

Structured Bandit Optimization

Alexandre Proutiere (KTH)

Lunteren Conference 2015

Outline

1. Bandit optimization
2. Graphically unimodal bandits
3. Applications and extensions

From joint work with Richard Combes (Supelec)

1. Bandit optimization

Bandit optimization

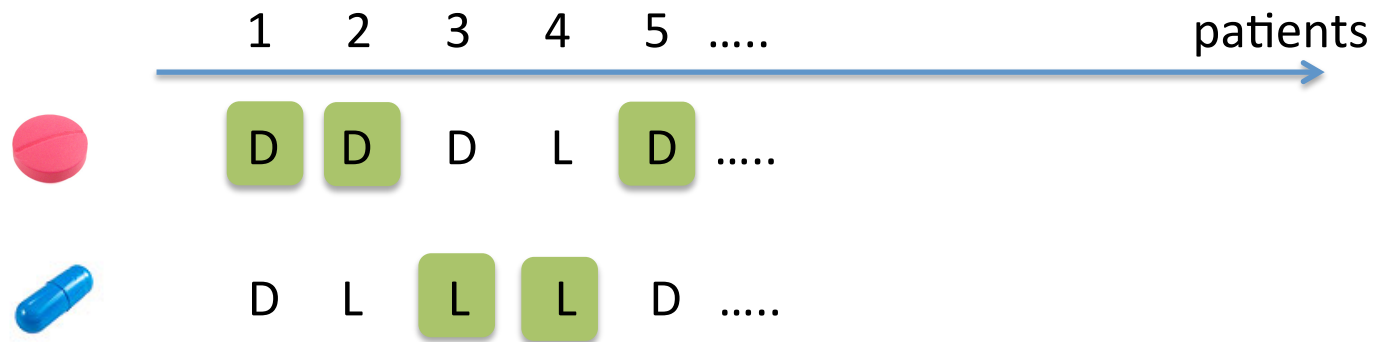
A sequential decision problem (**Thompson 1933**)



- A set of possible actions at each step
- Unknown sequence of rewards for each action

Bandit optimization

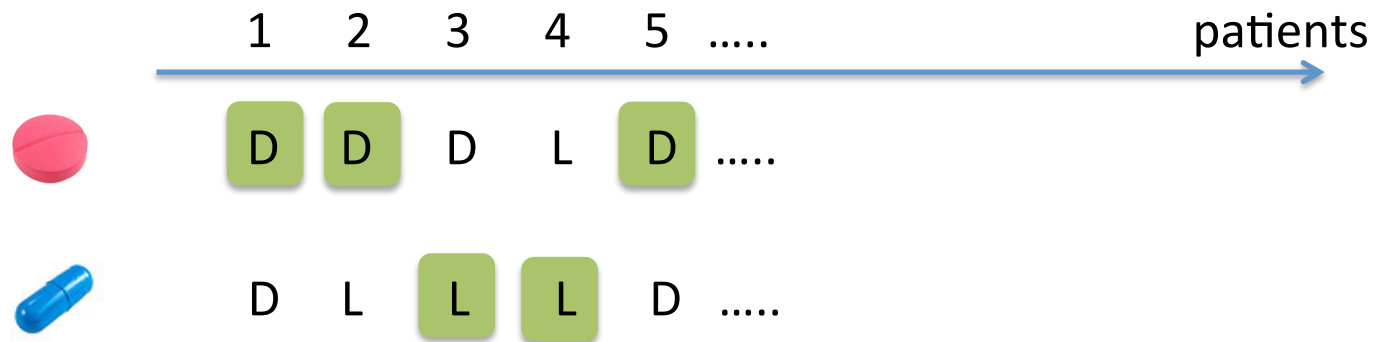
A sequential decision problem (**Thompson 1933**)



- A set of possible actions at each step
- Unknown sequence of rewards for each action
- Bandit feedback: only rewards of chosen actions are observed

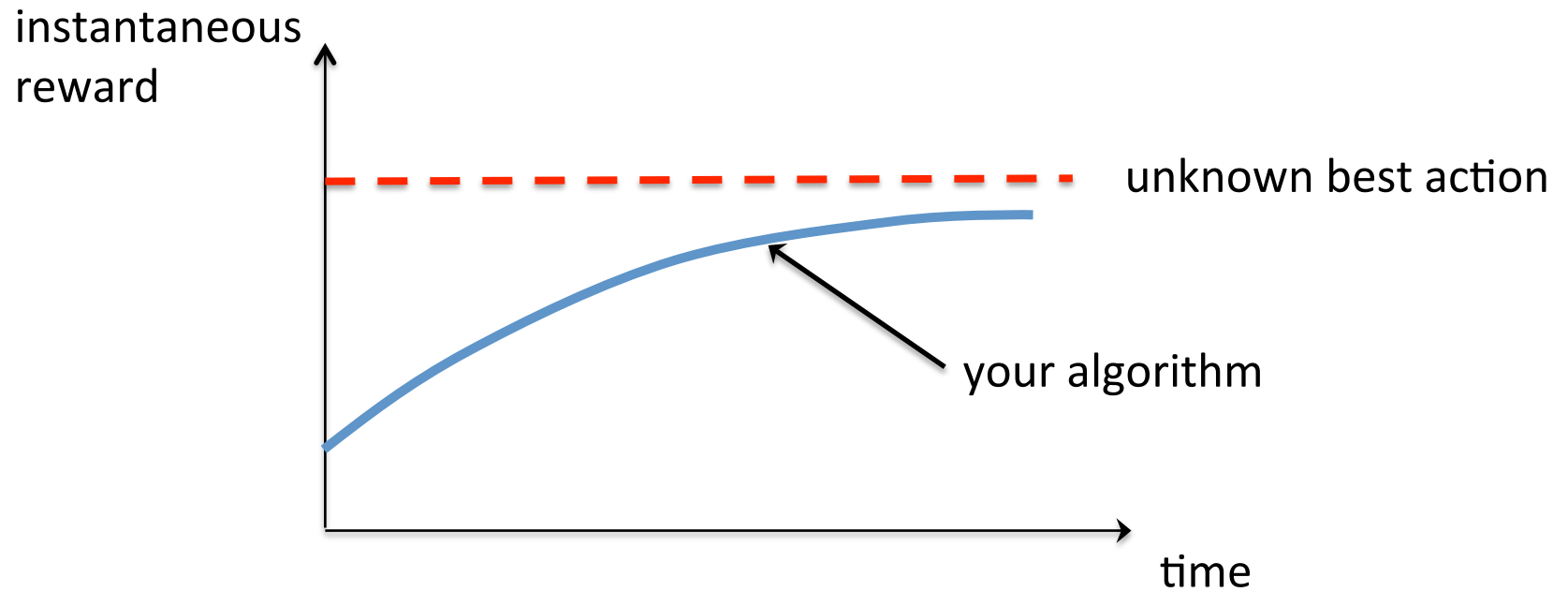
Bandit optimization

A sequential decision problem (**Thompson 1933**)

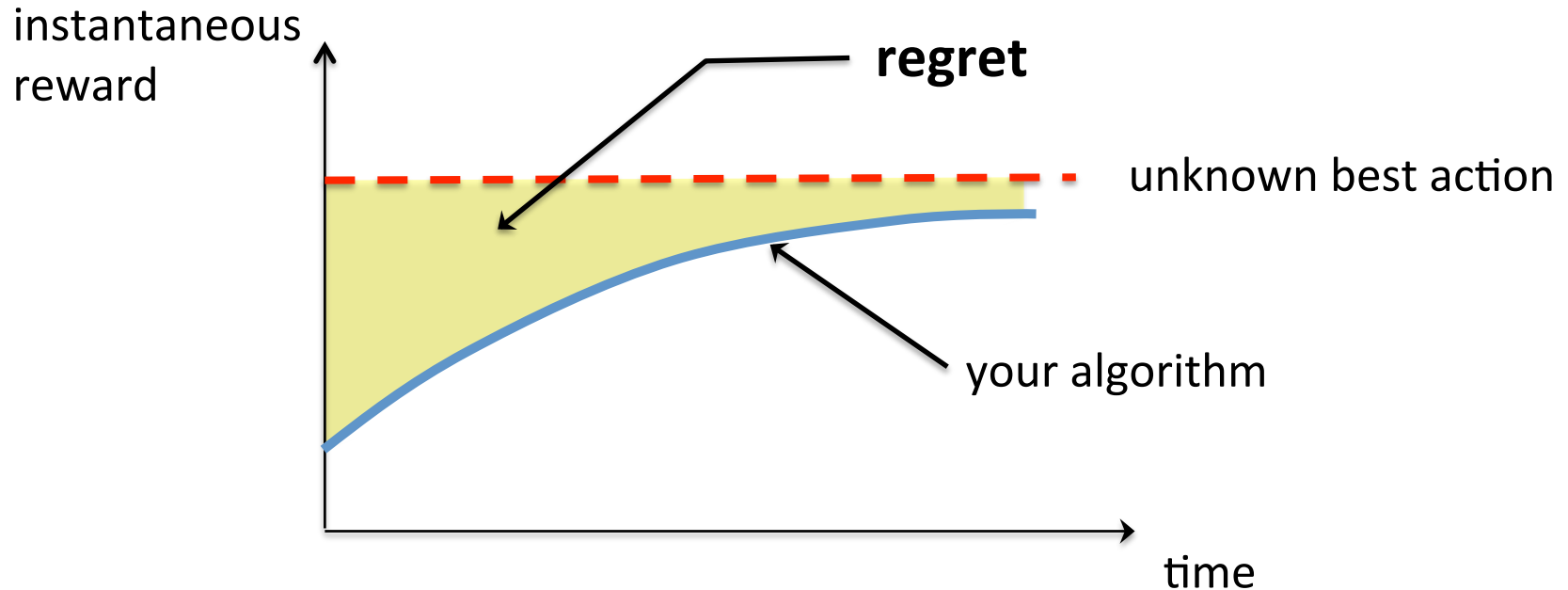


- A set of possible actions at each step
- Unknown sequence of rewards for each action
- Bandit feedback: only rewards of chosen actions are observed
- Goal: maximize the cumulative reward (up to step T), i.e., strike the optimal exploration-exploitation trade-off

Regret



Regret



Objective: to identify the best action with minimum exploration, i.e., to minimize regret (to maximize the “convergence rate”)

Particularly relevant when the best action evolves – for tracking problems

Stochastic Bandits

Robbins 1952

- K arms / decisions / actions
- Unknown i.i.d. rewards: $X_{i,t} \sim \text{Ber}(\mu_i)$, $\mu^* = \max_i \mu_i = \mu_{i^*}$
- Lack of structure: $\mu_i \in [0, 1]$, $\forall i \in \{1, \dots, K\}$
- Under online algorithm π , arm selected at time t :
 I_t^π function of history $(I_1^\pi, X_{I_1^\pi, 1}, \dots, I_{t-1}^\pi, X_{I_{t-1}^\pi, t-1})$
- Regret up to time T :

$$R^\pi(T) = \max_{i=1, \dots, K} \mathbb{E} \sum_{t=1}^T X_{i,t} - \mathbb{E} \sum_{t=1}^T X_{I_t^\pi, t}$$

Stochastic Bandits

- Asymptotic regret lower bound (no algorithm can beat this performance)
- Uniformly good algorithm: $\mathbb{E}[t_i(T)] = o(T^\alpha)$, $\forall \alpha > 0, \forall \mu, \forall i \neq i^*$

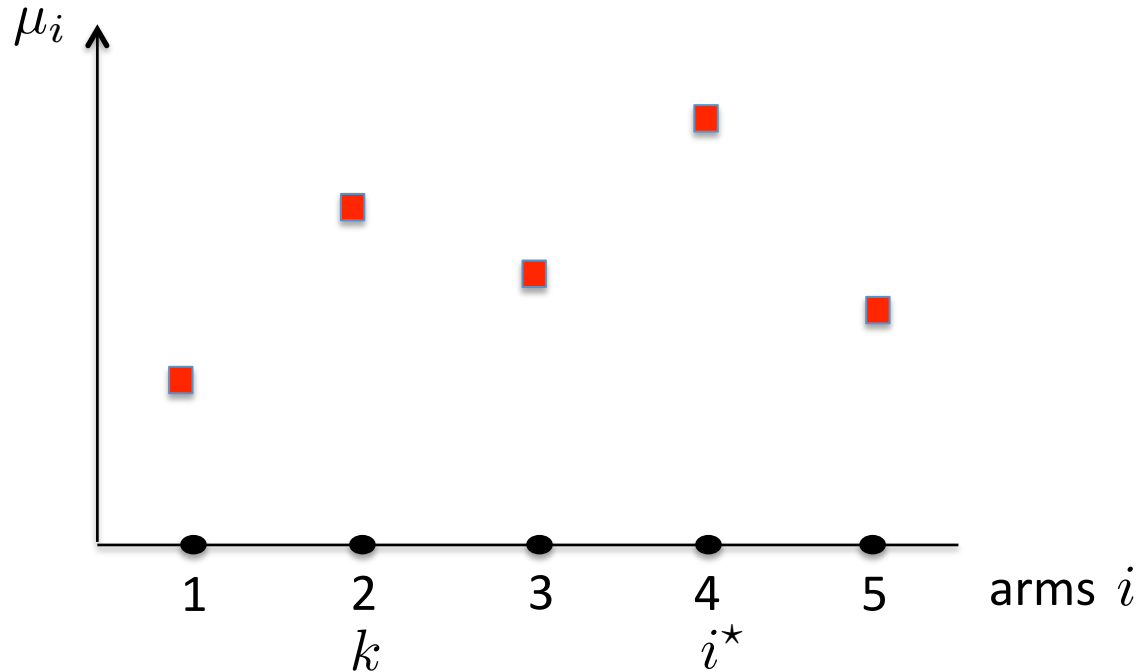
Theorem (Lai-Robbins 1985) For any uniformly good policy π

$$\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq \sum_{i \neq i^*} \frac{\mu^* - \mu_i}{\text{KL}(\mu_i, \mu^*)}$$

KL divergence number: $KL(p, q) = p \log\left(\frac{p}{q}\right) + (1 - p) \log\left(\frac{1 - p}{1 - q}\right)$

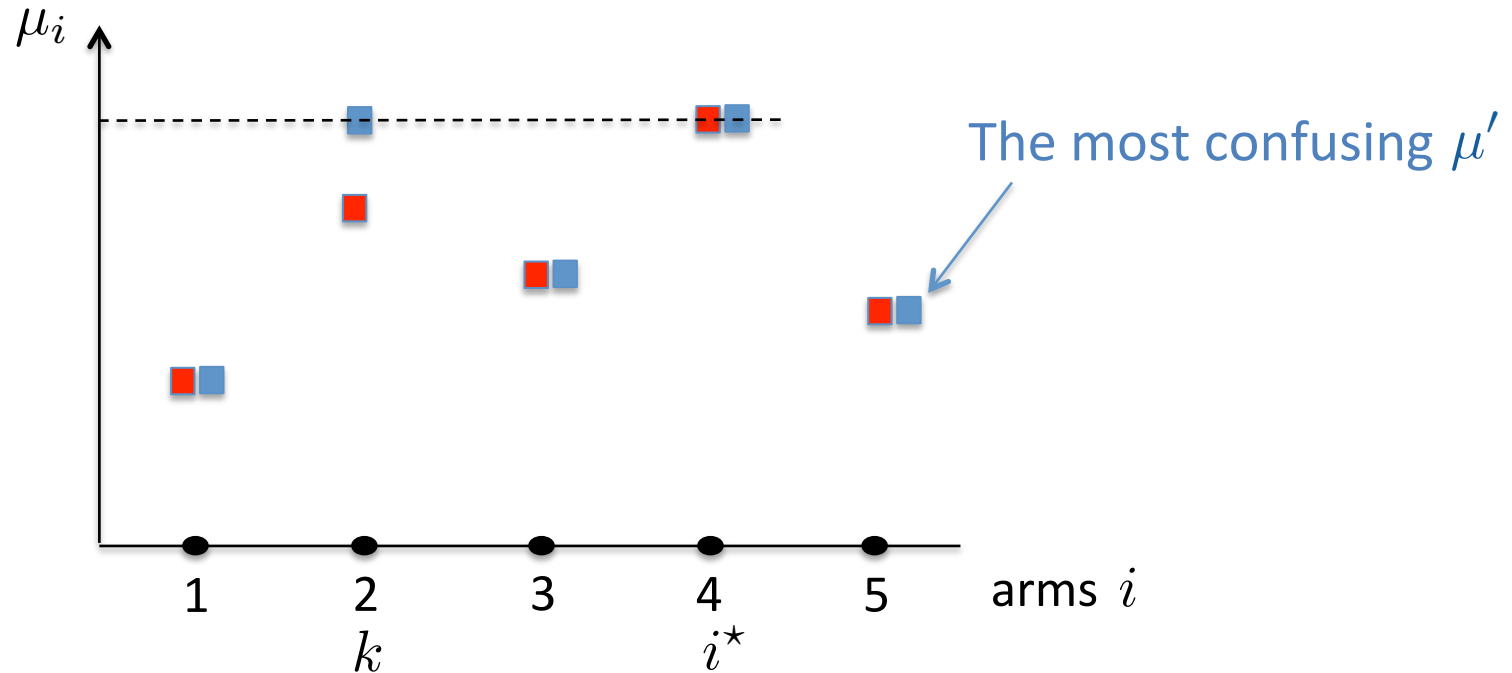
Regret linear in the number of arms, and proportional to $\frac{1}{(\mu^* - \mu_i)}$

The change-of-measure argument



To identify the minimum number of times sub-optimal arm k must be played, find the most confusing parameters

The change-of-measure argument



To identify the minimum number of times sub-optimal arm k must be played, find the most confusing parameters

$$k \text{ played} \sim \frac{\log(T)}{\text{KL}(\mu_k, \mu^*)} \text{ times, yielding a regret } \frac{\mu^* - \mu_k}{\text{KL}(\mu_k, \mu^*)} \log(T)$$

Algorithms

- Optimal but complicated policy (**Lai-Robbins** 1985)
- Simpler and optimal algorithms (**Agrawal** 1995)
- ϵ -greedy algorithm, $\epsilon=1/t$ logarithmic regret
- UCB algorithm (**Auer et al.** 2002): a simple suboptimal index policy

$$\frac{1}{n_i(t)} \sum_{t=1}^{n_i(t)} X_{i,t} + \sqrt{\frac{2 \log(t)}{n_i(t)}}$$

- KL-UCB (**Lai** 1987, **Garivier-Cappe** 2011): claiming back optimality. Index-based policy, index of arm i

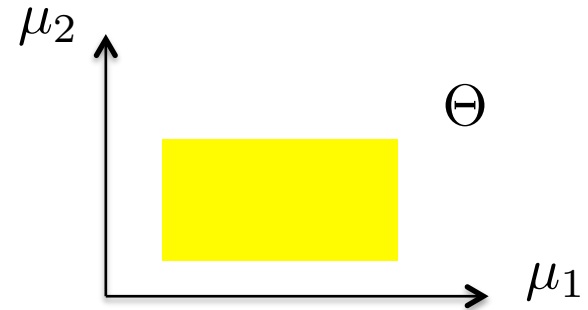
$$\max \left\{ q \leq 1 : n_i(t) \text{KL}(\hat{\mu}_i(t), q) \leq \log(t) + 3 \log \log(t) \right\}$$

Bandit classification

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

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

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



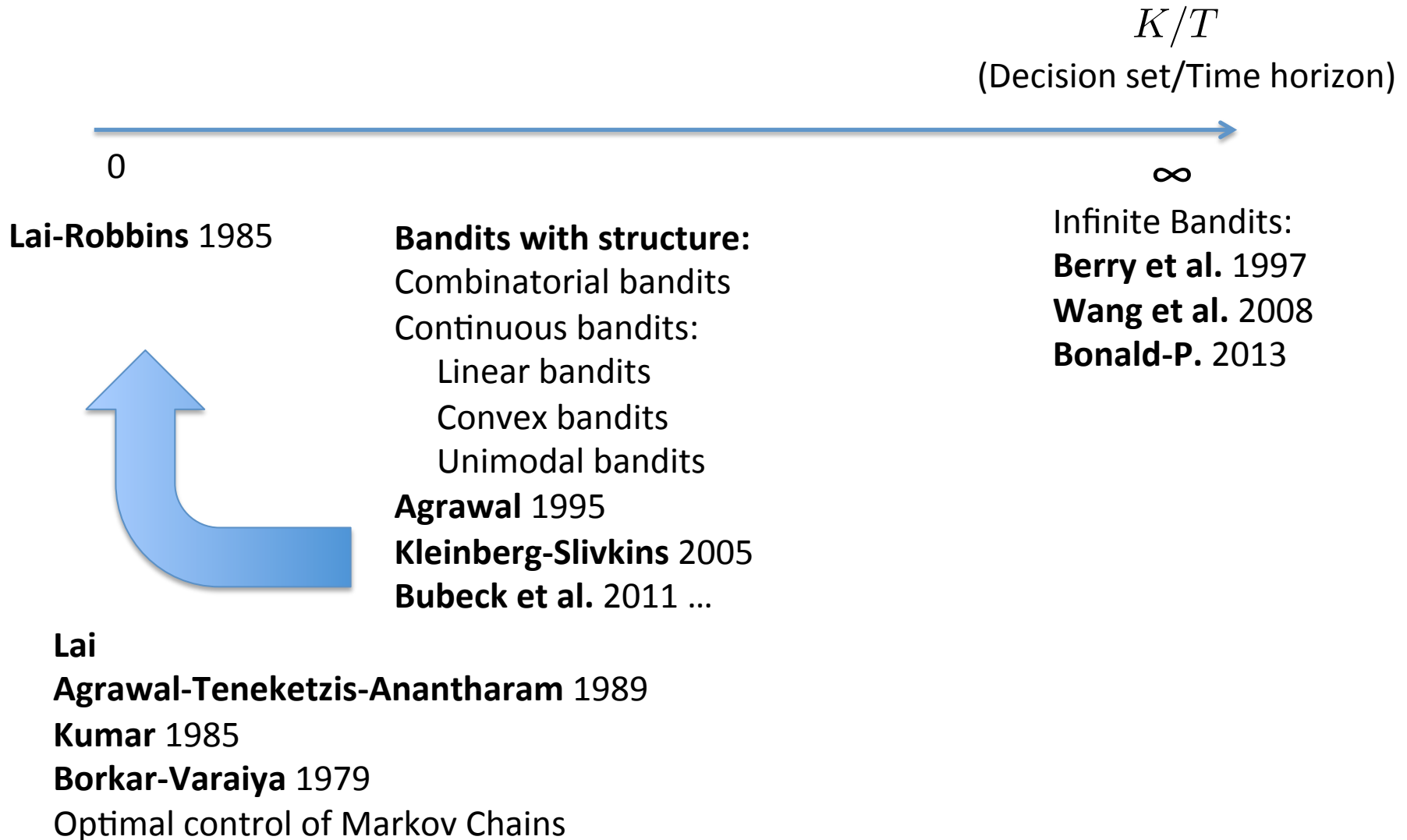
- **Structured bandits:** the decision maker knows that average rewards are related

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



- 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?

Bandit classification



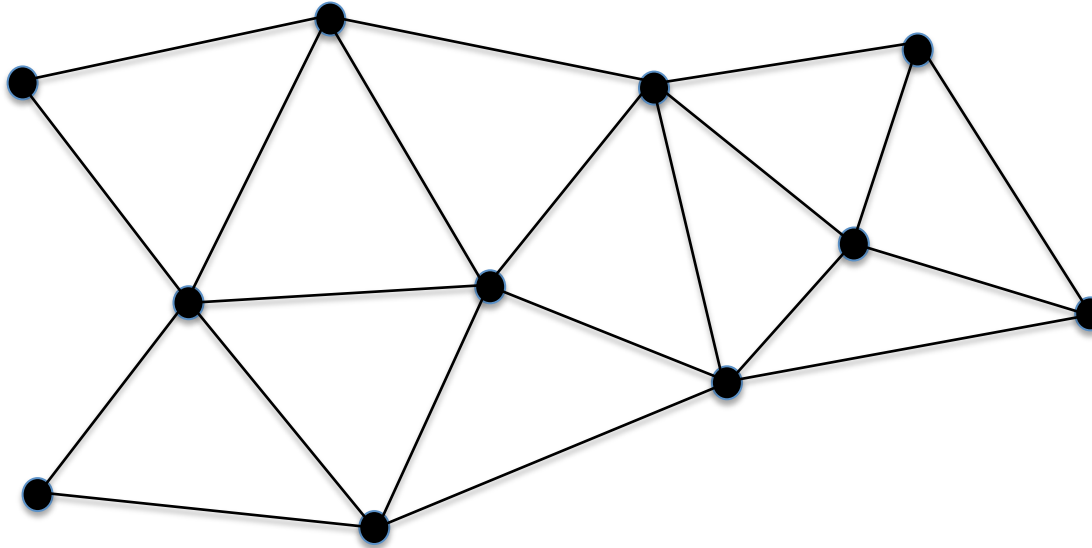
A few papers since 2005

[Abbasi-Yadkori, 2009] [Abernethy, Hazan, Rakhlin, 2008] [Abernethy, Bartlett, Rakhlin, Tewari, 2008] [Abernethy, Agarwal, Bartlett, Rakhlin, 2009] [Audibert, Bubeck, 2010] [Audibert, Munos, Szepesvári, 2009] [Audibert, Bubeck, Lugosi, 2011] [Auer, Ortner, Szepesvári, 2007] [Auer, Ortner, 2010] [Awerbuch, Kleinberg, 2008] [Bartlett, Hazan, Rakhlin, 2007] [Bartlett, Dani, Hayes, Kakade, Rakhlin, Tewari, 2008] [Bartlett, Tewari, 2009] [Ben-David, Pal, Shalev-Shwartz, 2009] [Blum, Mansour, 2007] [Bubeck, 2010] [Bubeck, Munos, 2010] [Bubeck, Munos, Stoltz, 2009] [Bubeck, Munos, Stoltz, Szepesvári, 2008] [Cesa-Bianchi, Lugosi, 2006] [Cesa-Bianchi, Lugosi, 2009] [Chakrabarti, Kumar, Radlinski, Upfal, 2008] [Chu, Li, Reyzin, Schapire, 2011] [Coquelin, Munos, 2007] [Dani, Hayes, Kakade, 2008] [Dorard, Glowacka, Shawe-Taylor, 2009] [Filippi, 2010] [Filippi, Cappé, Garivier, Szepesvári, 2010] [Flaxman, Kalai, McMahan, 2005] [Garivier, Cappé, 2011] [Grünewälder, Audibert, Opper, Shawe-Taylor, 2010] [Guha, Munagala, Shi, 2007] [Hazan, Agarwal, Kale, 2006] [Hazan, Kale, 2009] [Hazan, Megiddo, 2007] [Honda, Takemura, 2010] [Jaksch, Ortner, Auer, 2010] [Kakade, Shalev-Shwartz, Tewari, 2008] [Kakade, Kalai, 2005] [Kale, Reyzin, Schapire, 2010] [Kanade, McMahan, Bryan, 2009] [Kleinberg, 2005] [Kleinberg, Slivkins, 2010] [Kleinberg, Niculescu-Mizil, Sharma, 2008] [Kleinberg, Slivkins, Upfal, 2008] [Kocsis, Szepesvári, 2006] [Langford, Zhang, 2007] [Lazaric, Munos, 2009] [Li, Chu, Langford, Schapire, 2010] [Li, Chu, Langford, Wang, 2011] [Lu, Pàl, Pàl, 2010] [Maillard, 2011] [Maillard, Munos, 2010] [Maillard, Munos, Stoltz, 2011] [McMahan, Streeter, 2009] [Narayanan, Rakhlin, 2010] [Ortner, 2008] [Pandey, Agarwal, Chakrabarti, Josifovski, 2007] [Poland, 2008] [Radlinski, Kleinberg, Joachims, 2008] [Rakhlin, Sridharan, Tewari, 2010] [Rigollet, Zeevi, 2010] [Rusmevichientong, Tsitsiklis, 2010] [Shalev-Shwartz, 2007] [Slivkins, Upfal, 2008] [Slivkins, 2011] [Srinivas, Krause, Kakade, Seeger, 2010] [Stoltz, 2005] [Sundaram, 2005] [Wang, Kulkarni, Poor, 2005] [Wang, Audibert, Munos, 2008]

See ICML tutorial 2011: **Audibert-Munos**

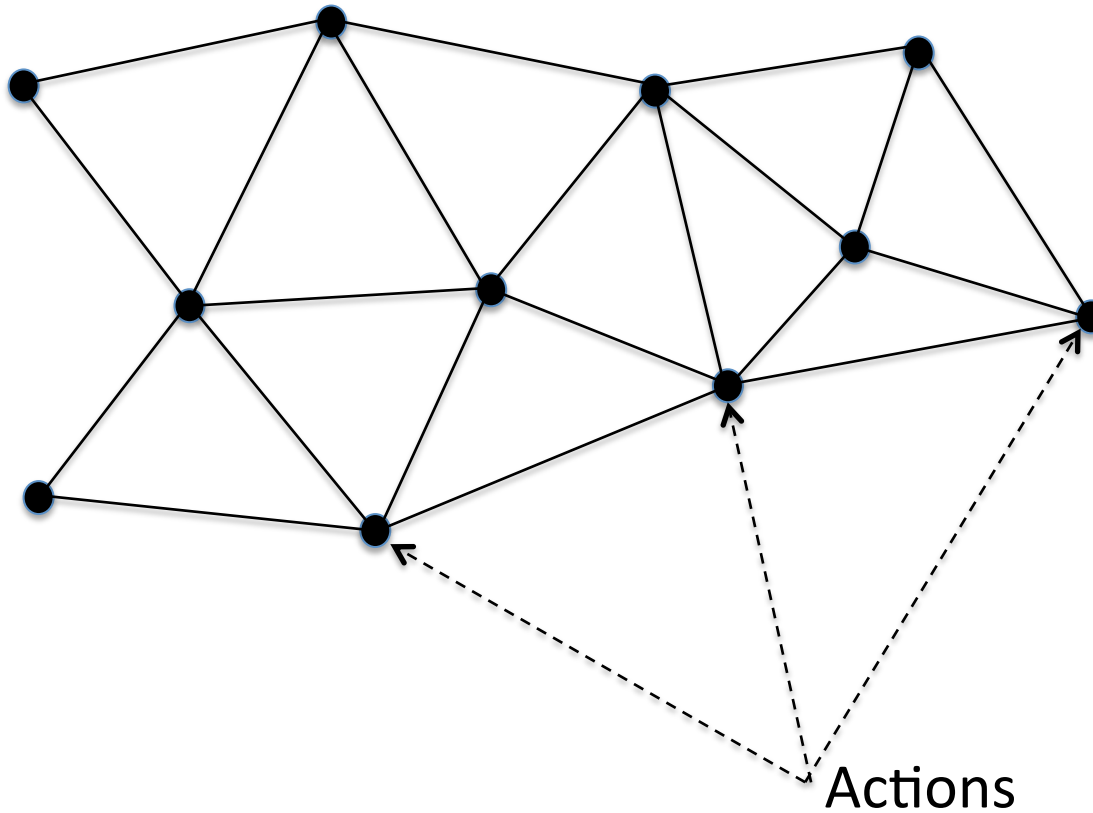
2. Graphically unimodal bandits

Actions and rewards



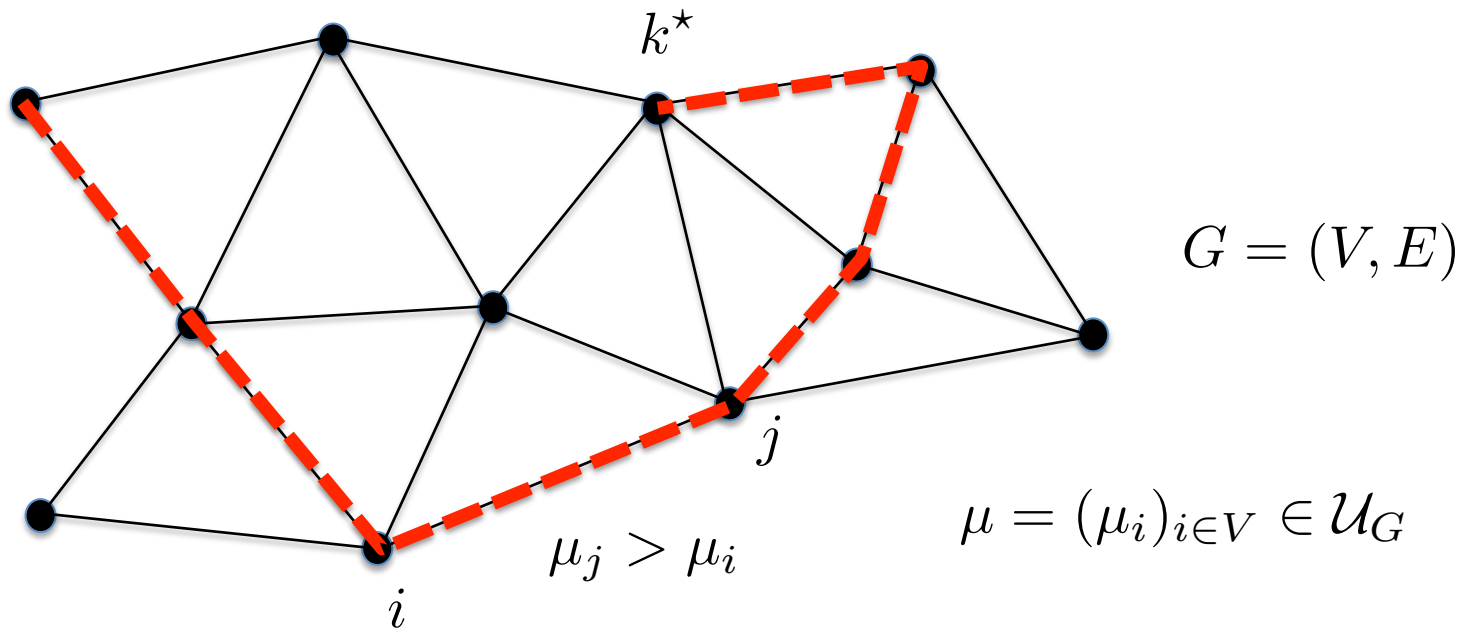
$$G = (V, E)$$

Actions and rewards



$$G = (V, E)$$

Actions and rewards



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

Related work: **Yu-Mannor** 2011 (sub-optimal algorithms)

Average rewards

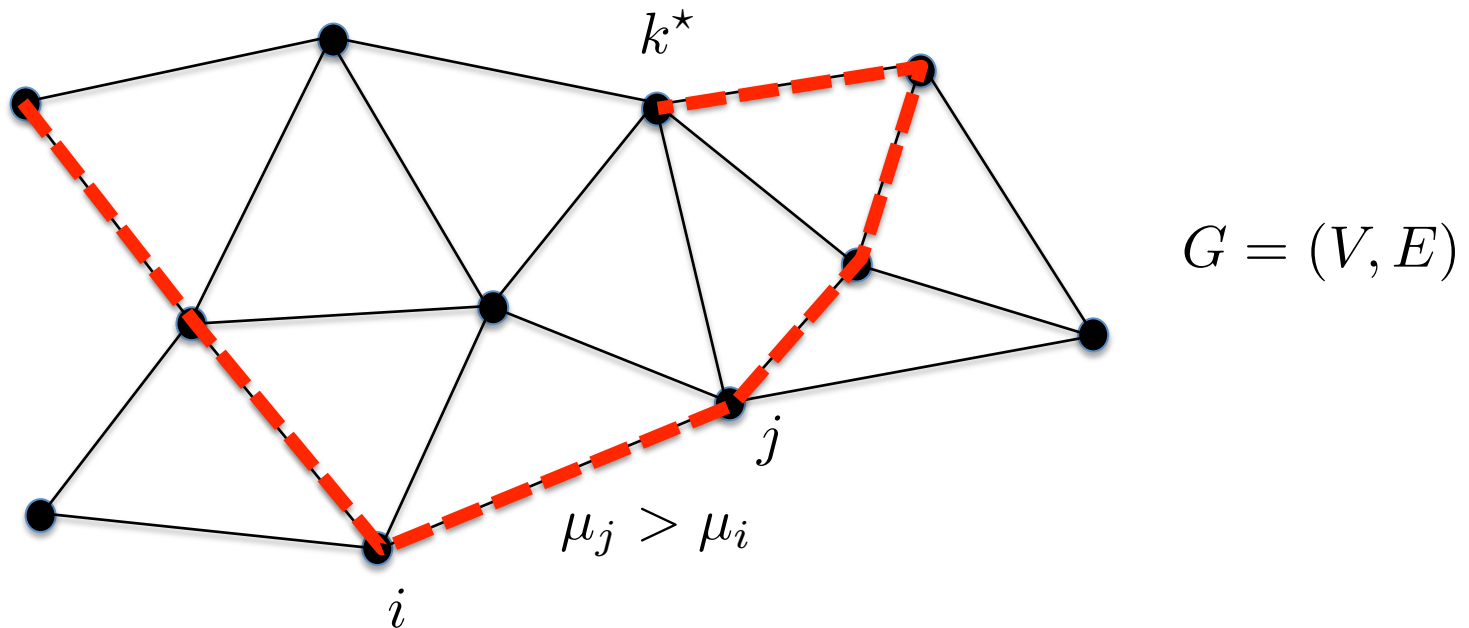
- Linear structure: at time t , action i yields a reward $r_i X_{i,t}$
 $X_{i,t} \sim \text{Ber}(\theta_i)$

r_1	r_2	$\dots$	r_N	known
θ_1	θ_2	$\dots$	θ_N	unknown $\in [0, 1]$
μ_1	μ_2	$\dots$	μ_N	Average rewards
$\mu_i = r_i \theta_i$				

Optimal action: k^* , $\mu^* = \mu_{k^*}$

- Graphical unimodality w.r.t. some known graph G

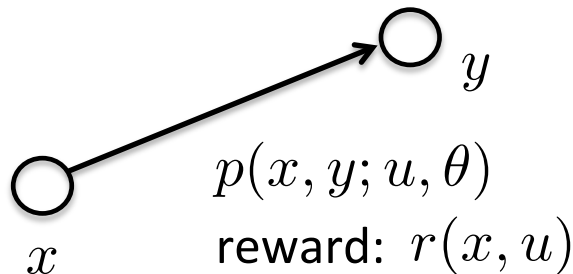
Graphically unimodal bandits



How can the reward structure be optimally exploited?
How does regret scale with the graph size and topology?

Controlled Markov Chains

Kumar, Lai, Borkar, Varaiya, ...



- Finite state and action spaces $\theta \in \Theta$
- Unknown parameter Θ : compact metric space

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

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

- 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 (Graves-Lai'97): $\liminf_{T \rightarrow \infty} \frac{R^{\pi}(T)}{\log(T)} \geq c(\theta)$

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

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

Application to graphically unimodal bandits

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

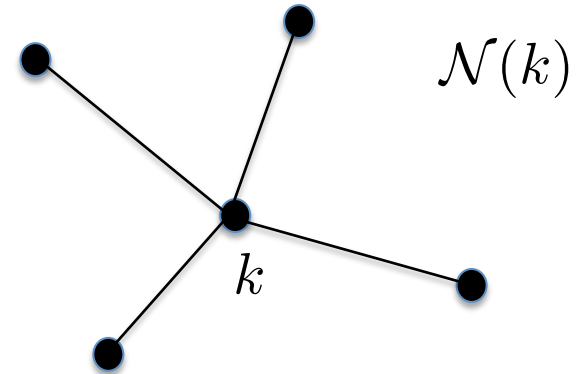
$$p(x, y : k, \theta) = \begin{cases} \theta_k & \text{if } y = r_k \\ 1 - \theta_k & \text{if } y = 0 \end{cases}$$

- Average rewards: $g = k$

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

Fundamental performance limit

$$N(k) = \{l \in \mathcal{N}(k) : r_k \theta_k \leq r_l\}$$



Theorem: For any uniformly good algorithm π

$$\liminf_{T \rightarrow \infty} \frac{R^\pi(T)}{\log(T)} \geq c_G(\theta) \quad c_G(\theta) = \sum_{k \in N(k^*)} \frac{r_{k^*} \theta_{k^*} - r_k \theta_k}{\text{KL}(\theta_k, \frac{r_{k^*} \theta_{k^*}}{r_k})}$$

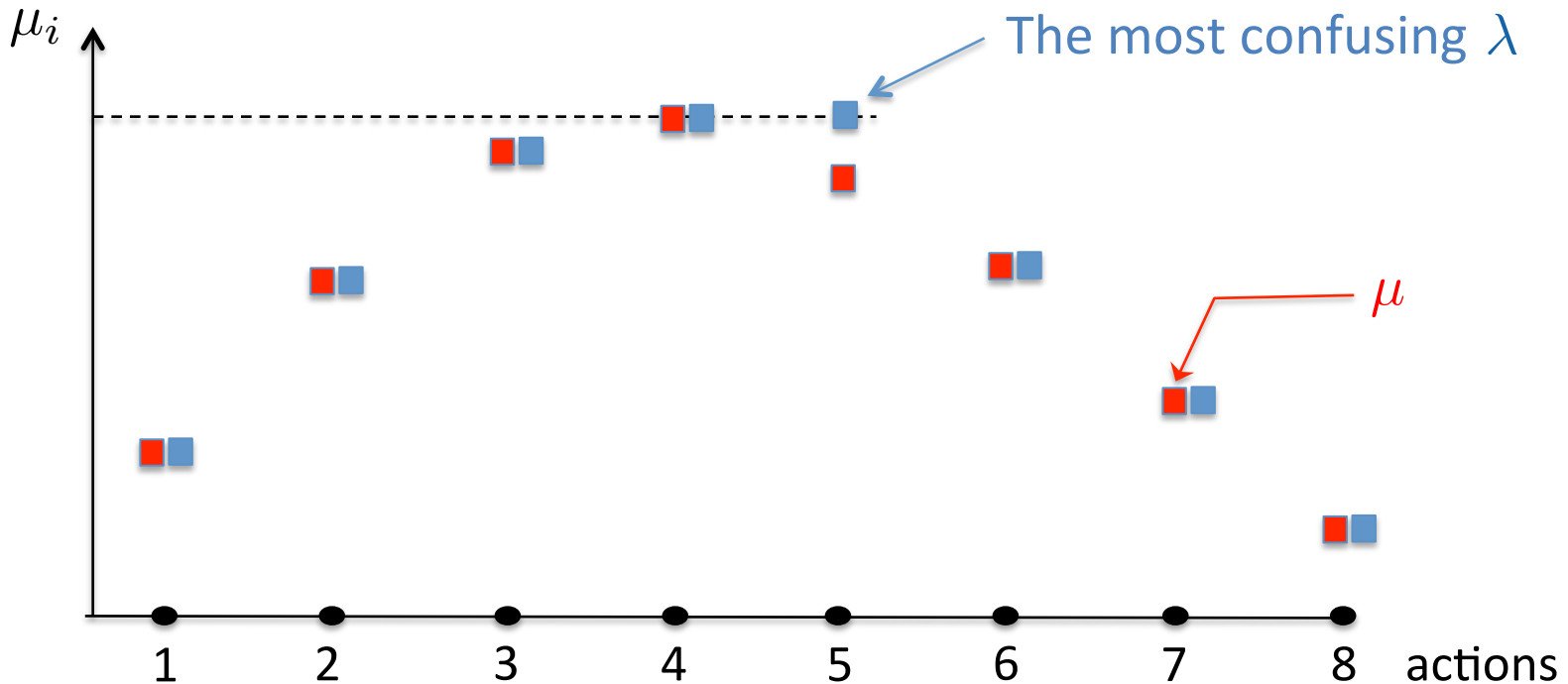
The performance limit does not depend on the size of the decision space! Structure could really help.

Proof

$$\inf_{g \neq g^*} \sum c_g (\mu_{g^*}(\theta) - \mu_g(\theta))$$

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

Example: classical unimodality



Optimal Action Sampling

- Empirical average reward: $\hat{\mu}_k(n) = \frac{1}{t_k(n)} \sum_{s=1}^{t_k(n)} r_k X_k(s)$
- 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 : $b_k(n) = \max \left\{ q \in [0, r_k] : t_k(n) \text{KL} \left(\frac{\hat{\mu}_k(n)}{r_k}, \frac{q}{r_k} \right) \leq \log(l_{L(n)}(n)) + c \log \log(l_{L(n)}(n)) \right\}$

Optimal Action Sampling

Algorithm – Optimal Action Sampling (OAS)

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

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

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

Theorem: For any $\mu \in \mathcal{U}_G$, $\limsup_{T \rightarrow \infty} \frac{R^{OAS}(T)}{\log(T)} \leq c_G(\theta).$

Proof

$$\begin{aligned} R^{OAS}(T) &\leq r_K \sum_{k \neq k^*} \mathbb{E}[l_k(T)] \\ &\quad + \sum_{k \in N(k^*)} (r_{k^*} \theta_{k^*} - r_k \theta_k) \mathbb{E} \left[\sum_{t=1}^T 1_{L(t)=k^*, k(t)=k} \right] \end{aligned}$$

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.

$t_n = \sum_{t=n_0}^n B_t$. $\phi \in \{n_0, \dots, T+1\}$ a $\mathcal{F}$ -stopping time with:
either $t_\phi \geq s$ or $\phi = T+1$. Then:

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

Non-stationary environments

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

$$R^\pi(T) = \sum_{t=1}^T (r_{k^*(t)} \theta_{k^*(t)}(t) - r_{k^\pi(t)} \theta_{k^\pi(t)}(t))$$

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

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{n=1}^T \sum_{k, k' \in N(k)} 1_{|r_k \theta_k(n) - r_{k'} \theta_{k'}(n)| \geq \Delta} \leq \phi(K) \Delta$$

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 $\pi = \text{SW-OAS}$

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

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)
- Sub-logarithmic terms are essential in the regret analysis
- Upper bound on regret per time unit:
 - Tends to zero when the evolution of average rewards gets smoother

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

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

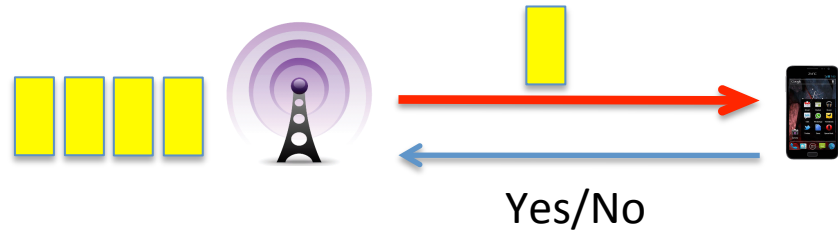
3. Rate adaptation in 802.11

with R. Combes, D. Yun, J. Ok, Y. Yi

Rate adaptation in 802.11

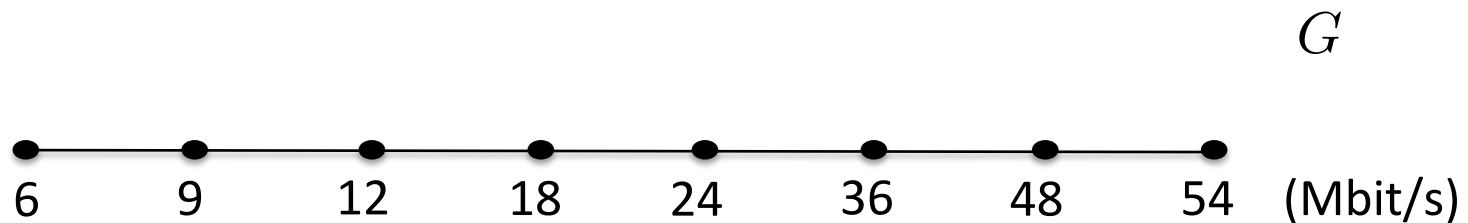
Adapting the modulation/coding scheme to the radio environment

- 802.11 a/b/g



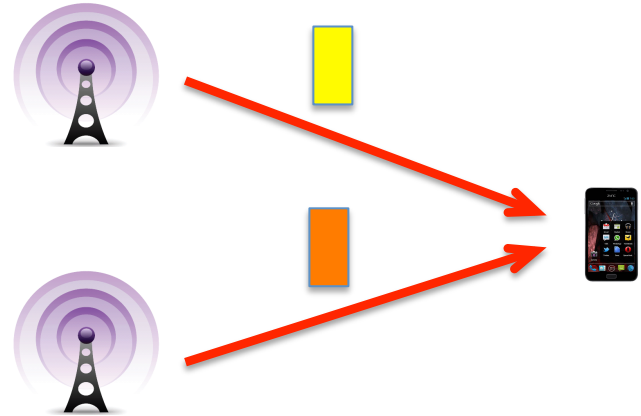
rates	r_1	r_2	$\dots$	r_N	
Success probabilities	θ_1	θ_2	$\dots$	θ_N	
Throughputs	μ_1	μ_2	$\dots$	μ_N	$\mu_i = r_i \theta_i$

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

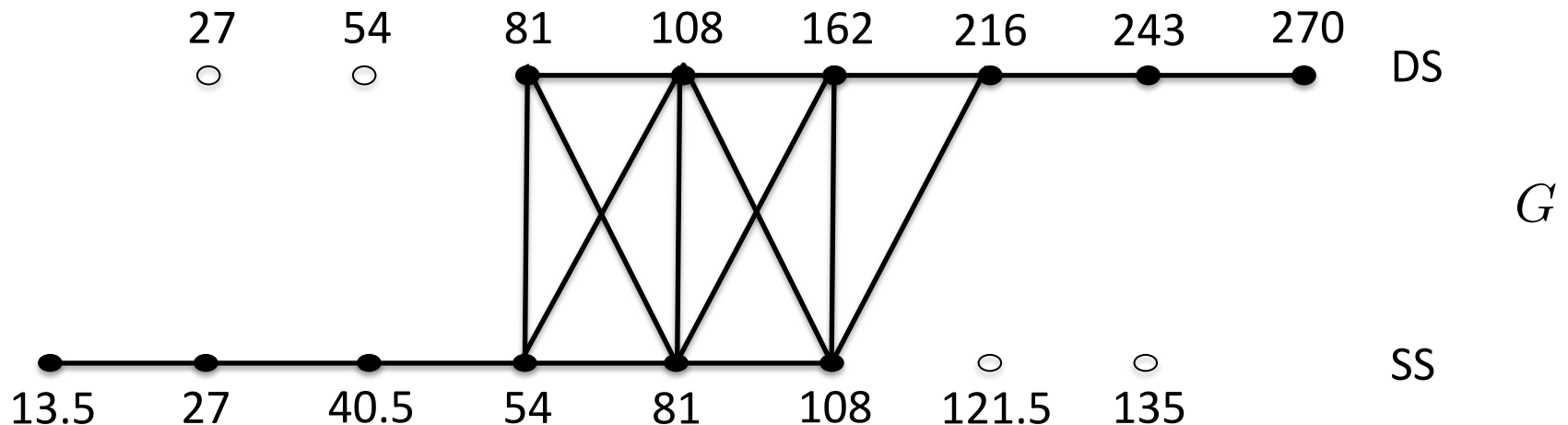


Rate adaptation in 802.11

- 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 $n = 1, \dots, K$, select action $k(n) = n$

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

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

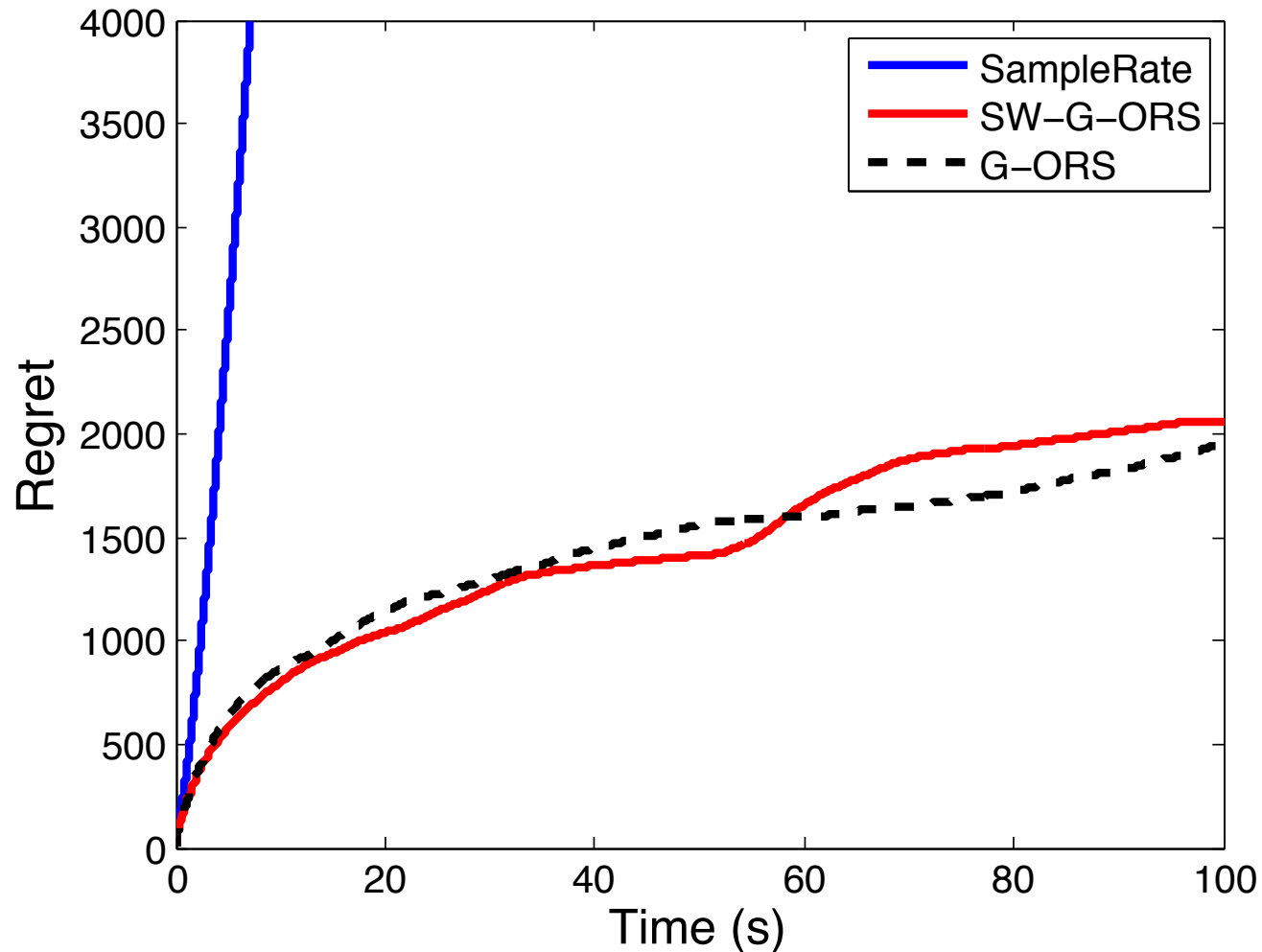
ORS is asymptotically optimal (minimizes regret)

Its performance does not depend on the number of possible rates!

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

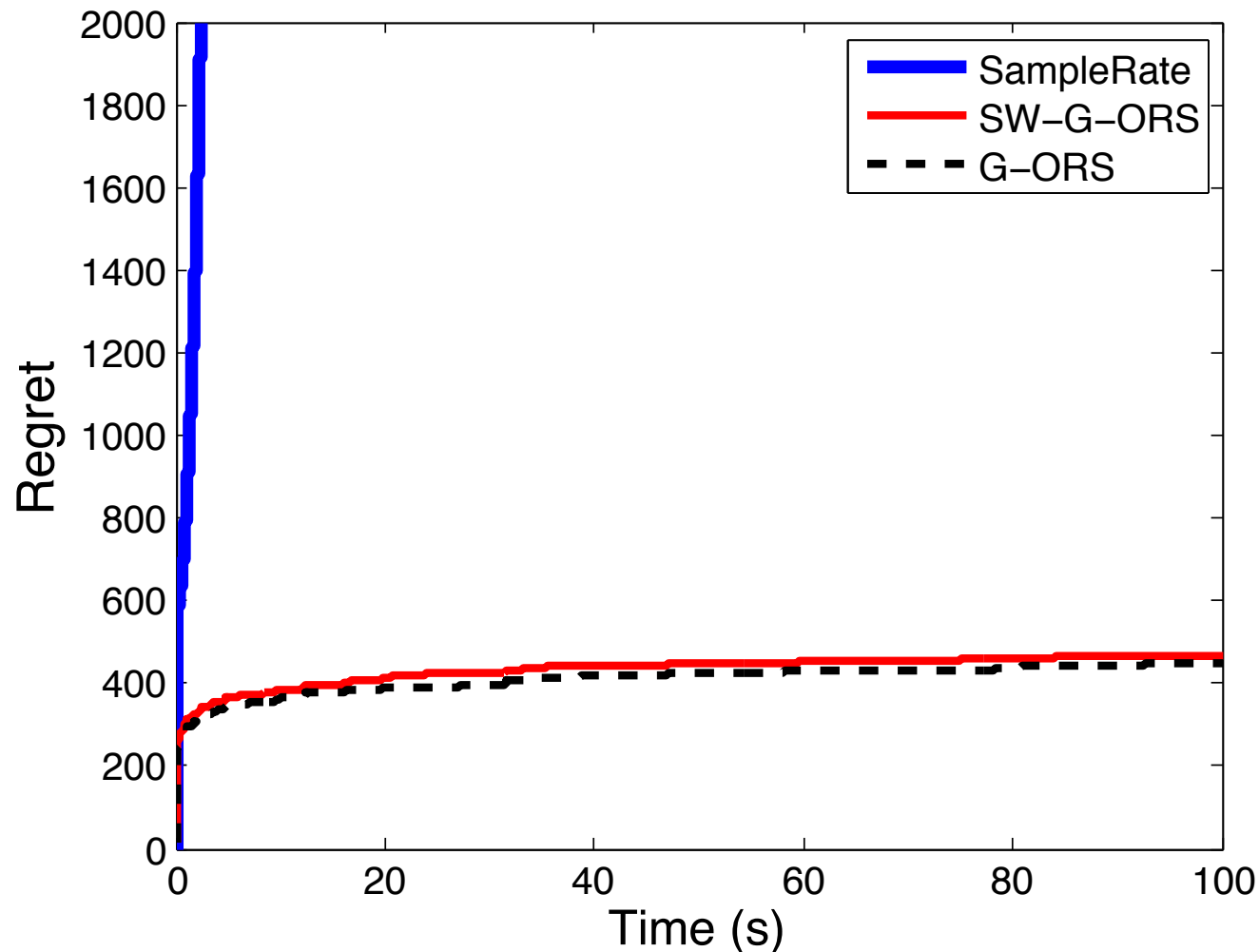
802.11g – stationary environment

GRADUAL (success prob. smoothly decreases with rate)



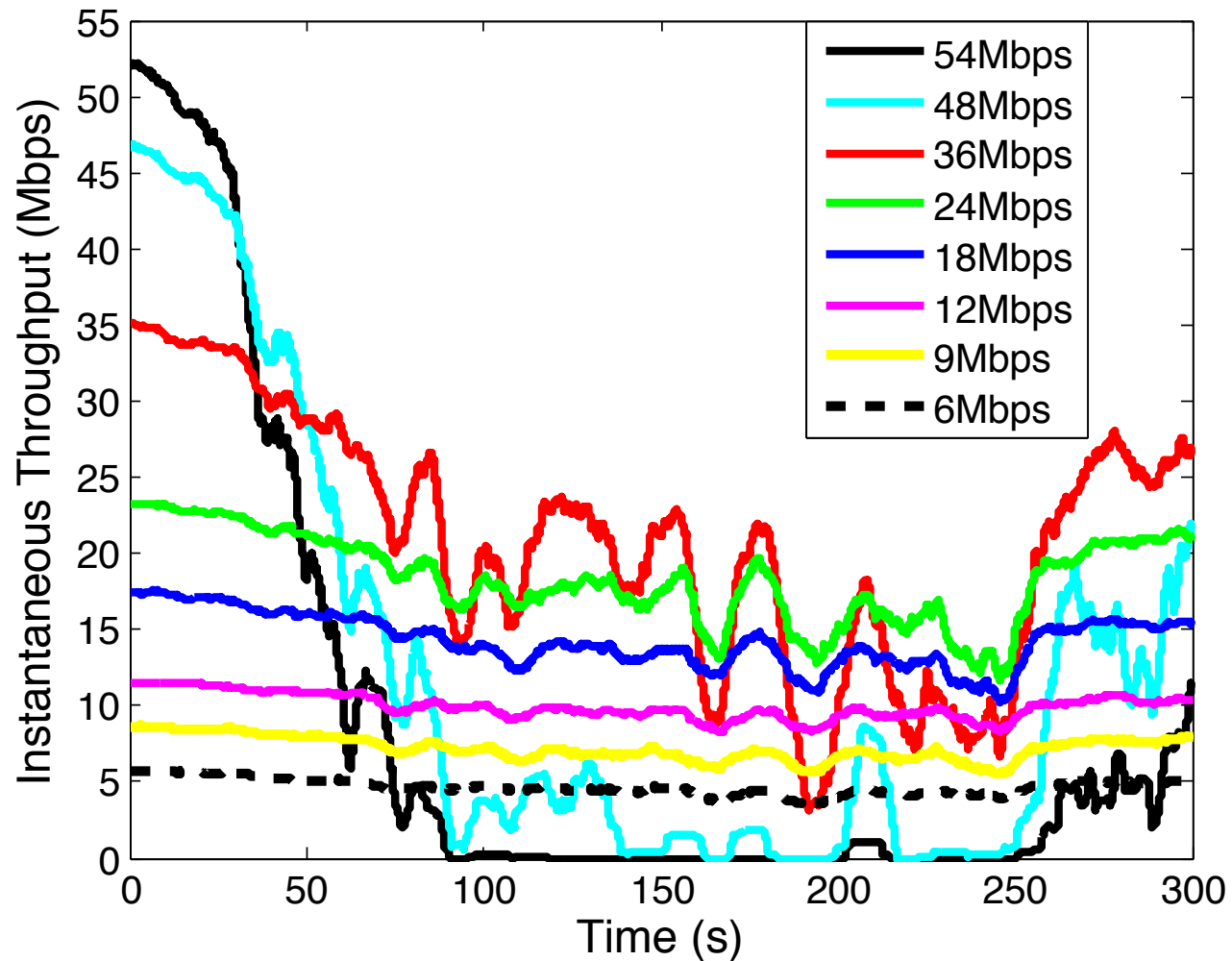
802.11g – stationary environment

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



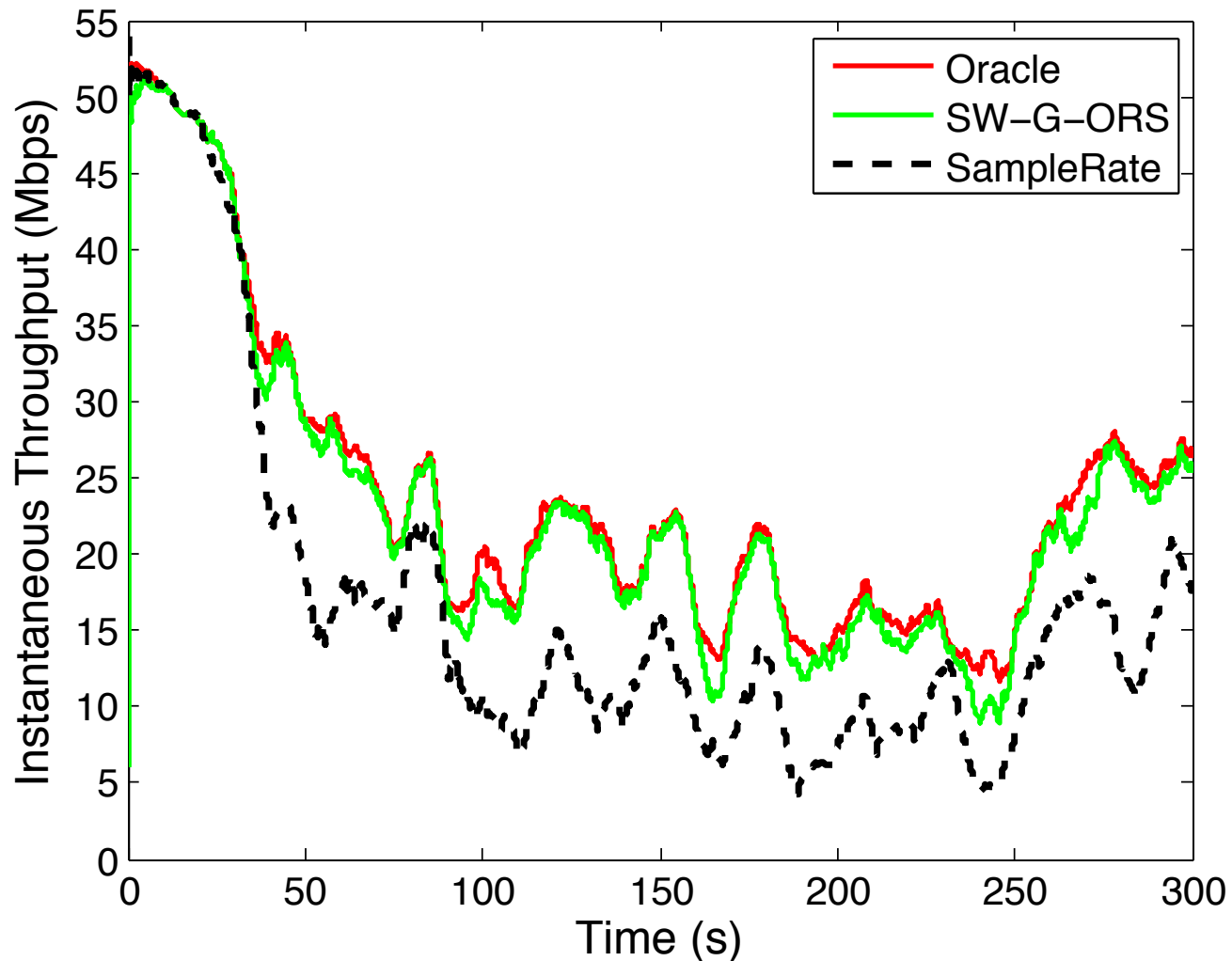
802.11g – non-stationary environment

TRACES



802.11g – non-stationary environment

RESULTS



Extensions

- The approach can be extended to other structured bandit problems
 - Lipschitz bandits
 - Linear bandits
 - Convex bandits
 - ...
- Continuous set of action
 - Unimodal: A stochastic golden search algorithm (see arxiv.org)

Thanks!

Papers

Combes-Proutiere, ICML 2014

Magureanu-Combes-Proutiere, COLT 2014