb{E}_\theta T_{\text{inf}}(n)$$

  - $T_{\text{inf}}(n)$ = number of times the inferior arm is pulled...

Why is this rate of regret best possible?

- for **any** strategy $\boldsymbol{\pi}$

$$\begin{aligned} \mathcal{R}_n(\boldsymbol{\pi}, \theta) &= (\mu_1 - \mu_2) \cdot \mathbb{E}_\theta \sum_{t=1}^{n} \mathbb{I}\{\boldsymbol{\pi_t} \neq \pi^*\} \\ &=: (\mu_1 - \mu_2) \cdot \mathbb{E}_\theta T_{\text{inf}}(n) \end{aligned}$$

  - $T_{\text{inf}}(n)$ = number of times the inferior arm is pulled...

- a "reasonable" policy needs to work well **for all** parameter configurations

  - can encode that objective using a minimax formulation

$$\sup_\theta \{\mathbb{E}_\theta T_{\text{inf}}(n)\}$$

Why is this rate of regret best possible?

Why is this rate of regret best possible?

- the following result is from Goldenshluger and Z (2011)

$$
\begin{aligned}
\sup_{(\theta_1,\theta_2)} \mathbb{E}_\theta T_{\inf}(n) &\geq \frac{1}{4}\sum_{t=1}^{n} \exp\{-\mathcal{K}_t(\theta_1\|\theta_2)\} \\
&\geq \frac{1}{4}\sum_{t=1}^{n} \exp\{-2\mathbb{E}_\theta T_{\inf}(t)\ \mathcal{K}(\theta_1\|\theta_2)\} \\
&\geq \frac{1}{4}\sum_{t=1}^{n} \exp\left\{-2\sup_\theta\{\mathbb{E}_\theta T_{\inf}(t)\}\ \mathcal{K}(\theta_1\|\theta_2)\right\} \\
&\geq \frac{1}{4} n \exp\left\{-2\sup_\theta\{\mathbb{E}_\theta T_{\inf}(n)\}\ \mathcal{K}(\theta_1\|\theta_2)\right\}
\end{aligned}
$$

- □ first step uses Fano's ineq. on probab. of error in hypothesis testing

Why is this rate of regret best possible?

- the following result is from Goldenshluger and Z (2011)

$$\begin{aligned}
\sup_{(\theta_1,\theta_2)} \mathbb{E}_\theta T_{\inf}(n) &\geq \frac{1}{4}\sum_{t=1}^{n} \exp\{-\mathcal{K}_t(\theta_1\|\theta_2)\} \\
&\geq \frac{1}{4}\sum_{t=1}^{n} \exp\{-2\mathbb{E}_\theta T_{\inf}(t)\ \mathcal{K}(\theta_1\|\theta_2)\} \\
&\geq \frac{1}{4}\sum_{t=1}^{n} \exp\left\{-2\sup_\theta\{\mathbb{E}_\theta T_{\inf}(t)\}\ \mathcal{K}(\theta_1\|\theta_2)\right\} \\
&\geq \frac{1}{4} n \exp\left\{-2\sup_\theta\{\mathbb{E}_\theta T_{\inf}(n)\}\ \mathcal{K}(\theta_1\|\theta_2)\right\}
\end{aligned}$$

  - first step uses Fano's ineq. on probab. of error in hypothesis testing

- observe that we have

$$L_n \geq \frac{1}{4}\, n\, \exp\{-2\mathcal{K}\, L_n\}$$

Why is this rate of regret best possible?

- recall, we had

$$L_n \;\geq\; \frac{1}{4}\, n\; \exp\{-2\mathcal{K}\; L_n\}$$

so

$$L_n \geq \frac{1}{2\mathcal{K}} \log n$$

Why is this rate of regret best possible?

- recall, we had

$$L_n \;\geq\; \frac{1}{4}\, n\; \exp\{-2\mathcal{K}\; L_n\}$$

so

$$L_n \geq \frac{1}{2\mathcal{K}} \log n$$

- hence we have proved that any policy π must satisfy

$$\sup_{\theta} \{\mathbb{E}_{\theta} T_{\inf}(n)\} \;\geq\; \frac{1}{2\mathcal{K}} \log n$$

Limitations of the standard Bandit formulation

Example 1: Allocating treatments in clinical trials

- patients enter sequentially
- receive one of two possible treatments
- treatment efficacy is yet to be determined...
- objective: allocate the "better" treatment to each patient

Limitations of the standard Bandit formulation

Example 1: Allocating treatments in clinical trials

- patients enter sequentially
- receive one of two possible treatments
- treatment efficacy is yet to be determined...
- objective: allocate the "better" treatment to each patient

Example 2: Interactive Marketing

- there are two unique marketing messages
- marketer can dynamically allocate message to each customer
- objective: maximize return over course of marketing campaign

Limitations of the standard Bandit formulation

Example 1: Allocating treatments in clinical trials

- patients enter sequentially
- receive one of two possible treatments
- treatment efficacy is yet to be determined...
- objective: allocate the "better" treatment to each patient

Example 2: Interactive Marketing

- there are two unique marketing messages
- marketer can dynamically allocate message to each customer
- objective: maximize return over course of marketing campaign

Main issue: response/reward is non-homogenous and depends on particulars of patient/consumer

Multi-armed bandits revisited...

- **mean response/reward:** function $f_i(x)$, $i = 1, 2$
  - *function is unknown* to decision maker
- **observable information and realized reward:** each time t
  - observe *side information* X_t
  - select arm i and receive $Y_t^i = f_i(X_t) + \varepsilon_t$

Multi-armed bandits revisited...

- **mean response/reward:** function $f_i(x)$, $i = 1, 2$
  - □ *function is unknown* to decision maker
- **observable information and realized reward:** each time t
  - □ observe *side information* X_t
  - □ select arm i and receive $Y_t^i = f_i(X_t) + \varepsilon_t$
- **strategy $\boldsymbol{\pi}$** based on past actions, *side information* and rewards.
- **total reward:** $r_n(\boldsymbol{\pi}, f) = \mathbb{E}_f \sum_{t=1}^{n} Y_t^{\pi_t}$

Multi-armed bandits revisited...

- **mean response/reward:** function $f_i(x)$, $i = 1, 2$
  - □ *function is unknown* to decision maker
- **observable information and realized reward:** each time t
  - □ observe *side information* X_t
  - □ select arm i and receive $Y_t^i = f_i(X_t) + \varepsilon_t$
- **strategy** $\boldsymbol{\pi}$ based on past actions, *side information* and rewards.
- **total reward:** $r_n(\boldsymbol{\pi}, f) = \mathbb{E}_f \sum_{t=1}^{n} Y_t^{\pi_t}$
- **regret:** loss relative to oracle... $\mathcal{R}(\boldsymbol{\pi}, f) = r_n^*(f) - r_n(\boldsymbol{\pi}, f)$

Multi-armed bandits revisited...

- **mean response/reward:** function $f_i(x)$, $i = 1, 2$
  - *function is unknown* to decision maker
- **observable information and realized reward:** each time t
  - observe *side information* X_t
  - select arm i and receive $Y_t^i = f_i(X_t) + \varepsilon_t$
- **strategy** $\boldsymbol{\pi}$ based on past actions, *side information* and rewards.
- **total reward:** $r_n(\boldsymbol{\pi}, f) = \mathbb{E}_f \sum_{t=1}^{n} Y_t^{\pi_t}$
- **regret:** loss relative to oracle... $\mathcal{R}(\boldsymbol{\pi}, f) = r_n^*(f) - r_n(\boldsymbol{\pi}, f)$
- **minimax regret objective:** seek policy $\boldsymbol{\pi}$ to **minimize**

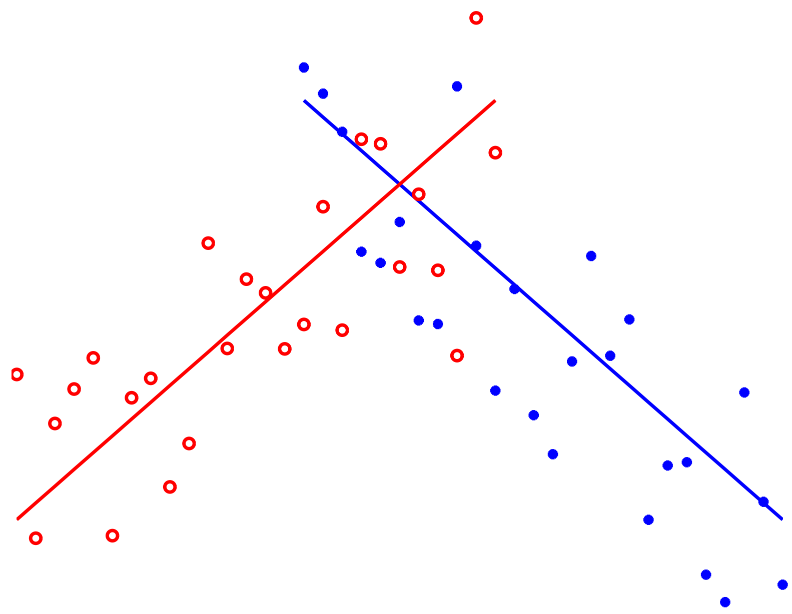
$$\sup_{f \in \mathcal{F}} \mathcal{R}(\boldsymbol{\pi}, f)$$

Illustrative example – Linear response

- **mean response/reward:** $\alpha_i x + \beta_i \quad i = 1, 2$
  - $\theta_i = (\alpha_i, \beta_i)$ unknown...
- **observable information and realized reward:** each time t
  - observe ***covariate*** X_t
  - select arm i and receive $Y_t^i = \alpha_i X_t + \beta_i + \varepsilon_t$

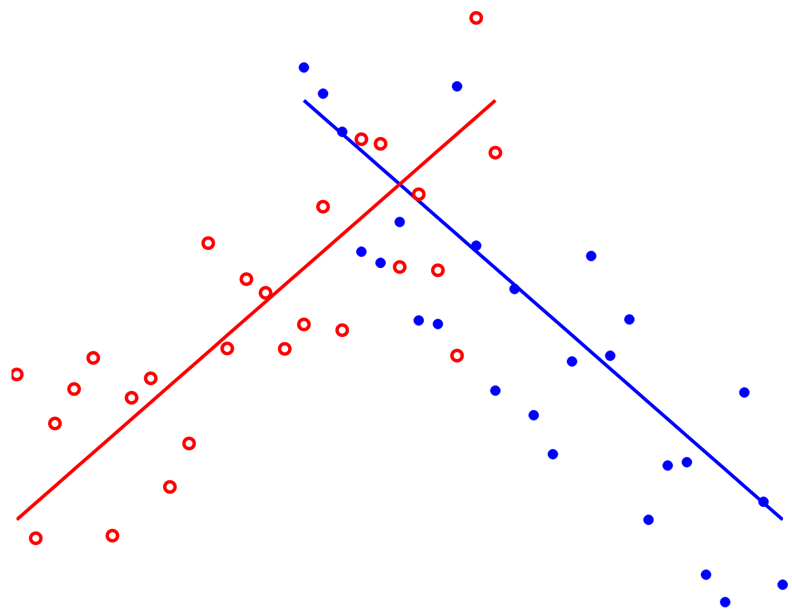
Illustrative example – Linear response

- **mean response/reward:** $\alpha_i x + \beta_i \quad i = 1, 2$
  - $\theta_i = (\alpha_i, \beta_i)$ unknown...
- **observable information and realized reward:** each time t
  - observe *covariate* X_t
  - select arm i and receive $Y_t^i = \alpha_i X_t + \beta_i + \varepsilon_t$



Illustrative example – Linear response

- **mean response/reward:** $\alpha_i x + \beta_i \quad i = 1, 2$
  - $\theta_i = (\alpha_i, \beta_i)$ unknown...
- **observable information and realized reward:** each time t
  - observe ***covariate*** X_t
  - select arm i and receive $Y_t^i = \alpha_i X_t + \beta_i + \varepsilon_t$



- see Goldenshluger and Z (2012) for analysis...

Further simplification: One-armed bandit

- basic idea goes back to Woodroofe [1979]

Further simplification: One-armed bandit

- basic idea goes back to Woodroofe [1979]

- **observation structure and realized reward:** each time t

  - <u>**observe side information**</u> X_t [i.i.d]

  - <u>select arm 1</u> and receive $Y_t^{(1)} = f(X_t; \boldsymbol{\theta}) + \varepsilon_t$

  - <u>or select arm 2</u> and receive $\boldsymbol{Y_t^{(2)} \equiv 0}$ [**known benchmark**]

Further simplification: One-armed bandit

- basic idea goes back to Woodroofe [1979]
- **observation structure and realized reward:** each time t
  - ☐ observe side information X_t [i.i.d]
  - ☐ select arm 1 and receive $Y_t^{(1)} = f(X_t; \boldsymbol{\theta}) + \varepsilon_t$
  - ☐ or select arm 2 and receive $\boldsymbol{Y_t^{(2)} \equiv 0}$ [**known benchmark**]
- **strategy** $\boldsymbol{\pi}$ depends on past actions, side info, and rewards
- **regret:** loss relative to oracle... $\mathcal{R}(\boldsymbol{\pi}, f) = r_n^*(f) - r_n(\boldsymbol{\pi}, f)$

Further simplification: One-armed bandit

- basic idea goes back to Woodroofe [1979]
- **observation structure and realized reward:** each time t
  - observe side information X_t [i.i.d]
  - select arm 1 and receive $Y_t^{(1)} = f(X_t; \boldsymbol{\theta}) + \varepsilon_t$
  - or select arm 2 and receive $\boldsymbol{Y_t^{(2)} \equiv 0}$ [**known benchmark**]
- **strategy** $\boldsymbol{\pi}$ depends on past actions, side info, and rewards
- **regret:** loss relative to oracle... $\mathcal{R}(\boldsymbol{\pi}, f) = r_n^*(f) - r_n(\boldsymbol{\pi}, f)$
- **minimax regret objective:** seek policy $\boldsymbol{\pi}$ to minimize

$$\sup_{f \in \mathcal{F}} \mathcal{R}(\boldsymbol{\pi}, f)$$

Illustrative example [to indicate subtlety...]

- **mean response/reward:** $f(x;\theta) = x - \theta$
- **observable information and realized reward:** each time t
  - observe X_t
  - select arm 1 and receive $Y_t = X_t - \theta + \varepsilon_t$
  - or select arm 2 and receive $Y_t \equiv 0$ [benchmark]

Illustrative example [to indicate subtlety...]

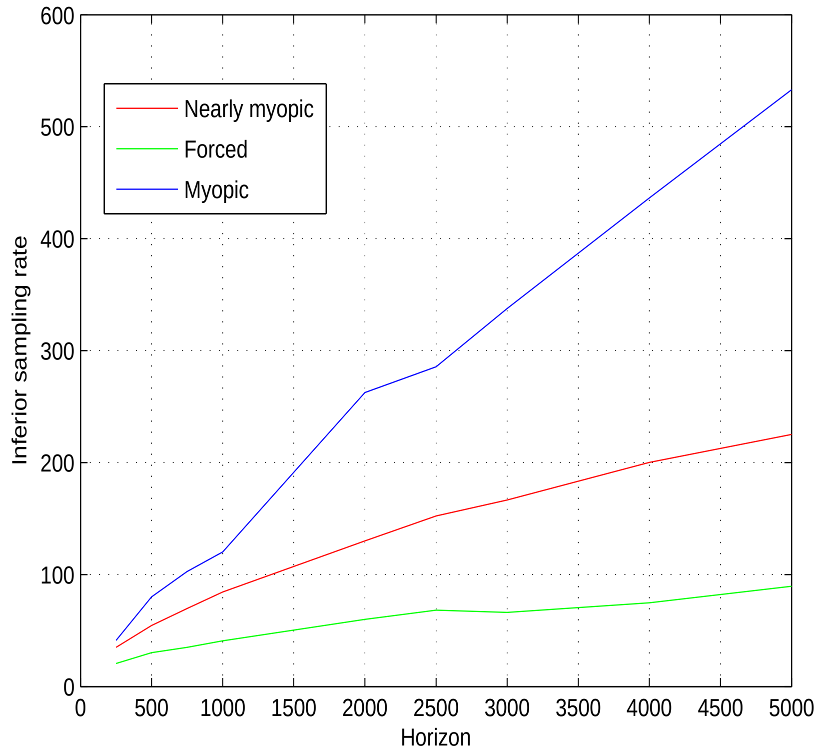
- **mean response/reward:** $f(x;\theta) = x - \theta$
- **observable information and realized reward:** each time t
  - observe X_t
  - select arm 1 and receive $Y_t = X_t - \theta + \varepsilon_t$
  - or select arm 2 and receive $Y_t \equiv 0$ [benchmark]

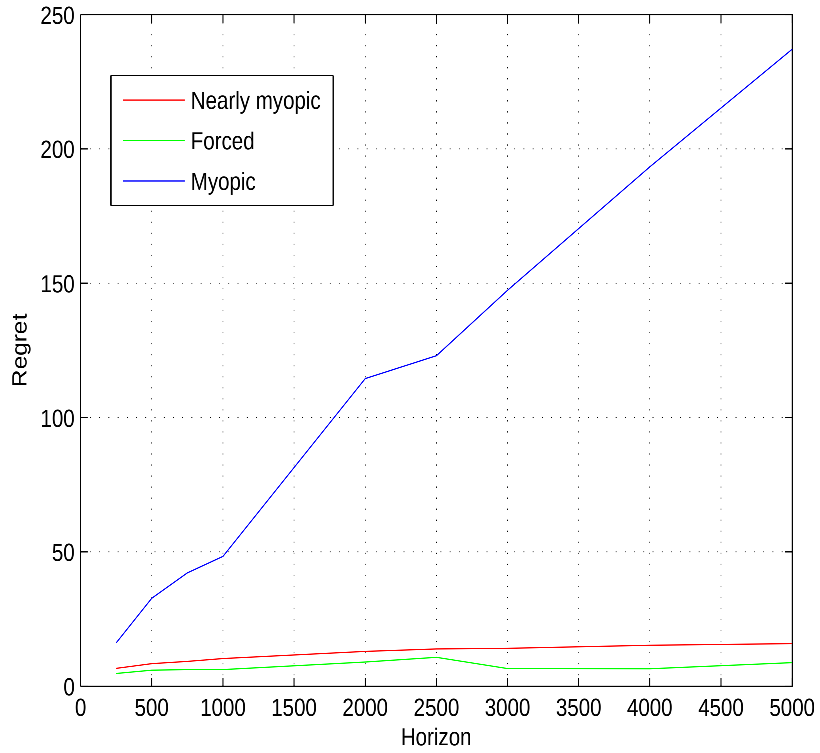
obvious strategy:

- estimate unknown arm parameter $\widehat{\theta}$
  - select arm 1 if $X_t \geq \widehat{\theta}$
  - select arm 2 if $X_t < \widehat{\theta}$
- simple myopic rule...

Numerical illustration

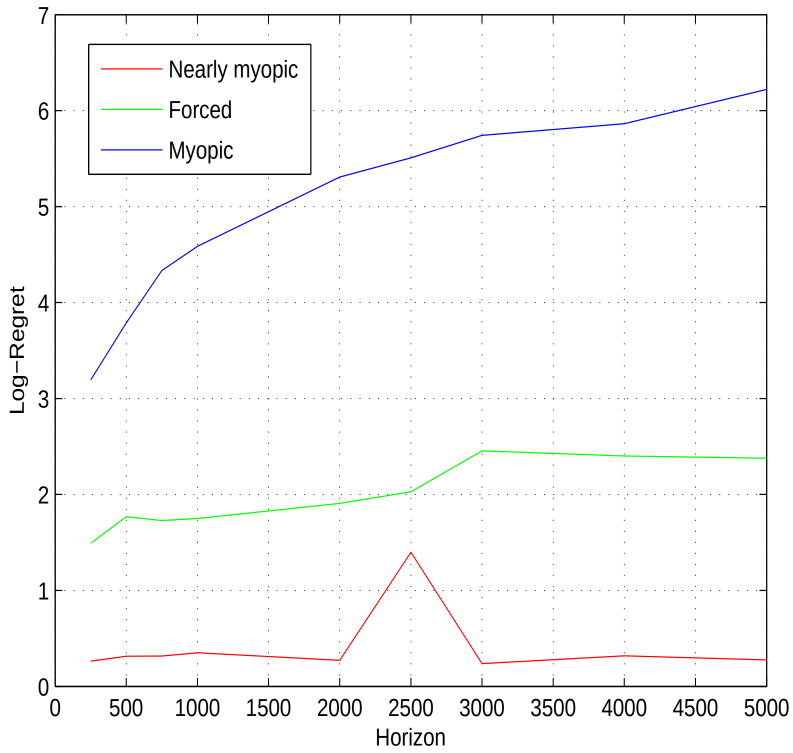
X_t uniform on $[-1, 1]$



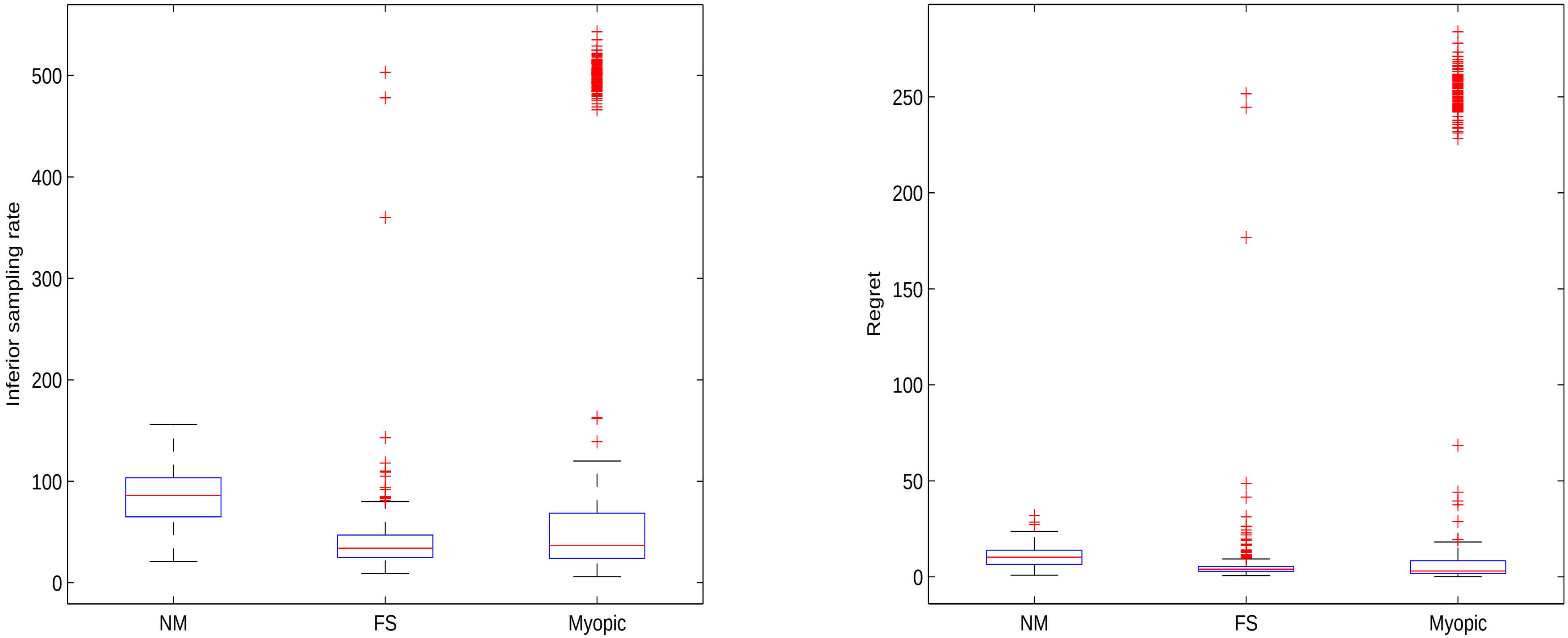


Numerical illustration

$X_t = \pm 1$ **with probability** $p = 1/2$



Boxplots


Inferior sampling rate
500
400
300
200
100
0
NM
FS
Myopic
Regret
250
200
150
100
50
0
NM
FS
Myopic

Algorithm 1: A nearly myopic rule

- **Initialize:** pull arm 1 twice
- **Estimate parameter:** based on X_t and response Y_t for $t = 1, 2$.

Algorithm 1: A nearly myopic rule

- **Initialize:** pull arm 1 twice
- **Estimate parameter:** based on X_t and response Y_t for $t = 1, 2$.
- **Arm selection:**
  - if

    $$X_t \geq \widehat{\theta}_t - \boldsymbol{\delta_t} \quad [\,\text{sequence } \boldsymbol{\delta_t} \text{ is history dependent}\,]$$

    then pull arm 1
  - o.w. pull arm 2 [benchmark action]

Algorithm 1: A nearly myopic rule

- **Initialize:** pull arm 1 twice
- **Estimate parameter:** based on X_t and response Y_t for $t = 1, 2$.
- **Arm selection:**
  - if

$$X_t \geq \widehat{\theta}_t - \boldsymbol{\delta_t} \quad [\,\text{sequence } \boldsymbol{\delta_t} \text{ is history dependent}\,]$$

    then pull arm 1
  - o.w. pull arm 2 [benchmark action]
- **Update estimates:** At each time t
  - update parameter estimates $\widehat{\theta}_t \mapsto \widehat{\theta}_{t+1} \ldots$
- **Repeat**

Algorithm 2: Forced (randomized) sampling

- **Initialize:** pull arm 1 twice
- **Estimate parameter:** based on X_t and response Y_t for $t = 1, 2$.

Algorithm 2: Forced (randomized) sampling

- **Initialize:** pull arm 1 twice
- **Estimate parameter:** based on X_t and response Y_t for $t = 1, 2$.
- **Arm selection:** at each step t
  - **with probability** $1 - \gamma_t$: follow myopic rule
  - **with probability** γ_t: sample from arm 1...

Algorithm 2: Forced (randomized) sampling

- **Initialize:** pull arm 1 twice
- **Estimate parameter:** based on X_t and response Y_t for $t = 1, 2$.
- **Arm selection:** at each step t
  - **with probability $1 - \gamma_t$**: follow myopic rule
  - **with probability γ_t**: sample from arm 1...
- **Update estimates:** At each time t
  - update parameter estimates $\widehat{\theta}_t \mapsto \widehat{\theta}_{t+1} \ldots$
  - update randomization sequence $\gamma_t \mapsto \gamma_{t+1}$.
- **Repeat**

Analysis of proposed algorithms

Q. Why do we have two algorithms?

Analysis of proposed algorithms

Q. Why do we have two algorithms?

Thm. *If distribution of side information is **discrete**, then regret of Algorithm 1 is **bounded** [independent of horizon n].*

Analysis of proposed algorithms

Q. Why do we have two algorithms?

Thm. *If distribution of side information is **discrete**, then regret of Algorithm 1 is **bounded** [independent of horizon n].*

Thm. *If distribution of side information is **continuous**, then regret of Algorithm 2 is of order $\log n$*

Analysis of proposed algorithms

Q. Why do we have two algorithms?

> **Thm.** *If distribution of side information is **discrete**, then regret of Algorithm 1 is **bounded** [independent of horizon n].*

> **Thm.** *If distribution of side information is **continuous**, then regret of Algorithm 2 is of order $\log n$*

- # of wrong pulls ≈ $\sqrt{n}$...
  - contrast with $\log n$ in L-R case

Optimality of the algorithm?

Thm. *Regret cannot diminish faster than* $C \log n$ *uniformly over target class*

Optimality of the algorithm?

Thm. *Regret cannot diminish faster than* $C\log n$ *uniformly over target class*

Logic is very different than L-R problem... [see Goldenshulger and Z (2009)]

Optimality of the algorithm?

Thm. *Regret cannot diminish faster than $C\log n$ uniformly over target class*

Logic is very different than L-R problem... [see Goldenshulger and Z (2009)]

Proof (ideas).

- reduce to Bayesian estimation problem
  - under mean squared error criterion
- use the van Trees inequality (1968)
  - Bayesian version of Cramer-Rao inequality
  - pointwise bound is $\Theta(1/t)$...
  - so sum over horizon n is $\Theta(\log n)$

Final comments

Simple extensions: [Goldenshluger and Z (2012)]

- Multi-armed problem has some similar flavor...
- Higher dimensions also works the same

Final comments

Simple extensions: [Goldenshluger and Z (2012)]

- Multi-armed problem has some similar flavor...
- Higher dimensions also works the same

Not-so-simple extensions:

- Nonparametric case

Final comments

Simple extensions: [Goldenshluger and Z (2012)]

- Multi-armed problem has some similar flavor...
- Higher dimensions also works the same

Not-so-simple extensions:

- Nonparametric case

Tomorrow: applications to on-line advertising

- MAB with side information
- several added twists and turns...