Bandit Optimization: Theory and Applications
- Part 2 -

R. Combes, A. Proutiere

Part 2. Structured Bandits

Discrete Structured Bandits

1. Regret lower bounds
2. Examples
3. Efficient algorithms for some structures: unimodal, Lipschitz

Infinite Bandits

1. Regret lower bounds
2. Optimal algorithms

Continuous Structured Bandits

1. Regret lower bounds
2. Unimodal bandits
3. Lipschitz bandits

Conclusion and Open Problems

2-A. Discrete Structured Bandits

Discrete Structured Bandits

- *K* arms
- Reward distributions parametrized by $\theta = (\theta_1, \ldots, \theta_K)$
- Average reward of arm *k*: $\mu_k = \mu_k(\theta)$
- Most often, reward distributions are taken from a single parameter exponential family (e.g. Bernoulli, $\theta_k = \mu_k$)
- *K* can be very large – yielding a prohibitive regret if arms are independent, i.e., $\Theta(K \log(T))$
- Structure matters and has to be exploited!

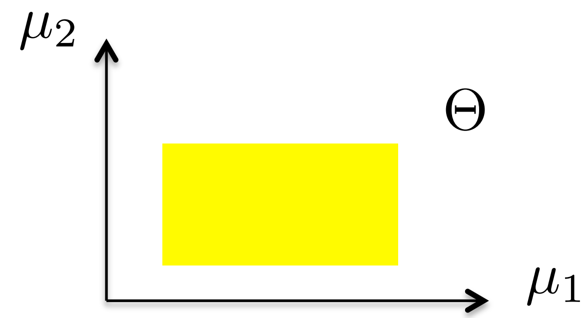
- Notation: $\mu^\star(\theta) = \max_k \mu_k(\theta) = \mu_{k^\star}(\theta)$

Discrete Structured Bandits

- **Unstructured bandits:** average rewards are not related

$$\mu = (\mu_1, \ldots, \mu_K) \in \Theta$$

$$\Theta = \prod_{i=1}^{K} [a_i, b_i]$$

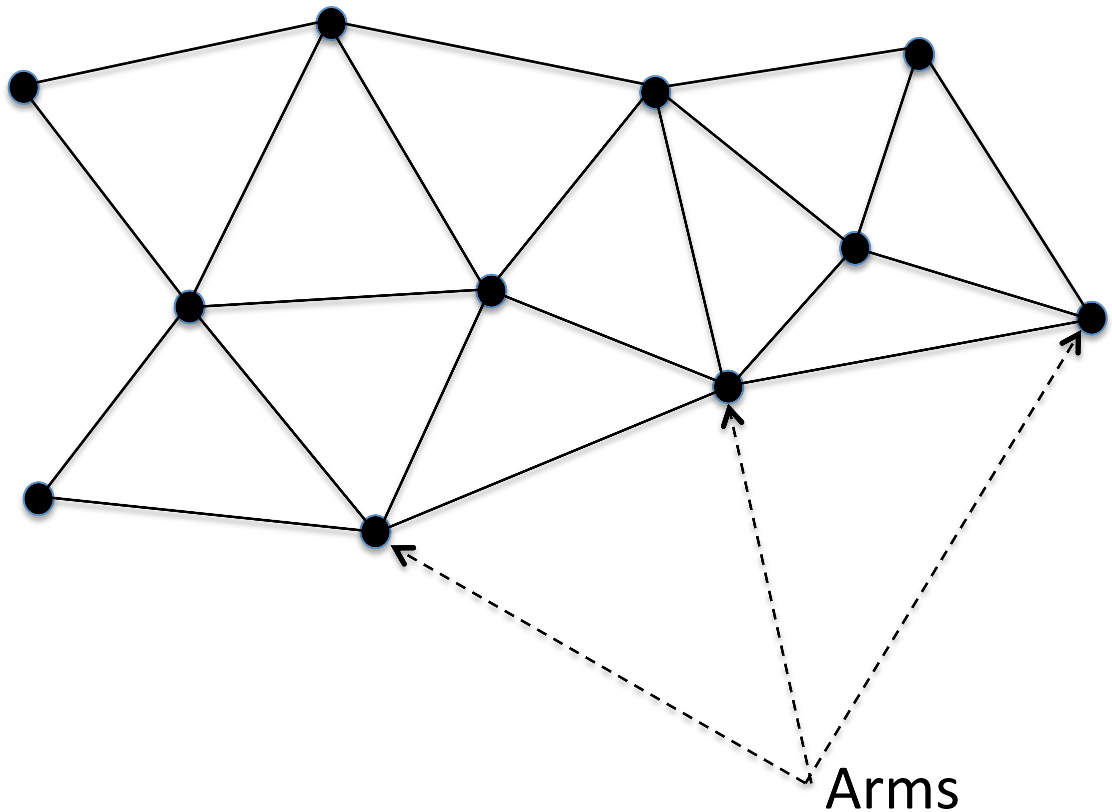


- **Structured bandits:** the decision maker knows that average rewards are related, i.e., that $\mu \in \Theta$

$$\Theta \neq \prod_{i=1}^{K} [a_i, b_i]$$

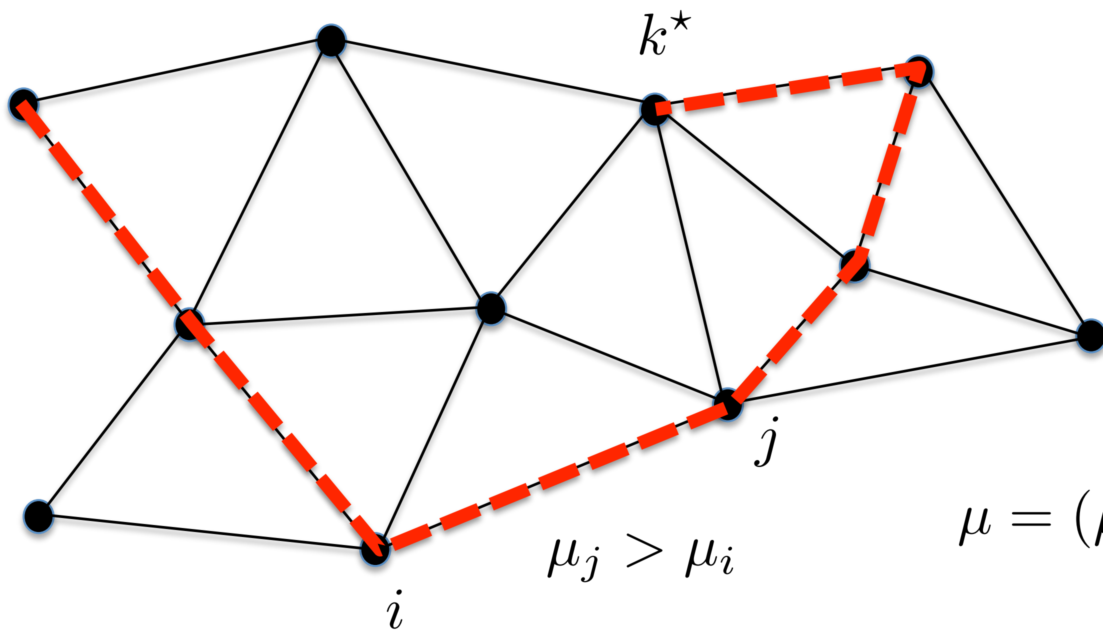
  - The rewards observed for a given action provide side-information about the average rewards of other actions
  - How can we exploit this side-information optimally?

Example 1: Graphical Unimodality



$G = (V, E)$

Example 1: Graphical Unimodality

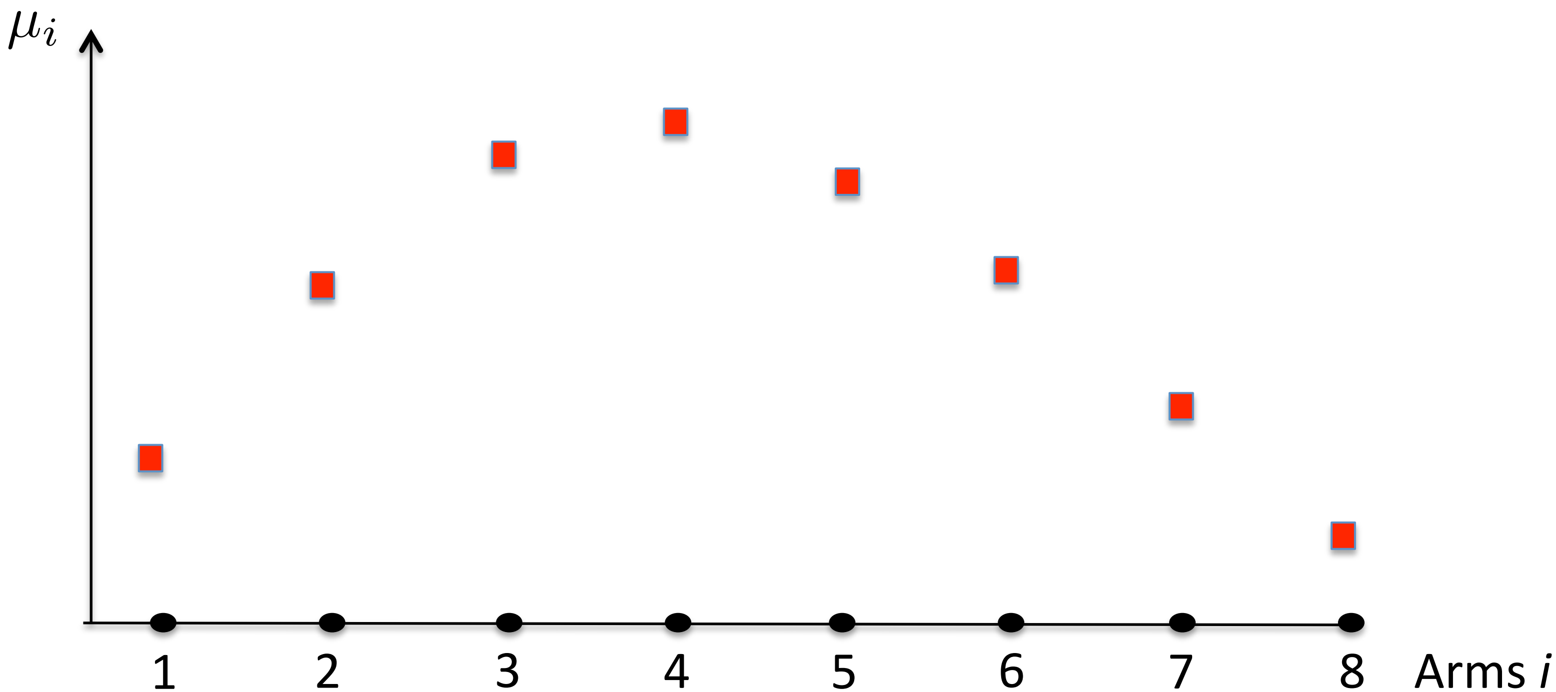


$$G = (V, E)$$

$$\mu = (\mu_i)_{i \in V} \in \mathcal{U}_G$$

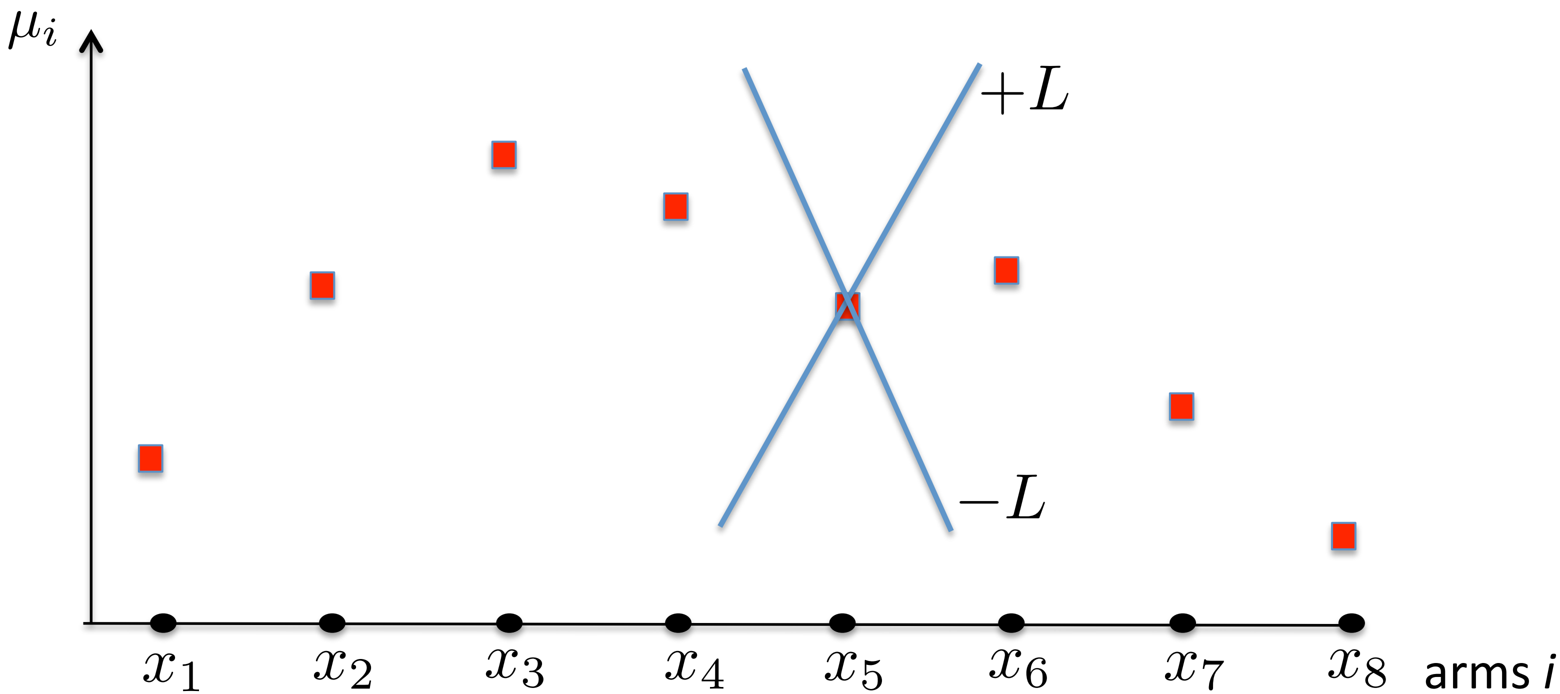
Graphical unimodality: from any vertex, there is a path with increasing rewards to the best vertex.

Example 1: Unimodality



Classical unimodality, graph = line

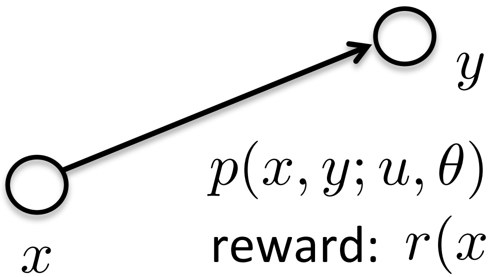
Example 2: Lipschitz



Let $x_1 < x_2 < \ldots < x_K$ denote the *positions* of the arms.
We assume that: $|\mu_k - \mu_{k'}| \leq L \times |x_k - x_{k'}|$.

A Markov Chain Control Perspective

Graves-Lai 1997



$p(x, y; u, \theta)$

reward: $r(x, u)$

- Finite state space $\mathcal{X}$ and action spaces $\theta \in \Theta$
- Unknown parameter

  Θ : compact metric space

- Control: finite set of irreducible control laws $g : \mathcal{X} \to \mathcal{U}$

$$\mu_g(\theta) = \sum_{x \in \mathcal{X}} \pi_\theta^g(x) r(x, g(x))$$

- Optimal control law: $g^\star$
- Regret: $R^\pi(T) = T\mu_{g^\star}(\theta) - \mathbb{E} \sum_{t=1}^{T} r(X_t, g^\pi(X_t))$

Regret lower bound

- KL number under policy g:

$$I^g(\theta, \lambda) = \sum_{x,y} \pi_\theta^g(x) p(x, y; g(x), \theta) \log \frac{p(x, y; g(x), \theta)}{p(x, y; g(x), \lambda)}$$

- Bad parameter set:

$$B(\theta) = \{\lambda \in \Theta : g^\star \text{ not opt.}, I^{g^\star}(\theta, \lambda) = 0\}$$

- Lower bound: $\liminf_{T \to \infty} \frac{R^\pi(T)}{\log(T)} \geq c(\theta)$

$$c(\theta) = \inf \sum_{g \neq g^\star} c_g(\mu_{g^\star}(\theta) - \mu_g(\theta))$$

$$\text{s.t.} \quad \inf_{\lambda \in B(\theta)} \sum_{g \neq g^\star} c_g I^g(\theta, \lambda) \geq 1$$

Application to Structured Bandits

- State space: set of possible rewards
- Control laws: constant mappings to the set of arms, e.g.
  $g = k$
- Transitions (i.i.d. process):

$$p(x, y; k, \theta) = \begin{cases} \theta_k & \text{if } y = 1 \\ 1 - \theta_k & \text{if } y = 0 \end{cases}$$

$$I^k(\theta, \lambda) = KL(\theta_k, \lambda_k)$$

- Average rewards: $g = k$

$$\mu_g(\theta) = \theta_k = \mu_k$$

Regret Lower Bound

- Lower bound: $\liminf_{T\to\infty} \frac{R^{\pi}(T)}{\log(T)} \geq c(\theta)$

$$c(\theta) = \inf_{c_k \geq 0, \forall k} c_k(\mu_{k^\star} - \mu_k)$$

$$\text{s.t.} \quad \inf_{\lambda \in B(\theta)} \sum_{k \neq k^\star} c_k I^k(\theta, \lambda) \geq 1$$

$$B(\theta) = \{\lambda \in \Theta : I^{k^\star}(\theta, \lambda) = 0, \mu^\star(\lambda) > \mu_{k^\star}(\lambda)\}$$

Regret Lower Bound

- Lower bound: $\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq c(\theta)$

$$c(\theta) = \inf_{c_k \geq 0, \forall k} c_k(\mu_{k^\star} - \mu_k)$$

$$\text{s.t.} \inf_{\lambda \in B(\theta)} \sum_{k \neq k^\star} c_k I^k(\theta, \lambda) \geq 1$$

$$B(\theta) = \{\lambda \in \Theta : I^{k^\star}(\theta, \lambda) = 0, \mu^\star(\lambda) > \mu_{k^\star}(\lambda)\}$$

- Identifying the *worst* λ can be challenging
- Examples where it is explicit: unimodal, Lipschitz. In this case, the regret lower solves an LP
- Interpretation: when optimal, an algorithm plays sub-optimal arm k $c_k \log(T)$ times

Asymptotically Optimal Algorithm

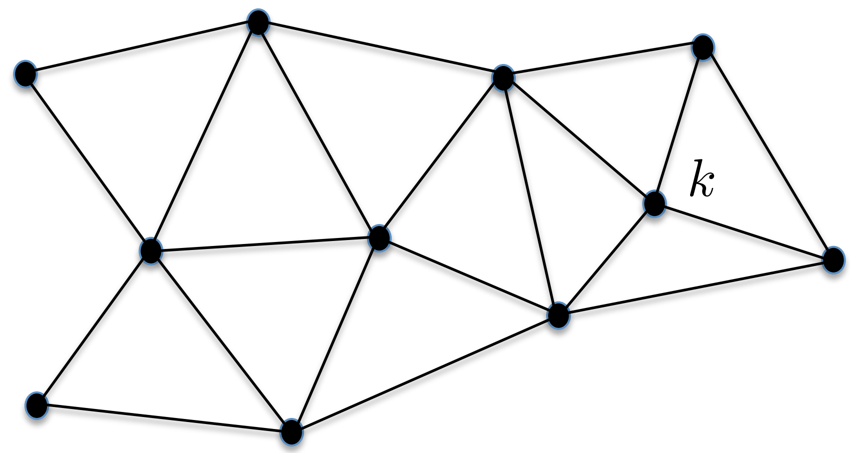
- Graves-Lai's algorithm
  - Uses the doubling trick
  - Needs to solve the regret lower bound problem repeatedly
  - Too complex, and inefficient for reasonable time horizons

2-A.1. Discrete Unimodal Bandits

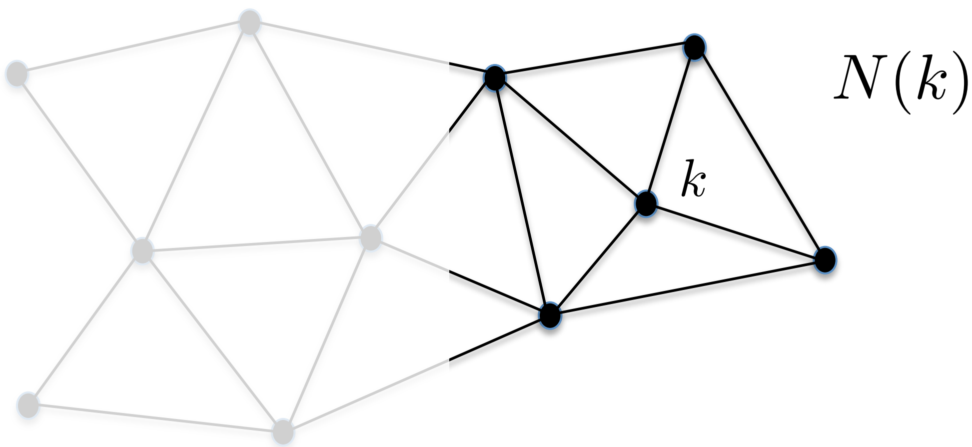
Combes, Proutiere. Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms, **ICML** 2014

Combes et al. Optimal Rate Sampling in 802.11 Systems, **IEEE Infocom** 2014

Regret Lower Bound



Regret Lower Bound



Theorem: For any uniformly good algorithm π

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq c_G(\theta) \qquad c_G(\theta) = \sum_{k\in N(k^\star)} \frac{\mu^\star - \mu_k(\theta)}{KL(\theta_k, \theta_{k^\star})}$$

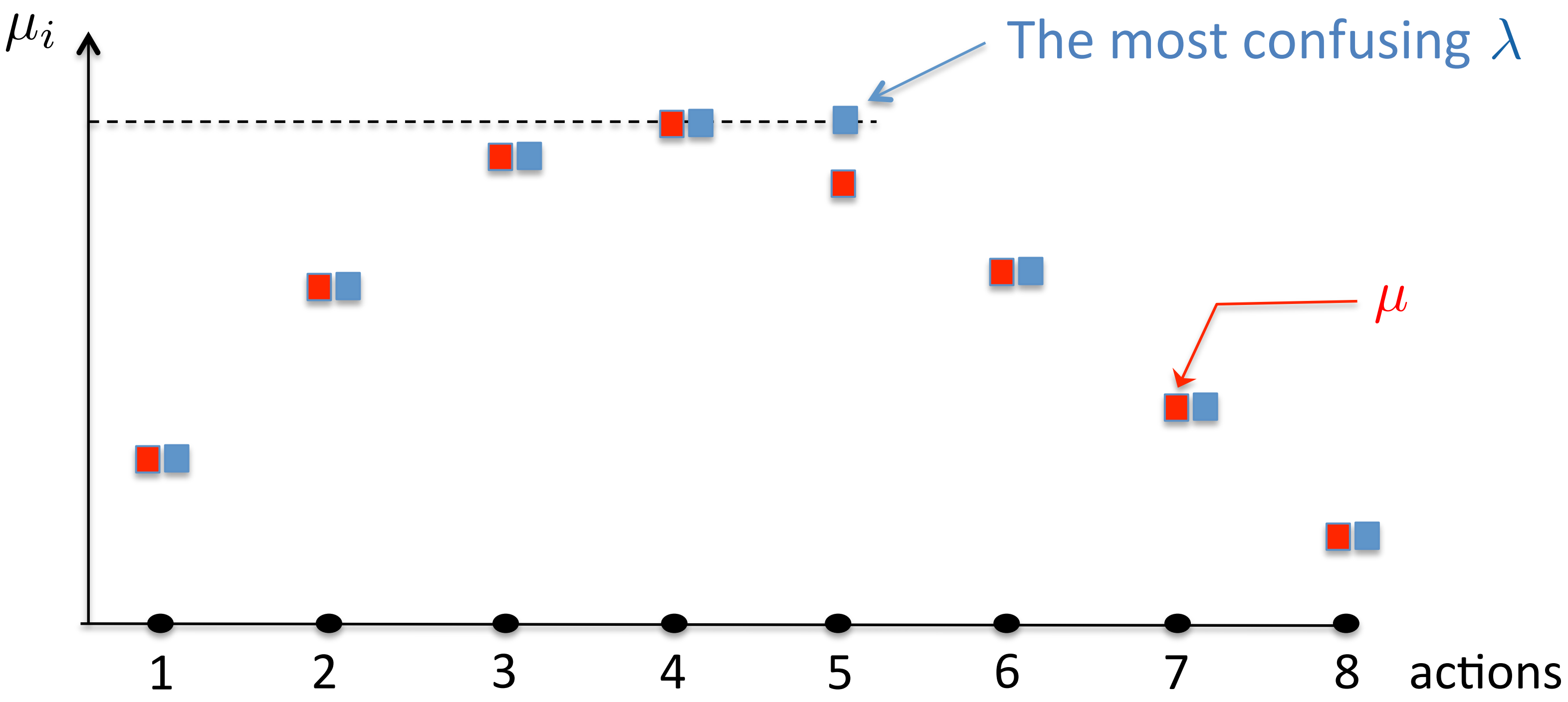
The performance limit does not depend on the size of the decision space! Structure could really help.

Proof

$$\inf \sum_{g \neq g^\star} c_g(\mu_{g^\star}(\theta) - \mu_g(\theta))$$

$$\text{s.t.} \quad \inf_{\lambda \in B(\theta)} \sum_{g \neq g^\star} c_g I^g(\theta, \lambda) \geq 1$$

<u>Example:</u> classical unimodality



Optimal Action Sampling

- Empirical average reward: $\hat{\mu}_k(n) = \dfrac{1}{t_k(n)} \sum_{s=1}^{t_k(n)} X_k(s)$

- Leader at time *n*: $L(n) \in \arg\max_k \hat{\mu}_k(n)$

- Number of times *k* has been the leader: $l_k(n) = \sum_{s=1}^{n} 1_{L(s)=k}$

- Index of *k*: $b_k(n) = \max\{q \in [0,1] : t_k(n) KL(\hat{\mu}_k(n), q) \leq \log(l_{L(n)}(n)) + c \log\log(l_{L(n)}(n))\}$

Optimal Action Sampling

Algorithm – Optimal Action Sampling (OAS)

For $n = 1, \ldots, K$, select action $k(n) = n$

For $n \geq K + 1$, select action $k(n)$:

$$k(n) = \begin{cases} L(n) & \text{if } (l_{L(n)}(n) - 1)/(\gamma + 1) \in \mathbb{N}, \\ \arg \max\limits_{k \in N(L(n))} b_k(n) & \text{otherwise.} \end{cases}$$

Theorem: For any $\mu \in \mathcal{U}_G$, $\quad \lim \sup\limits_{T \to \infty} \dfrac{R^{OAS}(T)}{\log(T)} \leq c_G(\theta).$

Proof

$$R^{OAS}(T) \leq \sum_{k \neq k^\star} \mathbb{E}[l_k(T)]$$

$$+ \sum_{k \in N(k^\star)} (\mu^\star - \mu_k(\theta))\mathbb{E}[\sum_{t=1}^{T} 1_{L(t)=k^\star, k(t)=k}]$$

First term $\leq O(\log\log(T))$

Second term $\leq (1+\epsilon)c(\theta)\log(T) + O(\log\log(T))$

Proof ingredients

1. Decomposition of the set of events

2. Deviation bounds (refined concentration inequalities), e.g.

Lemma. $\{Z_t\}_{t\in\mathbb{Z}}$ independent random variables in $[0, B]$. $\mathcal{F}_n = \sigma(\{Z_t\}_{t\leq n})$, $\mathcal{F} = (\mathcal{F}_n)_{n\in\mathbb{Z}}$. Let $s \in \mathbb{N}$, $n_0 \in \mathbb{Z}$ and $T \geq n_0$. $S_n = \sum_{t=n_0}^{n} B_t(Z_t - \mathbb{E}[Z_t])$, where $B_t \in \{0, 1\}$ is previsible. $t_n = \sum_{t=n_0}^{n} B_t$. $\phi \in \{n_0, \ldots, T+1\}$ a $\mathcal{F}$-stopping time with: either $t_\phi \geq s$ or $\phi = T+1$. Then:

$$\mathbb{P}[S_\phi \geq t_\phi \delta \;,\; \phi \leq T] \leq \exp(-\frac{2s\delta^2}{B^2}).$$

Non-stationary environments

- Average rewards may evolve over time: $\theta(t)$
- Best decision at time *t*: $k^\star(t)$
- Goal: track the best decision
- Regret:

$$R^\pi(T) = \sum_{t=1}^{T} (\mu_{k^\star(t)}(t) - \mu_{k^\pi(t)}(t))$$

- Sub-linear regret cannot be achieved (**Garivier-Moulines** 2011)
- Assumptions: $\theta(t)$ σ-Lipschitz (w.r.t. time), and separation

$$\lim \sup_{T\to\infty} \frac{1}{T} \sum_{n=1}^{T} \sum_{k,k' \in N(k)} 1_{|\theta_k(n) - \theta_{k'}(n)| < \Delta} \leq \phi(K)\Delta$$

OAS with Sliding Window

- SW-OAS (applies OAS over a sliding window of size τ)
- Graphical unimodality holds at any time
- Parameters:

$$\tau = \sigma^{-3/4}\log(1/\sigma)/8, \quad \Delta = \sigma^{1/4}\log(1/\sigma)$$

Theorem: Under $\pi =$ SW-OAS

$$\limsup_T \frac{R^\pi(T)}{T} \le C\phi(K)\sigma^{\frac{1}{4}}\log(1/\sigma)(1+Ko(1)), \quad \sigma \to 0^+$$

OAS with Sliding Window

- Analysis made complicated by the smoothness of the rewards vs. time (previous analysis by **Garivier-Moulines** assumes separation of rewards at any time)
- Upper bound on regret per time unit:
  - Tends to zero when the evolution of average rewards gets smoother

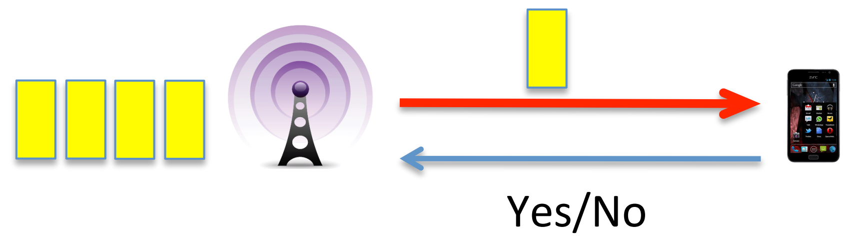
$$\sigma^{1/4}\log(1/\sigma) \to 0, \quad \text{as } \sigma \to 0^+$$

  - Does not depend on the size of the decision space if $\phi(K) \leq C$

Application: Rate adaptation in 802.11

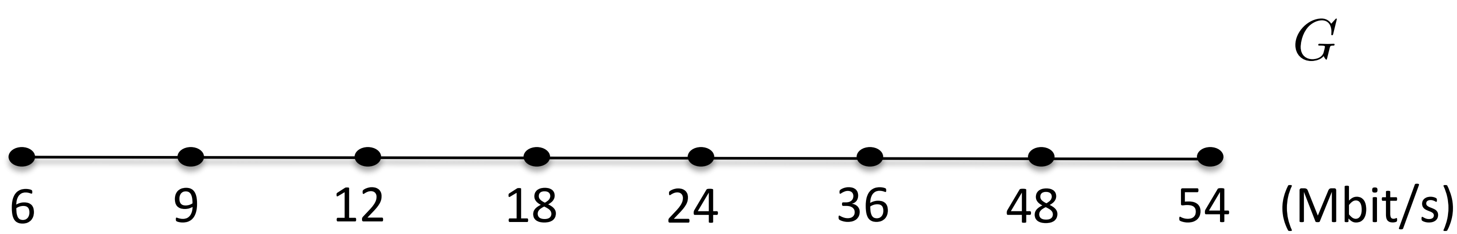
Adapting the modulation/coding scheme to the radio environment

- 802.11 a/b/g



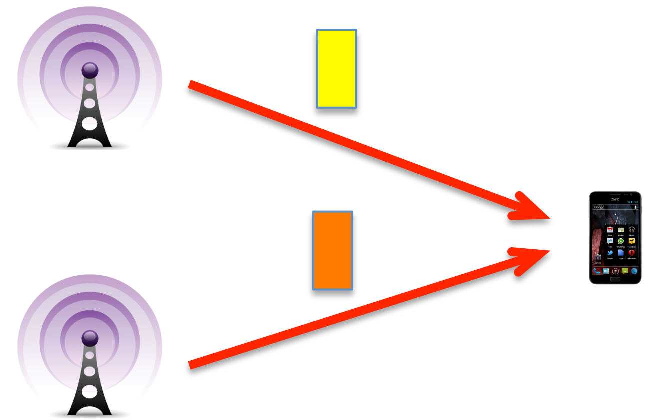
rates	r_1	r_2	$\ldots$	r_N	
Success probabilities	θ_1	θ_2	$\ldots$	θ_N	
Throughputs	μ_1	μ_2	$\ldots$	μ_N	$\mu_i = r_i\theta_i$

- Structure: unimodality + $\theta_1 > \theta_2 > \ldots > \theta_N$

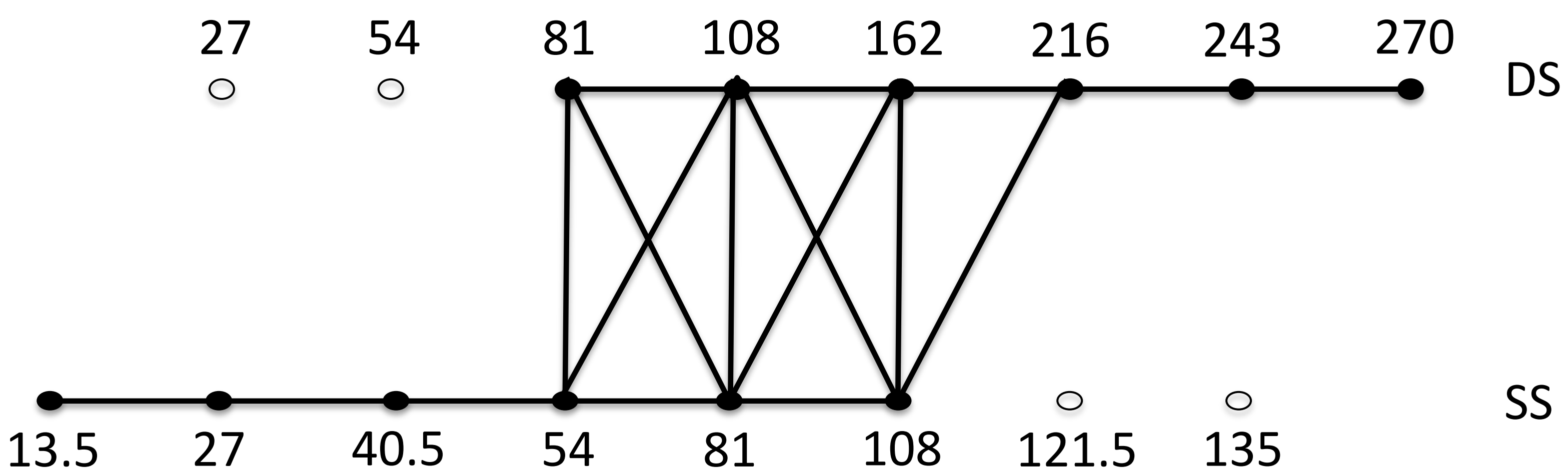


Rate adaptation in 802.11

- 802.11 n/ac MIMO
  Rate + MIMO mode
  (32 combinations in n)



- Example: two modes, single-stream (SS) or double-stream (DS)



State-of-the-art

- ARF (Auto Rate Fallback): after *n* successive successes, probe a higher rate; after two consecutive failures reduce the rate
- AARF: vary *n* dynamically depending on the speed at which the radio environment evolves
- SampleRate: based on achieved throughputs over a sliding window, explore a new rate every 10 packets
- Measurement based approaches: Map SNR to packet error rate (does not work – OFDM): RBAR, OAR, CHARM, …
- 802.11n MIMO: MiRA, RAMAS, …

All existing algorithms are heuristics.

Rate adaptation design: a graphically unimodal bandit with large strategy set

Optimal Rate Sampling

Algorithm – Optimal Rate Sampling (ORS)
For $n = 1, \ldots, K$, select action $k(n) = n$
For $n \geq K + 1$, select action $k(n)$:

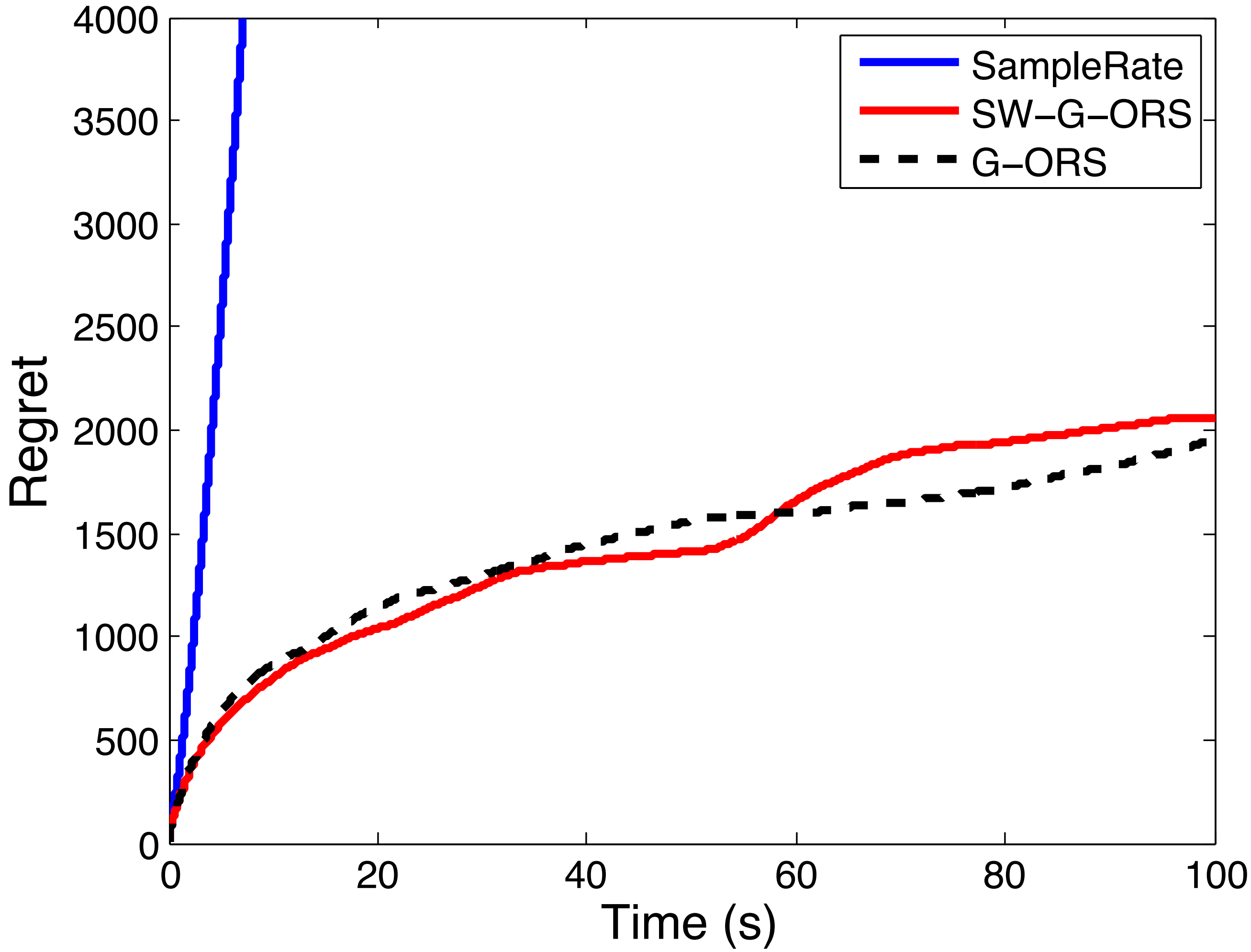
$$k(n) = \begin{cases} L(n) & \text{if } (l_{L(n)}(n) - 1)/(\gamma + 1) \in \mathbb{N}, \\ \arg \max_{k \in N(L(n))} b_k(n) & \text{otherwise.} \end{cases}$$

ORS is asymptotically optimal (minimizes regret)
Its performance does not depend on the number of possible rates!

For non-stationary environments: SW-ORS (ORS with sliding window)

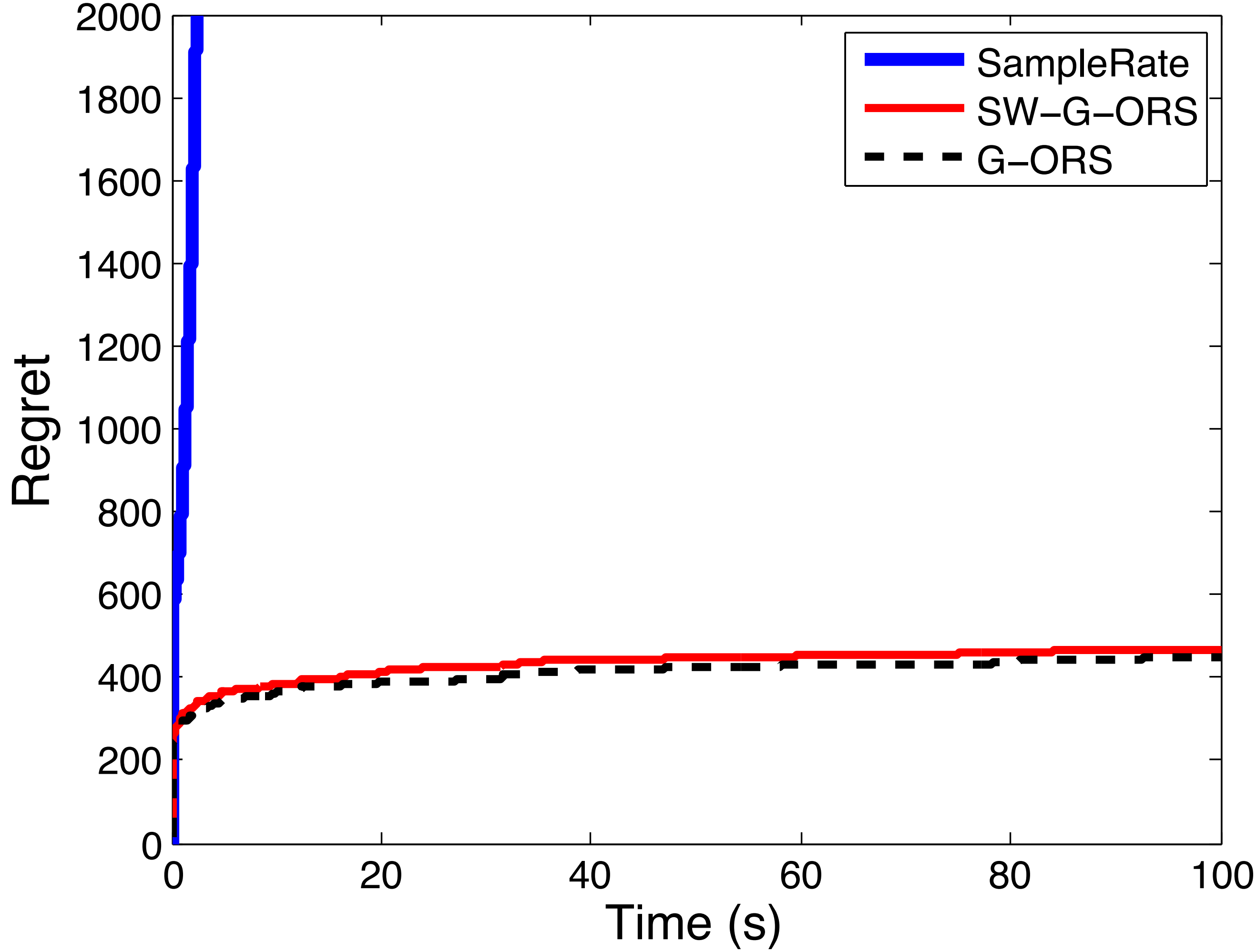
802.11g – stationary environment

GRADUAL (success prob. smoothly decreases with rate)



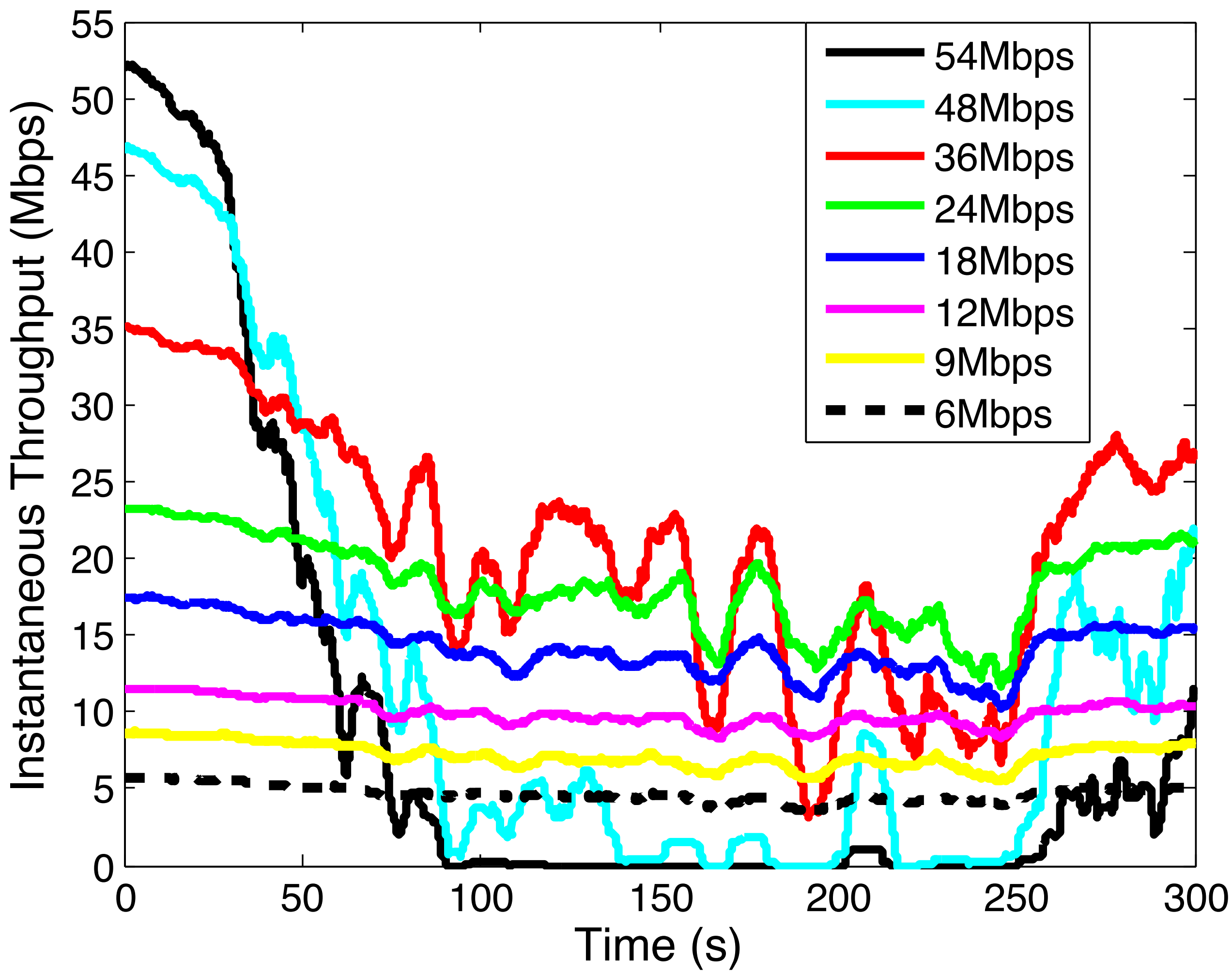
802.11g – stationary environment

STEEP (success prob. is either close to 1 or to 0)



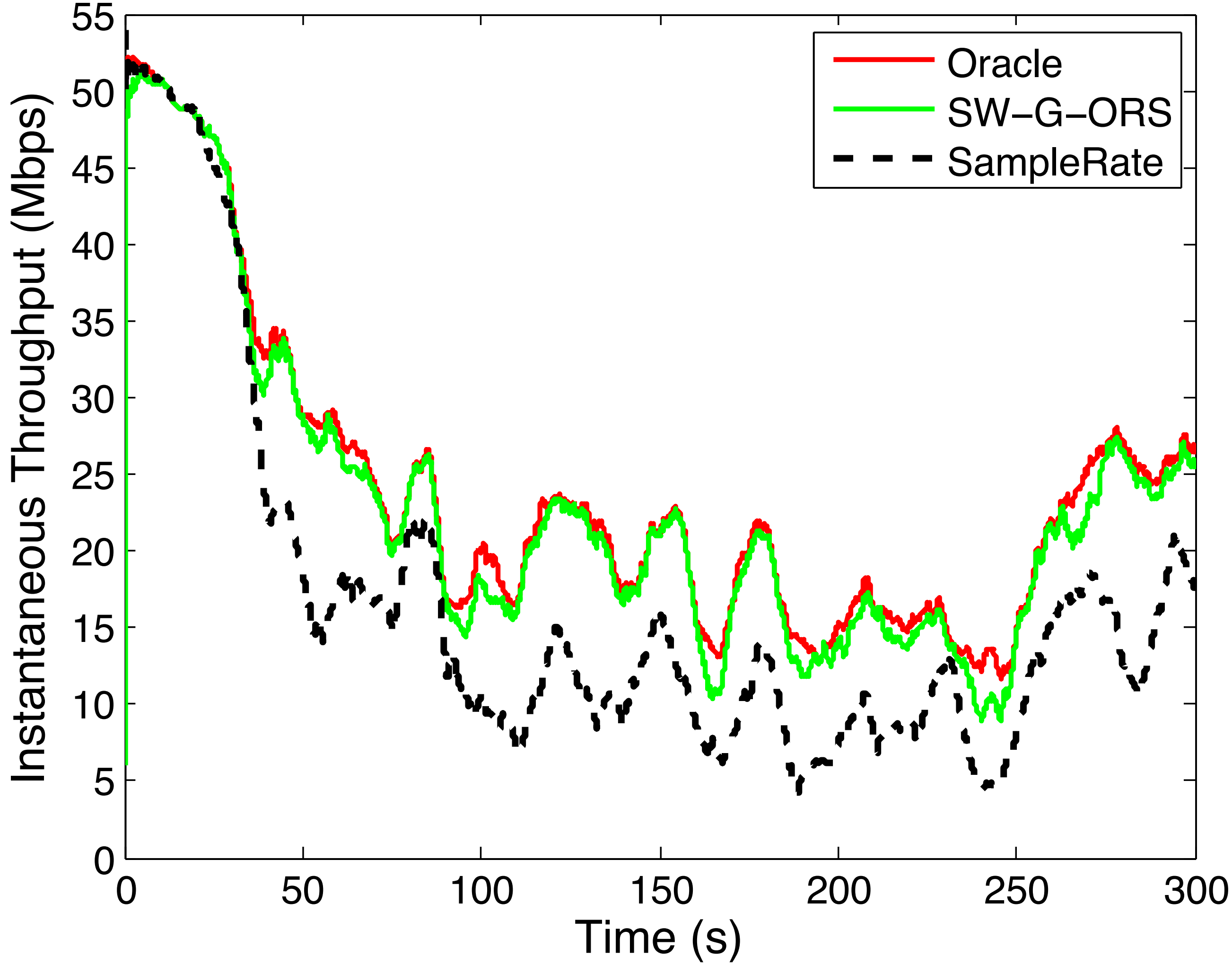
802.11g – non-stationary environment

TRACES



802.11g – non-stationary environment

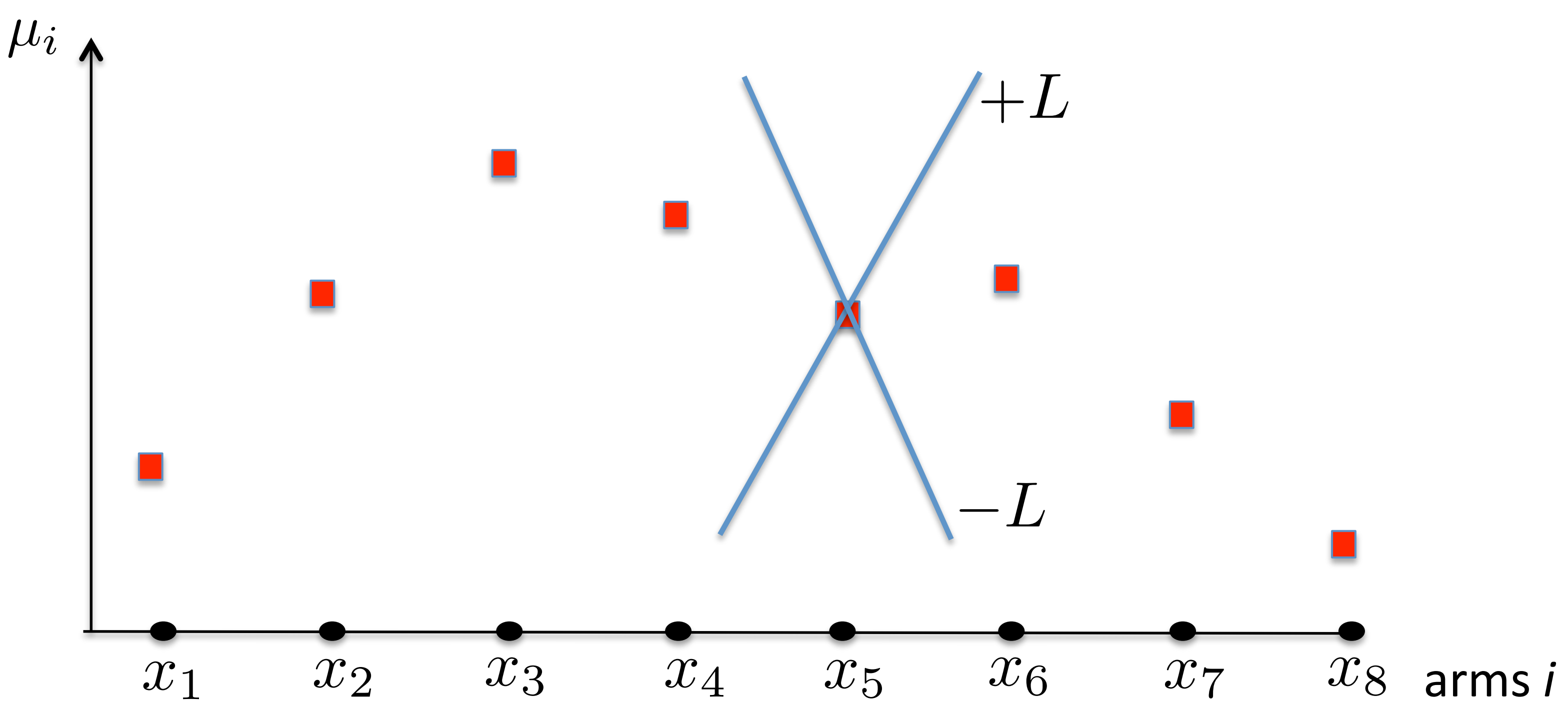
RESULTS



2-A.2. Discrete Lipschitz Bandits

Combes, Magureanu, Proutiere. Lipschitz Bandits: Regret Lower Bounds and Optimal Algorithms, **COLT** 2014

Discrete Lipschitz Bandits



Let $x_1 < x_2 < \ldots < x_K$ denote the *positions* of the arms.
We assume that: $|\mu_k - \mu_{k'}| \leq L \times |x_k - x_{k'}|$.

Related work

- Continuous set of actions (e.g. [0,1]): **Agrawal** 1995, **Kleinberg** 2004, **Kleinberg-Slivkins-Upfal** 2008, **Bubeck-Munos-Stolz-Szepesvári** 2008, …

Regret lower bound

Theorem: For any uniformly good algorithm π

$$\lim \inf_{T \to \infty} \frac{R^{\pi}(T)}{\log(T)} \geq C(\theta)$$

where $C(\theta)$ is the minimal value of:

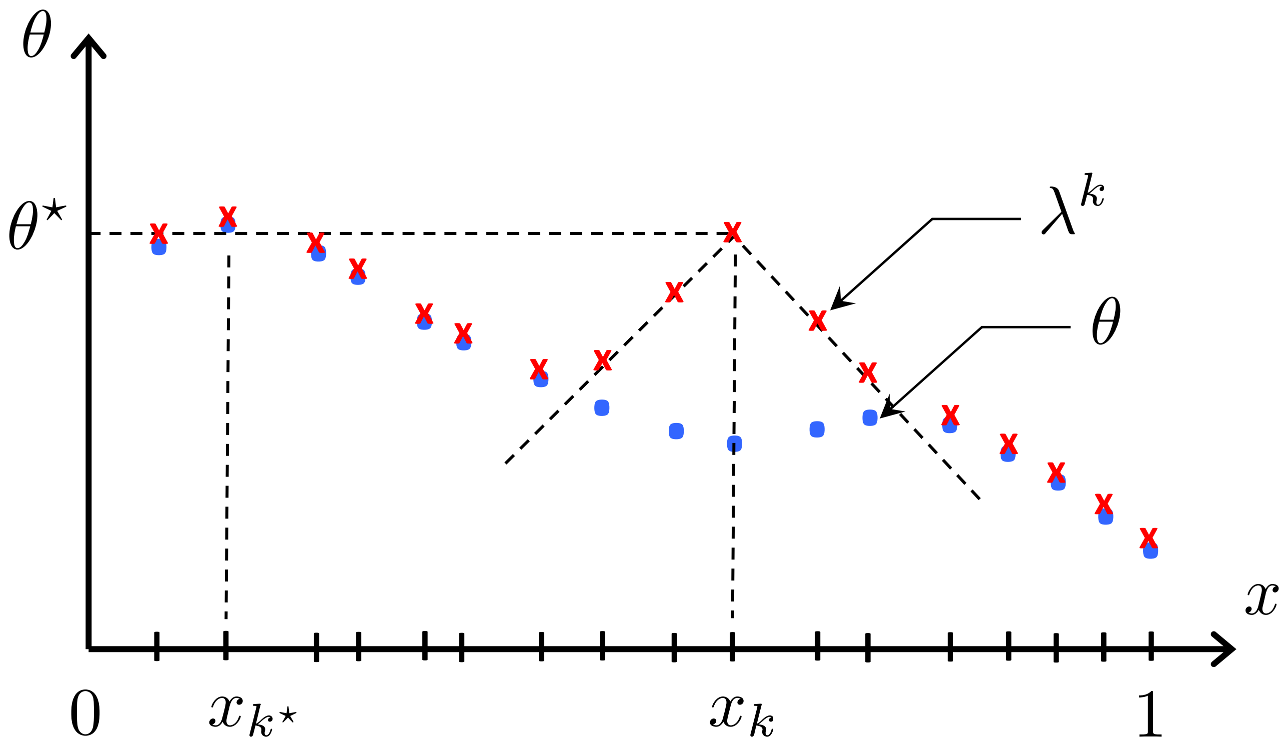
$$\min_{c_k \geq 0, \forall k \in \mathcal{K}^-} \sum_{k \in \mathcal{K}^-} c_k \times (\theta^\star - \theta_k)$$

$$\text{s.t. } \forall k \in \mathcal{K}^-, \sum_{i \in \mathcal{K}} c_i I(\theta_i, \lambda_i^k) \geq 1.$$

Regret lower bound

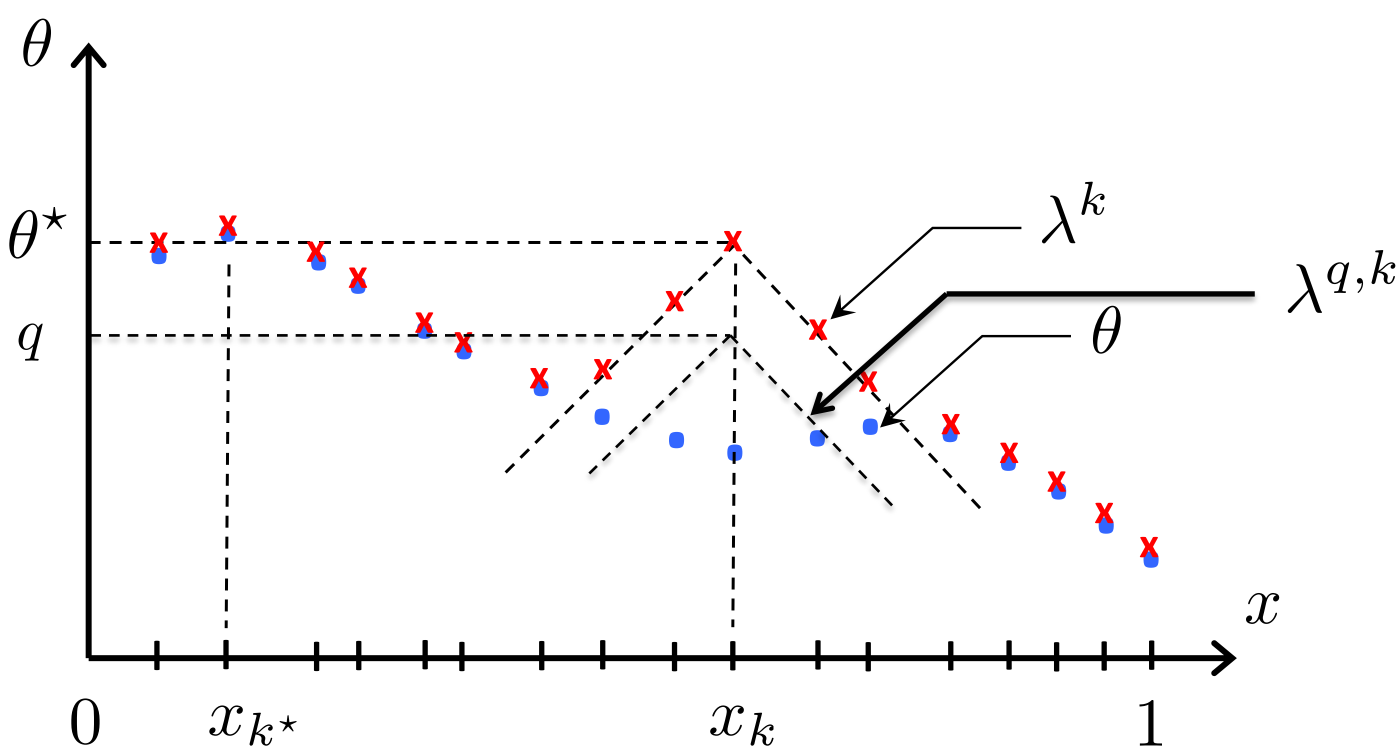
$$\min_{c_k \geq 0, \forall k \in \mathcal{K}^-} \sum_{k \in \mathcal{K}^-} c_k \times (\theta^\star - \theta_k)$$

$$\text{s.t. } \forall k \in \mathcal{K}^-, \ \sum_{i \in \mathcal{K}} c_i I(\theta_i, \lambda_i^k) \geq 1.$$



Algorithms

$$b_k(n) = \sup\{q \in [\hat{\theta}_k(n), 1] : \sum_{k'=1}^{K} t_{k'}(n) I^+(\hat{\theta}_{k'}(n), \lambda_{k'}^{q,k}) \le \log(n) + 3\log\log(n)\}.$$



The OSLB algorithm

- *Apparently* optimal arm sampling rate. Regret lower bound replacing θ by $\hat{\theta}(n)$: $c_k(n)$
- Set of arms apparently under-sampled:

$$\mathcal{K}_e(n) = \{k \in \mathcal{K}^-(n) : t_k(n) \leq c_k(n)\log(n)\}$$

$$\overline{k}(n) = \arg\min_{k \in \mathcal{K}_e(n)} t_k(n)$$

$$\underline{k}(n) = \arg\min_{k} t_k(n)$$

Algorithm -- OSLB

Select the leader if $\hat{\theta}_{L(n)}(n) \geq \max_{k \neq L(n)} b_k(n)$

Else

if $t_{\underline{k}(n)}(n) < \frac{\epsilon}{K} t_{\overline{k}(n)}(n)$, select $\underline{k}(n)$

else select $\overline{k}(n)$

A Simplified Algorithm

Algorithm -- CKL-UCB

Select the leader if it has the highest index

Else select the least explored arm with an index higher than the leader

Regret under OSLB and CKL-UCB

Theorem: For any $\theta \in \Theta_L$, under $\pi = \mathrm{OSLB}(\epsilon)$, we have:
For all $\delta > 0$, and all T,

$$R^\pi(T) \le C^\delta(\theta)(1+\epsilon)\log(T) + C_1 \log\log(T) + K^3\epsilon^{-1}\delta^{-2} + 3K\delta^{-2}$$

where $C^\delta(\theta) \to C(\theta)$ as $\delta \to 0^+$.

Theorem: For any $\theta \in \Theta_L$, under $\pi = \mathrm{CKL\text{-}UCB}$, we have:

$$\lim\sup_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \le C'(\theta),$$

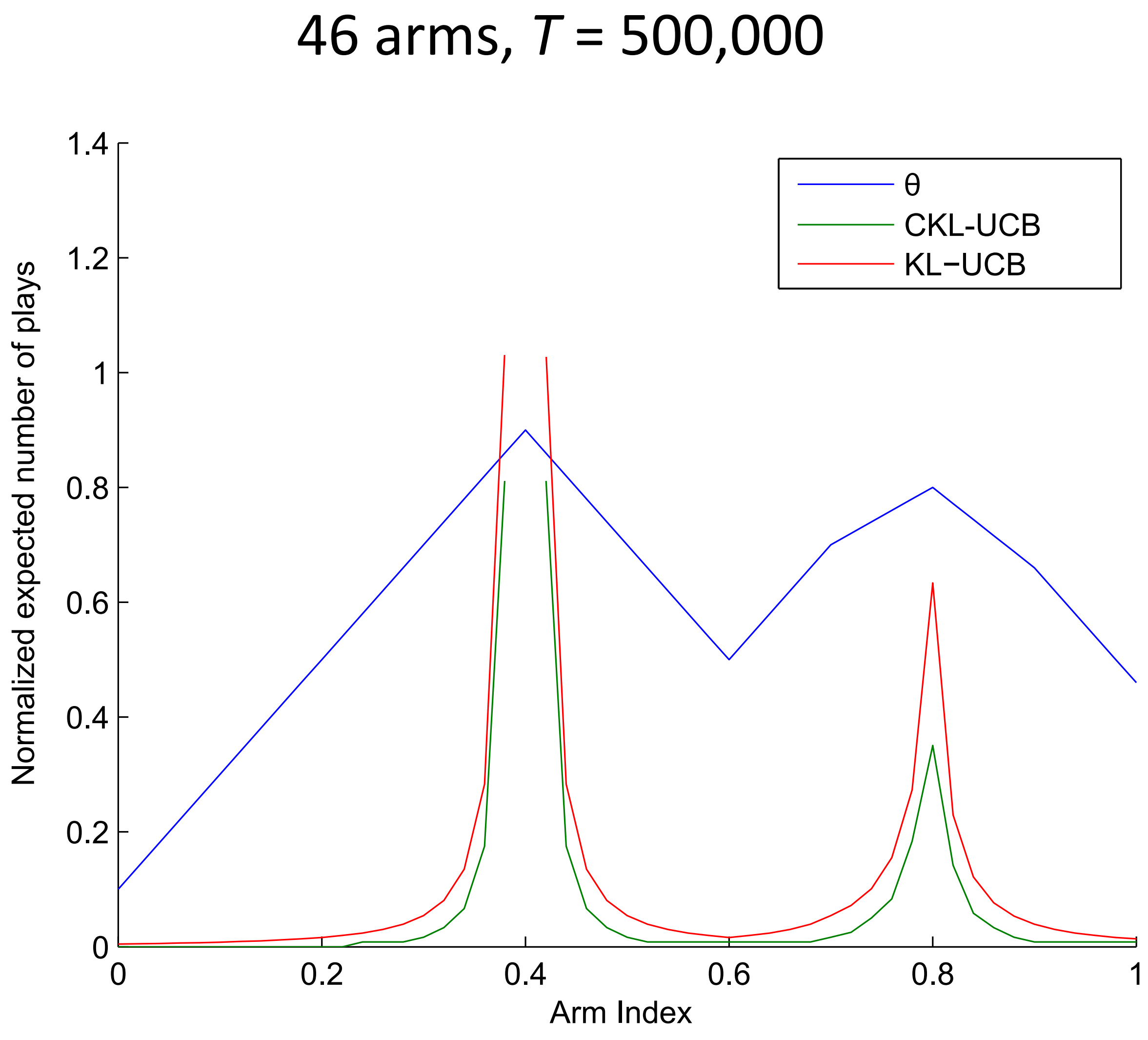
where $C'(\theta)$ is the minimal value of an optimization problem "close" to that providing the regret lower bound.

Proof ingredients

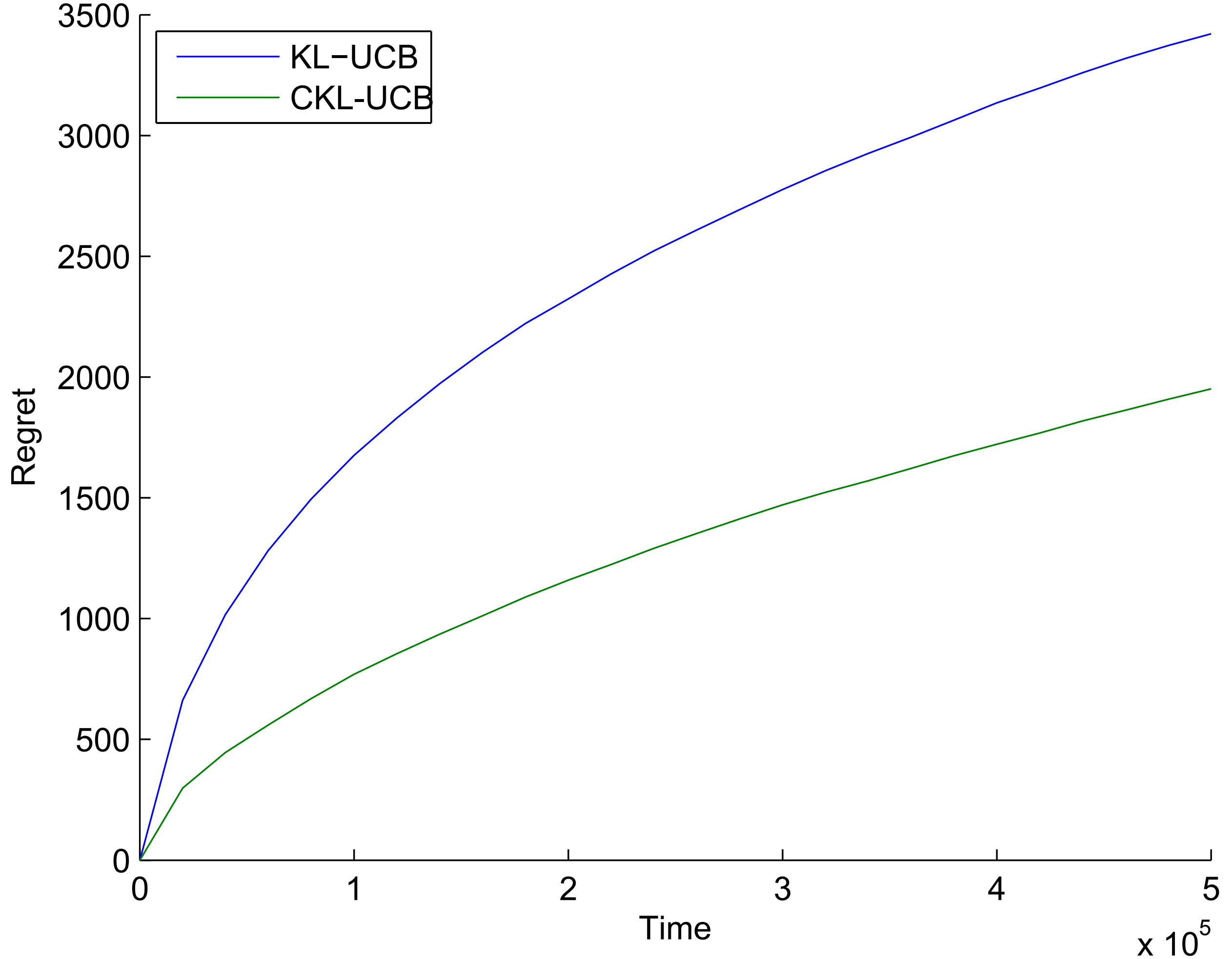
A concentration inequality for the sum of KL divergences:

$$\mathbb{P}\left[\sum_{k=1}^{K} t_k(n) I^+(\hat{\theta}_k(n), \theta_k) \geq \delta\right] \leq e^{-\delta} \left(\frac{\lceil \delta \log(n) \rceil \delta}{K}\right)^K e^{K+1}.$$

Example


46 arms, T = 500,000
Normalized expected number of plays
Arm Index
θ
CKL-UCB
KL−UCB
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

Example



KL−UCB
CKL-UCB
Regret
Time
x 10^5
0
500
1000
1500
2000
2500
3000
3500
0
1
2
3
4
5

Summary: Discrete Structured Bandits

- Regret lower bounds by Graves-Lai 1997: works for any structure
  - When is the solution explicit?
  - How does it scale with the dimension of the decision space?
  - When explicit, provides guidelines on the design of optimal algorithms – optimally exploiting the known structure
- Simple and efficient algorithm: Unimodal, and Lipschitz
- Other structures? Linear, Convex?
- Thompson Sampling
  - Is it always asymptotically optimal?
  - How to sample for the posterior?
- Complexity vs. Performance?

2-B. Infinite Bandits

Bonald, Proutiere. Two-Target Algorithm for Infinite-Armed Bandits, **NIPS** 2013

Actions and rewards

- An infinite number of Bernoulli arms
- Decision in each round: take a new arm, or play arms previously selected
- Bayesian setting: the expected reward θ_k of the *k*-th selected arm follows a *known* distribution

$$F(u) = \mathbb{P}[\theta_k > u]$$

$$F(u) \sim \alpha(1-u)^{\beta}, \quad \text{as } u \to 1-$$

- Regret: $R(T) = T - \mathbb{E}[\sum_{t=1}^{T} X_t]$

- More like a stopping time problem …

Related work

- **Mallows-Robbins** 1964, **Herschkorn-Pekoes-Ross** 1996: no-regret policies
- **Berry-Chen-Zame-Heat-Shepp** 1997: uniformly distributed parameter, policy with regret 2√T, conjectured to be optimal

  1-failure policy: keep the first arm with more than √T successive 1's

rewards	110	10	11110	11111110101011100…
arm	1	2	3	4

Related work

- **Mallows-Robbins** 1964, **Herschkorn-Pekoes-Ross** 1996: no-regret policies
- **Berry-Chen-Zame-Heat-Shepp** 1997: uniformly distributed parameter, policy with regret 2√T, conjectured to be optimal

  1-failure policy: keep the first arm with more than √T successive 1's

rewards	110	10	11110	11111110101011100…
arm	1	2	3	4

1-failure policies are actually sub-optimal …

Related work

- **Wang-Audibert-Munos** 2013: More general parameter distribution, regret scaling as $T^{\beta/(\beta+1)}$ up to log factors.

  Policy: select X arms and run UCB …

  Not a stopping rule. The number of arms tested does not depend on the realizations of the rewards.

Regret lower bound

Theorem: For any algorithm π knowing the time horizon,

$$\lim_{T\to\infty} \inf \frac{R^\pi(T)}{T^{\frac{\beta}{\beta+1}}} \geq \left(\frac{\beta+1}{\alpha}\right)^{\frac{1}{\beta+1}}$$

Conjecture: When the time horizon is unknown,

$$\lim_{T\to\infty} \inf \frac{R^\pi(T)}{T^{\frac{\beta}{\beta+1}}} \geq \frac{\beta+1}{\beta}\left(\frac{\beta}{\alpha}\right)^{\frac{1}{\beta+1}}$$

Example: parameter unif. distributed, $\sqrt{2T}, \quad 2\sqrt{T}.$

Two-target algorithms

Exploration of arm *k*:

Run 1 *Run2* *Run 3* *Run m+1*

11111111110 11110 111111110 ……. 111111111110

$\longleftrightarrow$ L_1 $\longleftrightarrow$ L_2

If $L_1 < \ell_1$, explore a new arm

Else if $L_2 < \ell_2$, explore a new arm

else keep it forever

Two-target algorithms

Theorem: Select $\ell_1 = \left\lfloor \left(\frac{\alpha n}{\beta+1}\right)^{\frac{1}{\beta+2}} \right\rfloor, \quad \ell_2 = \left\lfloor m \left(\frac{\alpha n}{\beta+1}\right)^{\frac{1}{\beta+1}} \right\rfloor.$

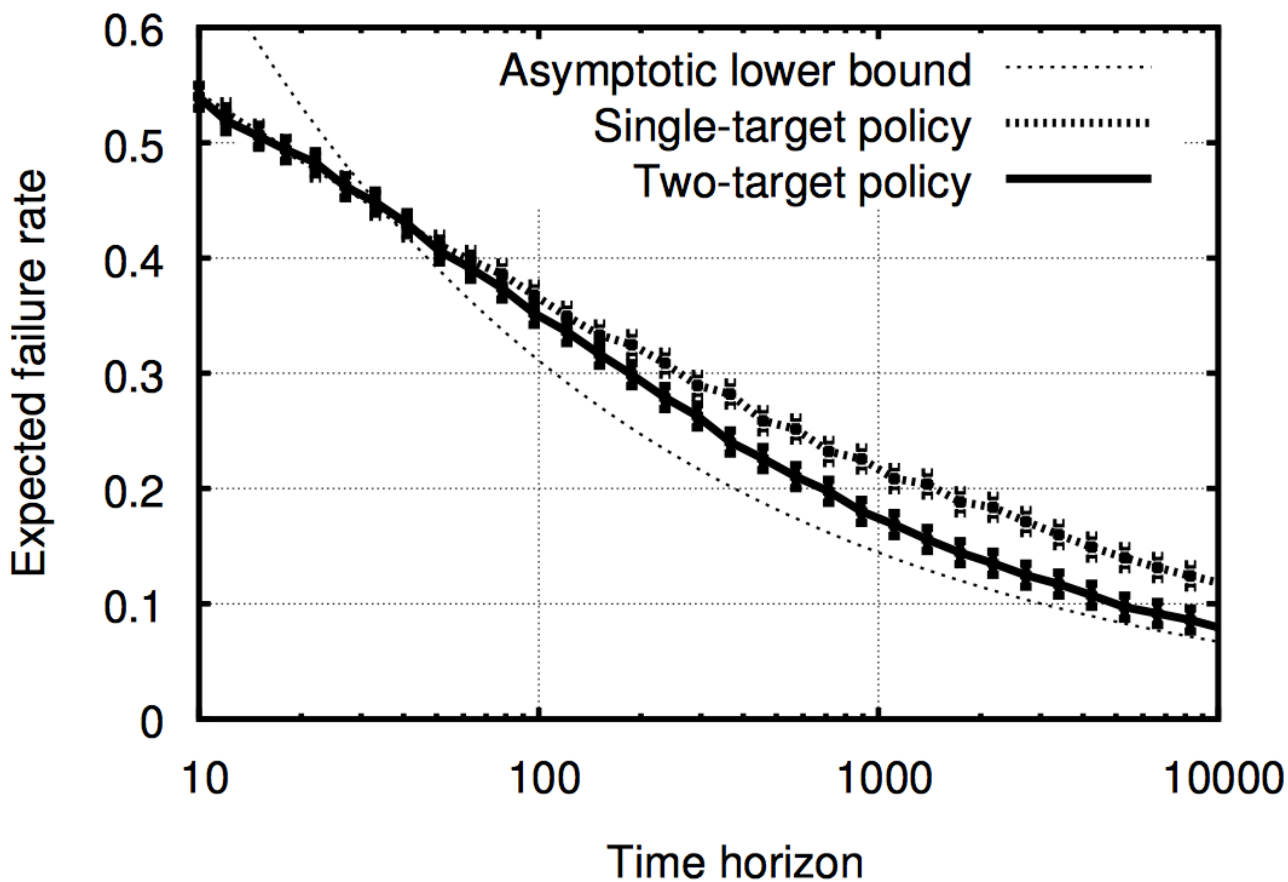
$$\lim \sup_{T\to\infty} \frac{R^\pi(T)}{T^{\frac{\beta}{\beta+1}}} \leq \left(\frac{\beta+1}{\alpha}\right)^{\frac{1}{\beta+1}} \left(1 + O(\frac{1}{m})\right).$$

Example: parameters for unif. distribution,

$$\ell_1 \sim (n/2)^{1/3}, \quad \ell_2 \sim m\sqrt{n/2}.$$

Numerical Example

- Beta(1,2) mean reward distribution
- Expected failure rate = mean regret per round



Summary: Infinite Bandits

- Regret lower bound and optimal algorithms when the support of the reward distribution is 1, and the time horizon is known
- What about unknown time horizon?
- What if the support of the reward distribution does not include 1?
- What if the reward distribution is only partially known?

2-C. Continuous Structured Bandits

Continuous Structured Bandits

- Set of arms: $[0, 1]$
- Bernoulli reward for arm x of mean $\mu(x)$
- Reward realizations: $(X_n(x), n \geq 1)$ i.i.d. over time, independent over arms
- Algorithm π: selects arm $x^\pi(n)$ in round n
- Bandit feedback: $X_n(x^\pi(n))$
- Regret:

$$R^\pi(T) = T\mu^\star - \sum_{n=1}^{T} \mu(x^\pi(n))$$

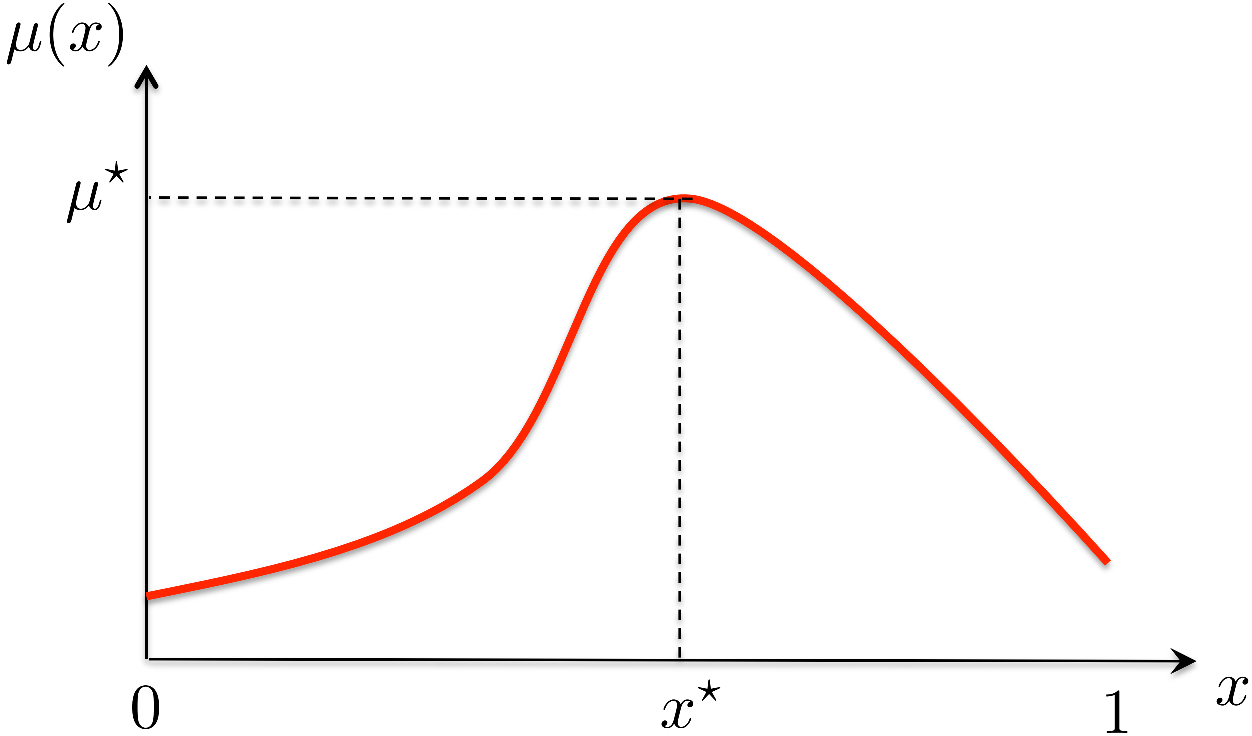
$$\mu^\star = \sup_{x \in [0,1]} \mu(x) = \mu(x^\star)$$

- Structure: $x \mapsto \mu(x)$ is unimodal, linear, concave, Lipschitz, …

2-C.1. Continuous Unimodal Bandits

Combes, Proutiere. Unimodal Bandits without Smoothness, **arxiv** 2014

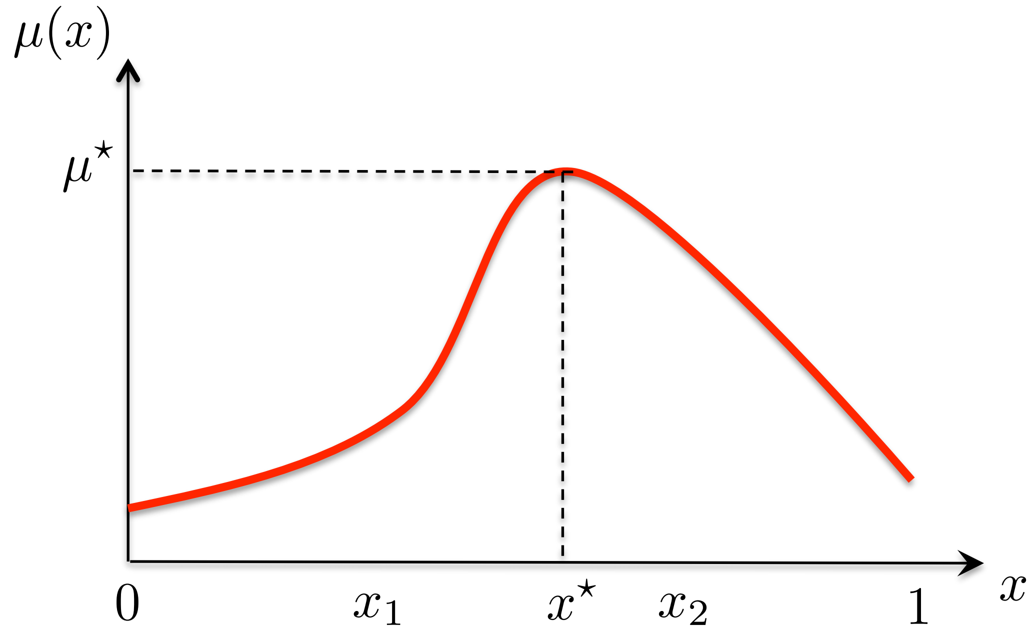
Continuous Unimodal Bandit



The mapping $x \mapsto \mu(x)$ is unimodal.

Golden Section Algorithm

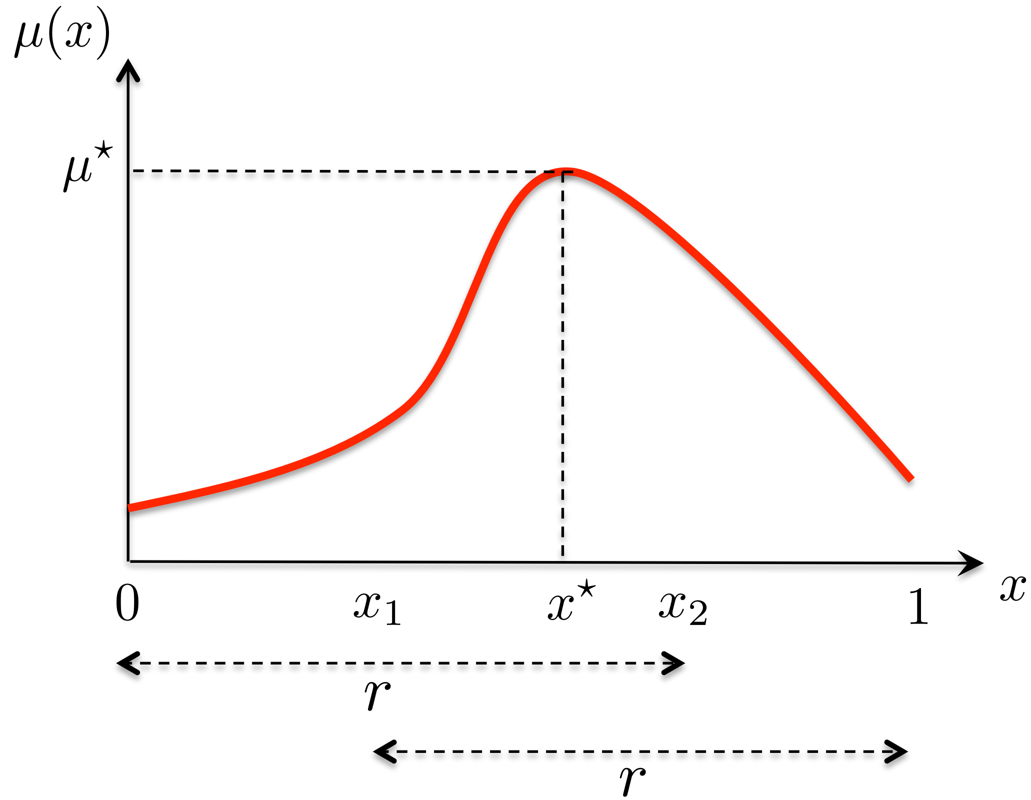
Kiefer 1953



- Deterministic setting
- Evaluate the function in points x_1, x_2
- If $\mu(x_1) < \mu(x_2)$, keep $[x_1, 1]$, else keep $[0, x_2]$
- Design choices: (i) the ratio of the lengths of the old and new new intervals is always r and (ii) we need to evaluate the function once in each step

Golden Section Algorithm

Kiefer 1953



$$\frac{1}{r} = \frac{r}{1-r} \Longrightarrow r = \frac{-1+\sqrt{5}}{2} \approx 0.618$$

Stochastic Setting – Related Work

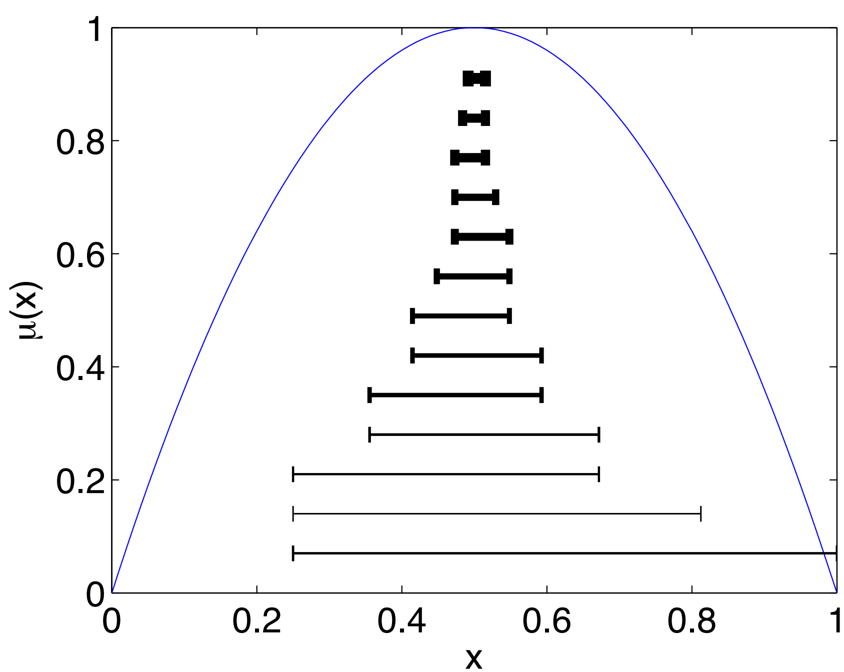
- Smoothness assumption:

$$|\mu(x) - \mu(x^\star)| \overset{x \to x^\star}{\sim} C|x - x^\star|^\alpha, \quad \alpha > 0$$

- Regret lower bound (Dani et al. 2008 – linear): $\Omega(\sqrt{T})$
- Existing approaches yielding a regret $\tilde{O}(\sqrt{T})$
  - Kleinberg 2004: discretization with step $(\log(T)/\sqrt{T})^{1/\alpha}$
  - Coppe 2009: stochastic gradient, works for $\alpha \geq 2$ only
  - Yu-Mannor 2011: stochastic version of the golden section algorithm, assume the knowledge of α, C
- Without any knowledge on the function smoothness: interval trimming algorithm yielding a regret $\tilde{O}(\sqrt{T})$, Combes-Proutiere 2014

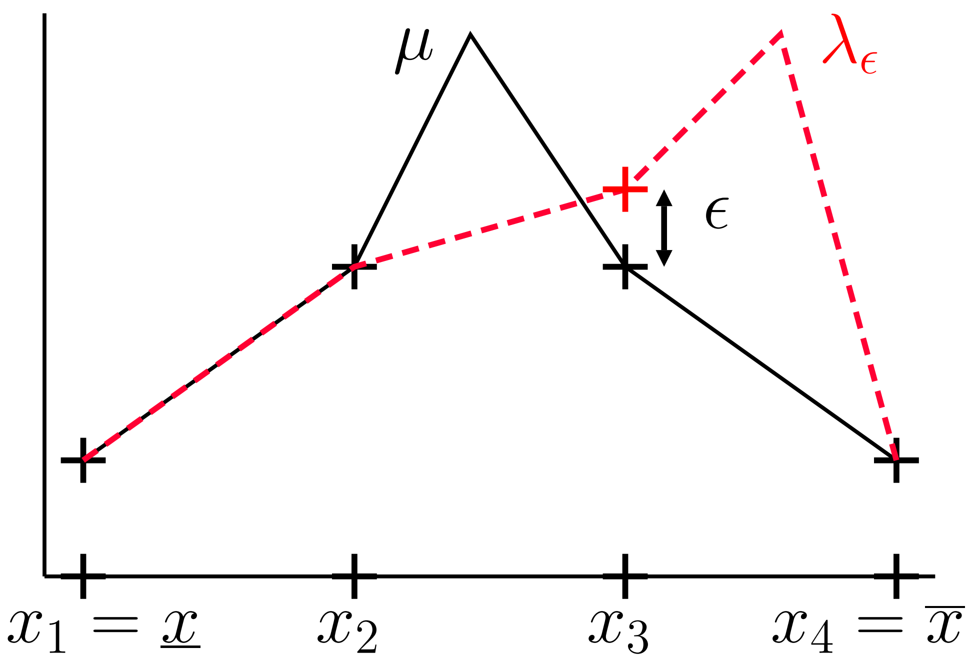
Interval Trimming

- Idea: construct a sequence of intervals $I^T \subset \ldots \subset I^0 = [0,1]$ with $x^\star \in \cap_{t=0}^{T} I^t$ with high probability
- Step *t*: start with $I^t = [\underline{x}, \overline{x}]$
  - Sample the function at *K* points $\underline{x} \leq x_1 \leq \ldots \leq x_K \leq \overline{x}$ until enough information is gathered to eliminate either the left or right part of I^t



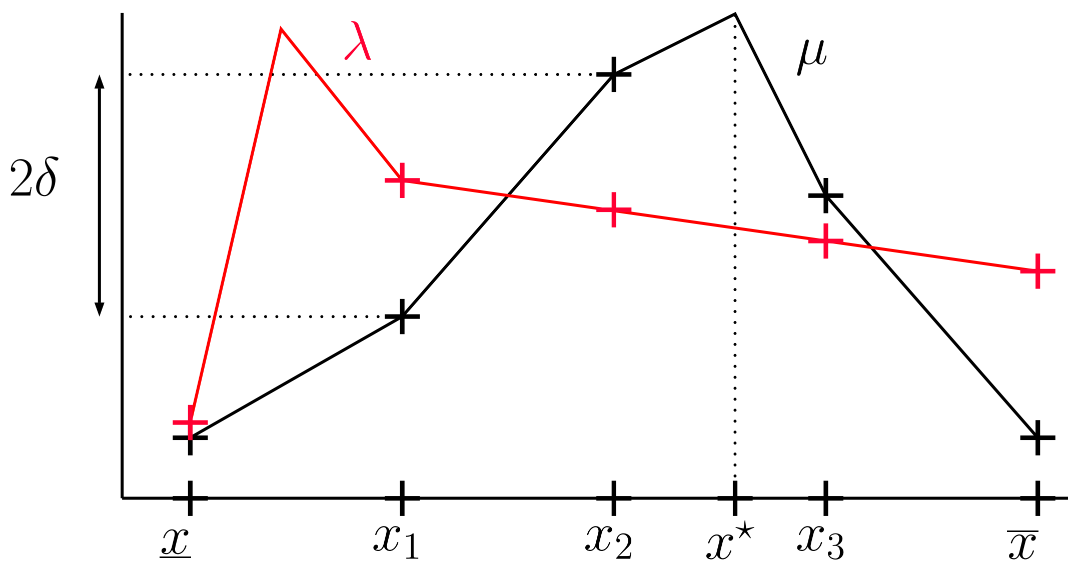
The Failure of Golden Section Algorithm -- Unknown Smoothness

- We need to sample at least 3 arms in the interior of the interval to be trimmed to guarantee that $x^\star \in \cap_{t=0}^{T} I^t$ with high probability



Optimal Interval Trimming

- Sample 3 points in the interior of the interval $x_1 < x_2 < x_3$
- If $x^\star > x_2$, and $\mu(x_1) < \mu(x_2)$ -- remove $[\underline{x}, x_1]$
- If $x^\star < x_1$, and $\mu(x_3) < \mu(x_2)$ -- remove $[x_3, \overline{x}]$
- Sample long enough until $\hat{\mu}(x_2) - \hat{\mu}(x_1)$ or $\hat{\mu}(x_2) - \hat{\mu}(x_3)$ is large enough



Optimal Interval Trimming

- Location test:

$$KL^\star(\mu_1, \mu_2) = 1_{\mu_1 < \mu_2} \left[KL(\mu_1, \frac{\mu_1 + \mu_2}{2}) + KL(\mu_2, \frac{\mu_1 + \mu_2}{2}) \right]$$

- Sample x_1, x_2, x_3 in a round robin fashion
- Stop when there is $m \in \{1, 3\}$ such that:

$$\underline{t}(n) KL^\star(\hat{\mu}_m(n), \hat{\mu}_2(n)) \geq \log(T)$$

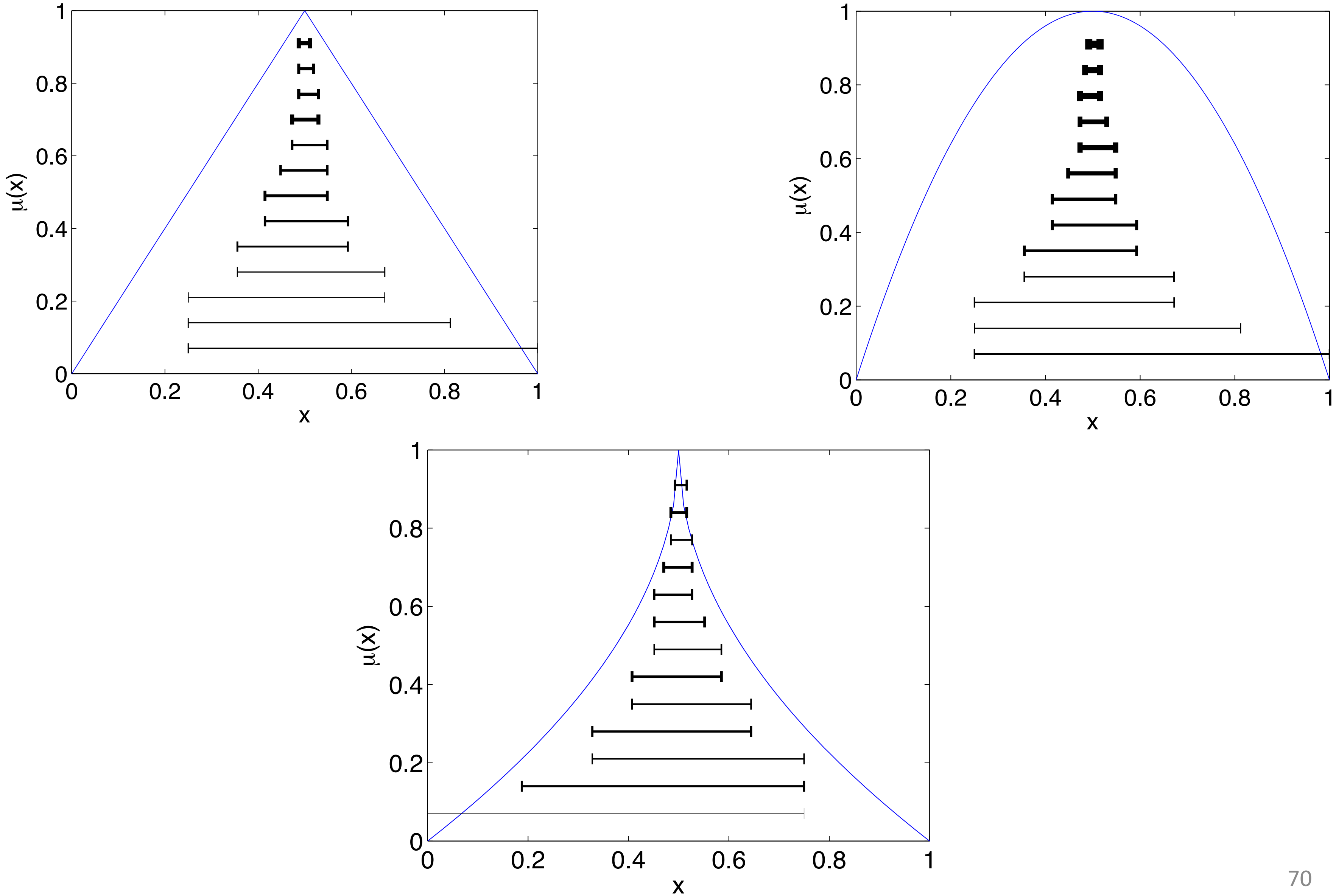
- If $m = 1$, remove $[\underline{x}, x_1]$
- If $m = 3$, remove $[x_3, \overline{x}]$

Performance

Theorem: Let $\delta = \mu(x_2) - \mu(x_1)$ if $x^\star \geq x_2$, and $\delta = \mu(x_2) - \mu(x_3)$ otherwise. The interval trimming procedure has length $O(\delta^{-2}\log(T))$ and risk $O(T^{-1})$.

Theorem: Assume $|\mu(x) - \mu(x^\star)| \overset{x \to x^\star}{\sim} C|x - x^\star|^\alpha, \quad \alpha > 0$. Then the proposed algorithm has regret $O(\sqrt{T\log(T)})$.

Examples



2-C.2. Continuous Lipschitz Bandits

Related work

- Continuous set of actions (e.g. [0,1]): **Agrawal** 1995, **Kleinberg** 2004, **Kleinberg-Slivkins-Upfal** 2008, **Bubeck-Munos-Stolz-Szepesvári** 2008, …

- For continuous bandits, algorithms should
  1. Adapt the subset of arms to sample from
  2. Optimally exploit the Lipschitz structure to select the arm based on ***all*** past observations

- Existing algorithms perform 1, but not 2. (for 2., simple UCB-like index are used …)

- Alternative approach: optimal algorithm for discrete bandits, and then optimal discretization of the set of arms

Zooming Algorithm

- **Kleinberg-Slivkins-Upfal** 2008

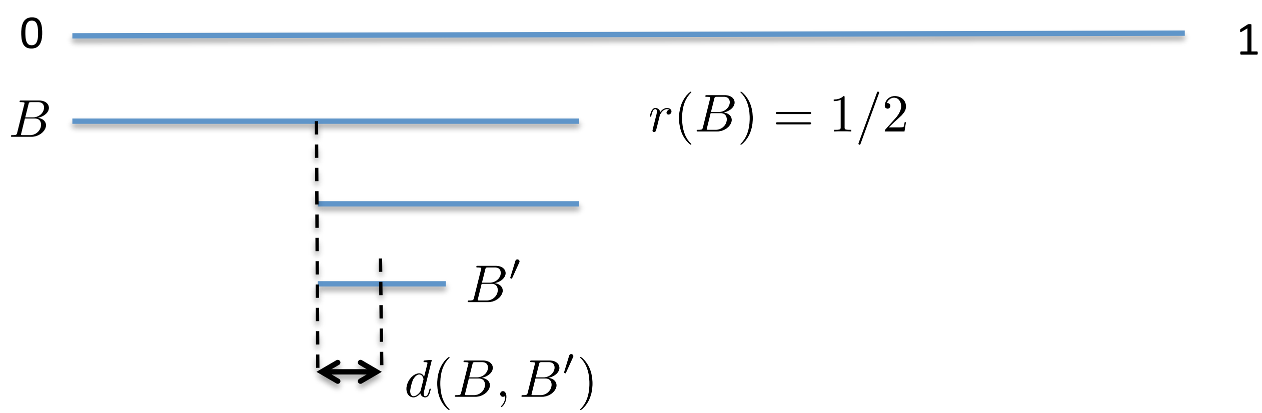
- Maintains a set of active balls: $\mathcal{A}_t$

$$\mathrm{conf}_t(B) = 4\sqrt{\frac{\log(T)}{1+n_t(B)}}$$

$$\mathrm{dom}_t(B) = B \setminus \cup_{B' \in \mathcal{A}_t : r(B') < r(B)} B'$$

Zooming Algorithm

- **Kleinberg-Slivkins-Upfal** 2008

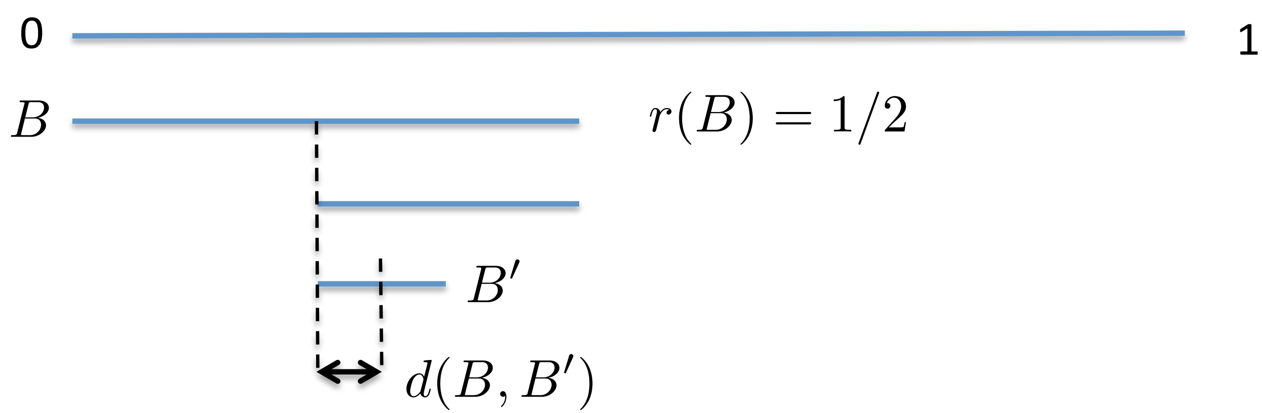


- Index of balls: $I_t(B) = r(B) + \min_{B' \in \mathcal{A}_t} (U_t(B') + d(B, B'))$

$$U_t(B) = \hat{\mu}_t(B) + r(B) + \mathrm{conf}_t(B)$$

Zooming Algorithm

- **Kleinberg-Slivkins-Upfal** 2008



- Algorithm:
  - Select a ball *B* with highest index and an arm *y* in *B*
  - If $\text{conf}_t(B) \leq r(B)$, activate the ball centered at *y* with radius $r(B)/2$

- Crude index, and sub-optimal structure exploitation

Optimal Discretized Algorithm

$$|\mu(x) - \mu(x^\star)| \overset{x \to x^\star}{\sim} C|x - x^\star|^\alpha, \quad \alpha > 0$$

Algorithm

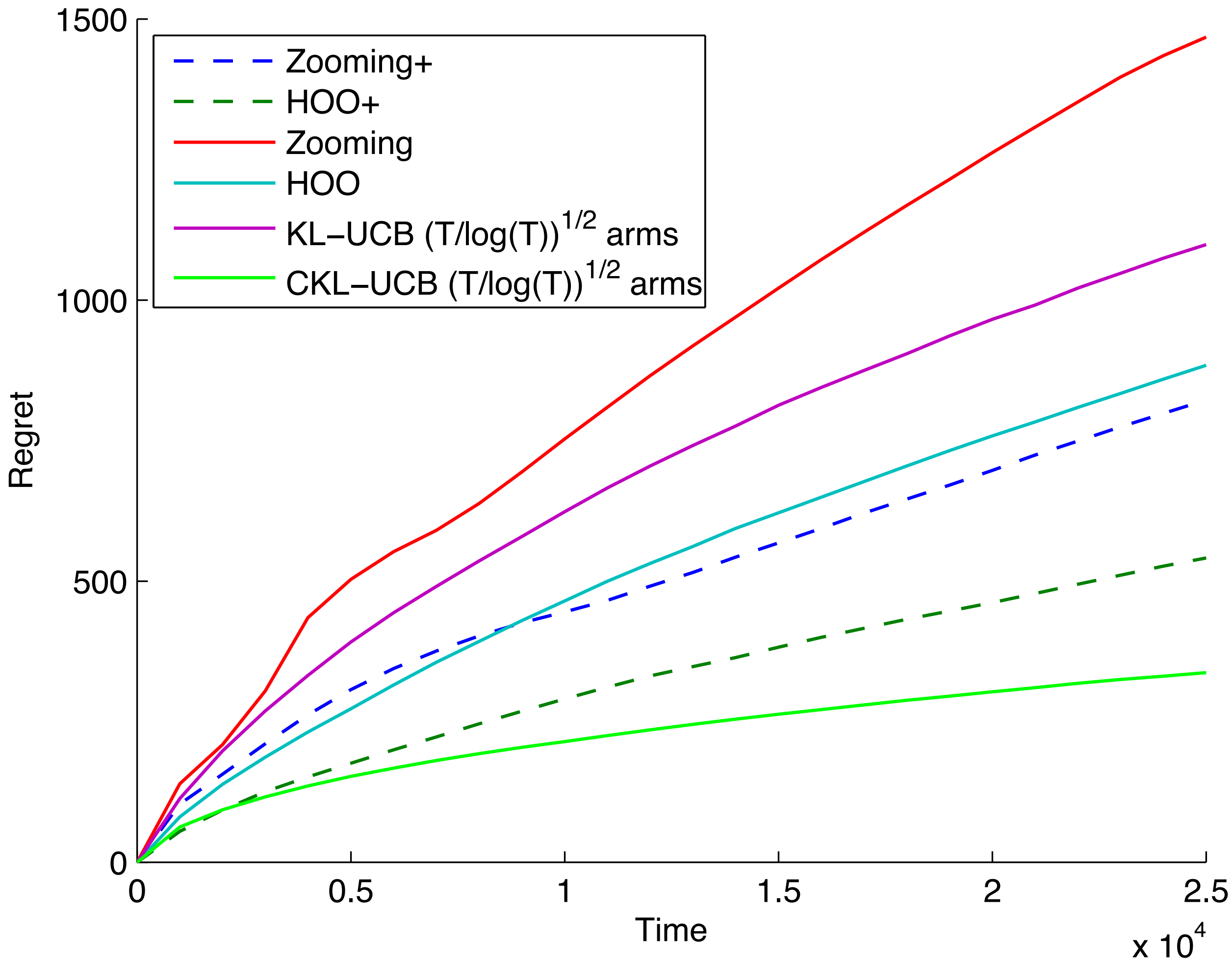
1. Discretization of the set of arms: step size $(\log(T)/\sqrt{T})^{1/\alpha}$
2. Apply discrete bandit algorithms

The above algorithm is order-optimal, as (discretization +KL-UCB), HOO algorithms, regret $\tilde{O}(T^{1/2})$

The zooming algorithm does not take the smoothness into account – in general sub-optimal, regret $\tilde{O}(T^{2/3})$

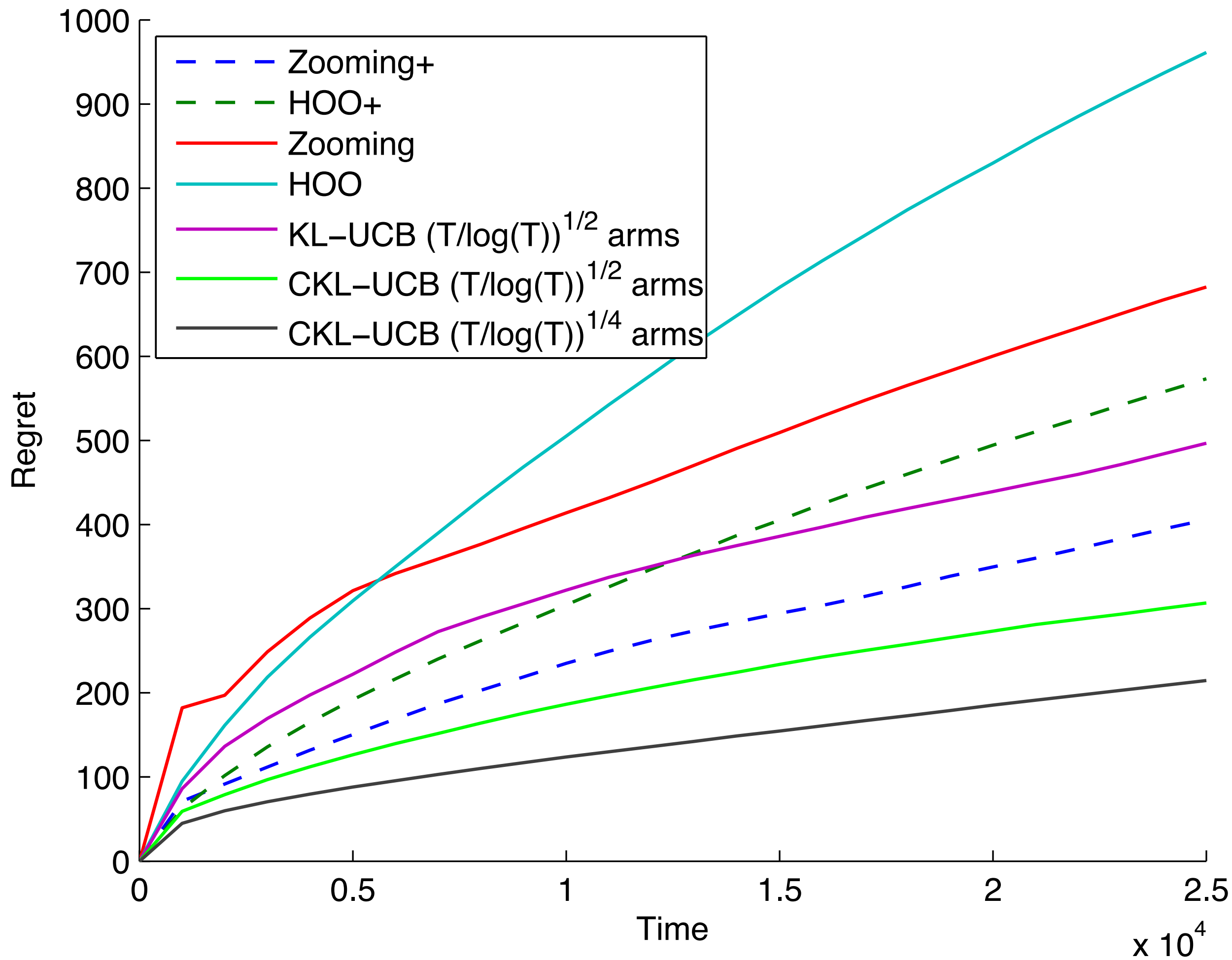
Example: Continuous set of arms

Triangular reward function



Example: Continuous set of arms

Quadratic reward function



Summary: Continuous Bandits

- State-of-the-art algorithms apply an appropriate discretization of the set of arms, and optimally exploit the structure
- Discretization: depends on the smoothness of the expected reward function
- Without smoothness: optimal location test + interval trimming approach
- No problem-specific regret lower bound

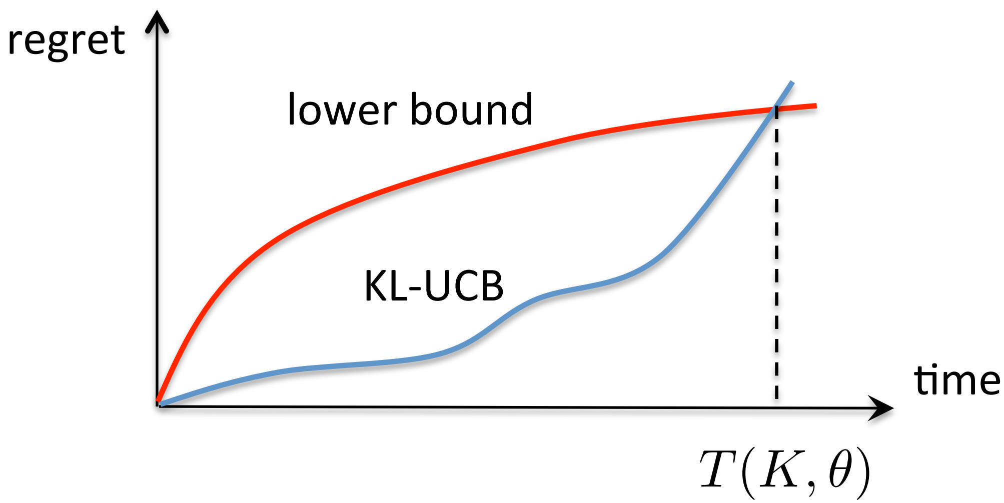
2-D. Conclusions and Open Problems

Conclusions: Stochastic Bandits

- Regret: the right performance metrics when dealing with uncertain and time-varying (non-stationary) environment
  - Tracking the best decision with minimum exploration cost
  - Many applications
- A well developed theory (essentially in the control and stat. communities, from the 70's to the late 90's)
- Further insights and new applications (ML community)
- Many open questions …

Anytime Regret Guarantees

- Classical unstructured discrete bandits: the asymptotic lower bound is not tight for small time horizons



$T(K,\theta)$ rapidly grow with *K*!

- Optimality for small time horizon?
- Preliminary result: **Guha** 2014 (COLT), Thompson sampling is 2-competitive for very specific problems

Discrete Structured Bandits

- Simple and yet asymptotically optimal algorithm for generic structure?
- Graves-Lai lower bound indicates the numbers of times sub-optimal arms should be selected
  - These numbers solve a complex optimization problem
  - … that we need to solve to get asymptotic optimality
  - What about the trade-off between complexity and regret?
- How does the lower bound scale with the number of arms?
- Example: combinatorial bandits (e.g. routing problems)
- Performance of Thompson sampling?

Continuous Structured Bandits

- Problem specific lower bounds?
- How to optimally exploit the structure? Linear, convex, and other structure?
- The optimal discretization depends on the structure and the smoothness of the expected reward function: is there an algorithm learning the structure and the smoothness?

Bibliography

- Graves and Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains, 1997
- Garivier and Moulines. On Upper-Confidence Bound Policies for Non-stationary Bandit Problems, 2011
- Agrawal. The Continuum-Armed Bandit Problem, 1995
- Kleinberg. Nearly tight bounds for the continuum-armed bandit problem, 2004
- Kleinberg, Slivkins, and Upfal, Multi-armed bandits in metric spaces, 2008
- Bubeck, Munos, Stoltz, Szepesvári. X-Armed Bandits, 2011
- Mallows, Robbins. Some Problems of Optimal Sampling Strategy, 1964
- Berry, Chen, Zame, Heath, and Shepp, Bandit problems with infinitely many arms, 1997
- Wang, Audibert, and Munos. Algorithms for infinitely many-armed bandits, 2008

Bibliography

- Kiefer. Sequential minimax search for a maximum, 1953
- Guha and Munagala. Stochastic Regret Minimization via Thompson Sampling, 2014

Thanks!

- Richard Combes: https://dl.dropboxusercontent.com/u/19365883/site/index.html
- Alexandre Proutiere: http://people.kth.se/~alepro/