Bandit Optimization: Theory and Applications

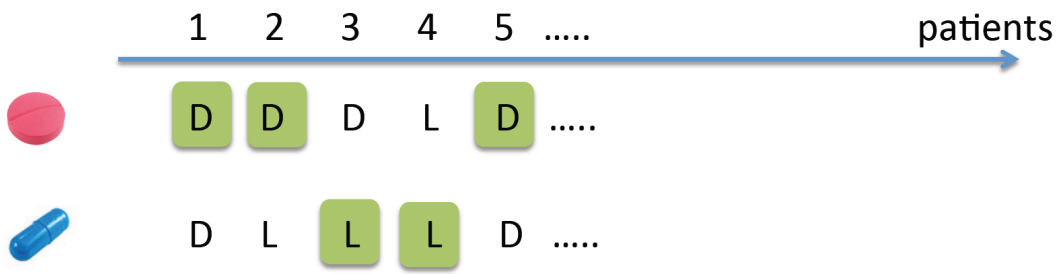
Richard Combes[1] and Alexandre Proutière[2]

[1]Centrale-Supélec / L2S, France
[2]KTH, Royal Institute of Technology, Sweden.

SIGMETRICS 2015

Outline

A first example: sequential treatment allocation

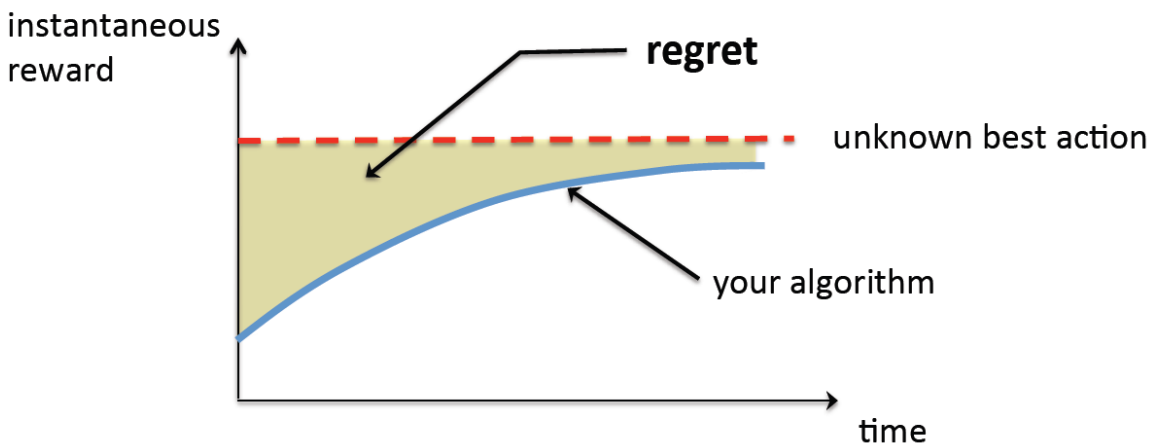


- There are T patients with the same symptoms awaiting treatment
- Two treatments exist, one is better than the other
- Based on past successes and failures which treatment should you use ?

The model

- At time n, choose action $x_n \in \mathcal{X}$, observe feedback $y_n(x_n) \in \mathcal{Y}$, and obtain reward $r_n(x_n) \in \mathbb{R}^+$.
- "Bandit feedback": rewards and feedback depend on actions (often $y_n \equiv r_n$)
- Admissible algorithm:
  $x_{n+1} = f_{n+1}(x_0, r_0(x_0), y_0(x_0), ..., x_n, r_n(x_n), r_n(y_n))$
- Performance metric: regret

$$R(T) = \underbrace{\max_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{n=1}^{T} r_n(x)\right]}_{\text{oracle}} - \underbrace{\mathbb{E}\left[\sum_{n=1}^{T} r_n(x_n)\right]}_{\text{your algorithm}}.$$



Bandit taxonomy: adversarial vs stochastic

Stochastic Bandit:

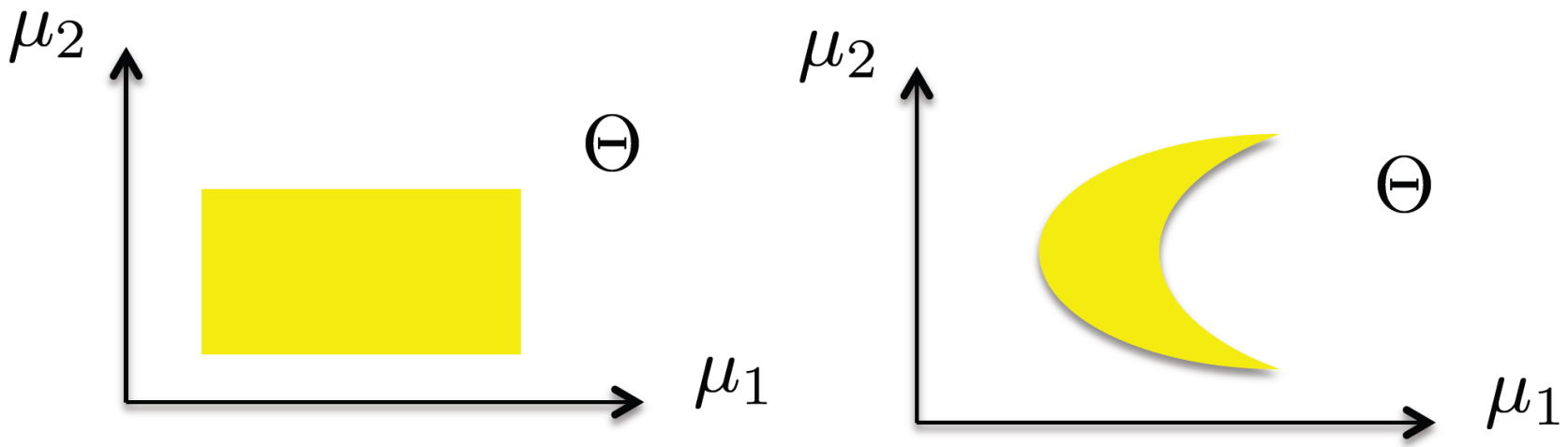
- Game against a stochastic environment
- Unknown parameters $\theta \in \Theta$
- $(r_n(x))_n$ is i.i.d with expectation θ_x

Adversarial Bandit:

- Game against a *non-adaptive* adversary
- For all x, $(r_n(x))_n$ arbitrary sequence in $\mathcal{X}$
- At time 0, the adversary "writes down $(r_n(x))_{n,x}$ in an envelope"

Engineering problems are mainly stochastic

Independent vs correlated arms



- Independent arms: $\Theta = [0,1]^K$
- Correlated arms: $\Theta \neq [0,1]^K$: choosing 1 gives information on 1 and 2

Correlation enables (sometimes much) faster learning.

Bandit taxonomy: frequentist vs bayesian

How to assess an algorithm applied to a *set* of problems ?

Frequentist (classical):

- Problem dependent regret: $R_\theta^\pi(T)$, θ fixed
- Minimax regret: $\max_{\theta \in \Theta} R_\theta^\pi(T)$
- Usually very different regret scaling

Bayesian:

- Prior distribution $\theta \sim P$, known to the algorithm
- Bayesian regret: $\mathbb{E}_{\theta \sim P}[R_\theta^\pi(T)]$
- P naturally includes information on the problem structure

Bandit taxonomy: cardinality of the set of arms

Discrete Bandits:

- $\mathcal{X} = \{1, ..., K\}$
- All arms can be sampled infinitely many times
- Regret $O(\log(T))$ (stochastic), $O(\sqrt{T})$ (adversarial)

Infinite Bandits:

- $\mathcal{X} = \mathbb{N}$, Bayesian setting (otherwise trivial)
- Explore $o(T)$ arms until a good one is found
- Regret: $O(\sqrt{T})$.

Continuous Bandits:

- $\mathcal{X} \subset \mathbb{R}^d$ convex, $x \mapsto \mu_\theta(x)$ has a *structure*
- Structures: convex, Lipschitz, linear, unimodal (quasi-convex) etc.
- Similar to derivative-free stochastic optimization
- Regret: $O(\mathbf{poly}(d)\sqrt{T})$.

Bandit taxonomy: regret minimization vs best arm identification

Sample arms and output the best arm with a given probability, similar to PAC learning

Fixed budget setting:

- T fixed, sample arms $x_1, ..., x_T$, and output $\hat{x}^T$
- Easier problem: estimation + budget allocation
- Goal: minimize $\mathbb{P}[\hat{x}^T \neq x^*]$

Fixed confidence setting:

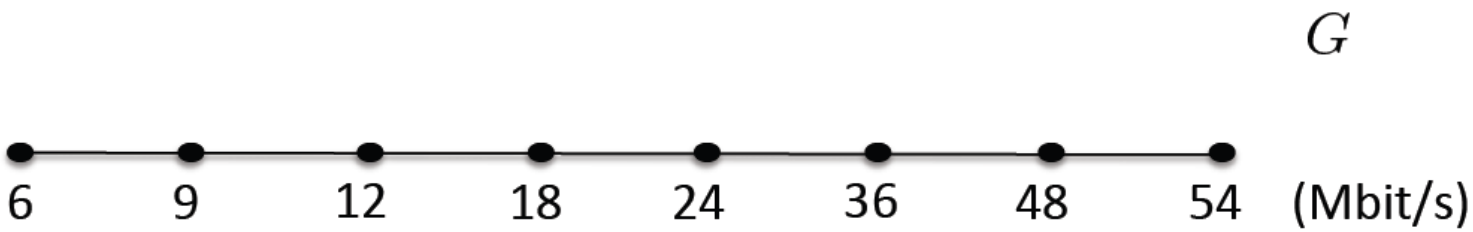
- δ fixed, sample arms $x_1, ..., x_\tau$ and output $\hat{x}^\tau$
- Harder problem: estimation + budget allocation + optimal stopping (τ is a stopping time)
- Goal: minimize $\mathbb{E}[\tau]$ s.t. $\mathbb{P}[\hat{x}^\tau \neq x^*] \leq \delta$

Example 1: Rate adaptation in wireless networks

- Adapting the modulation/coding scheme to the radio environment

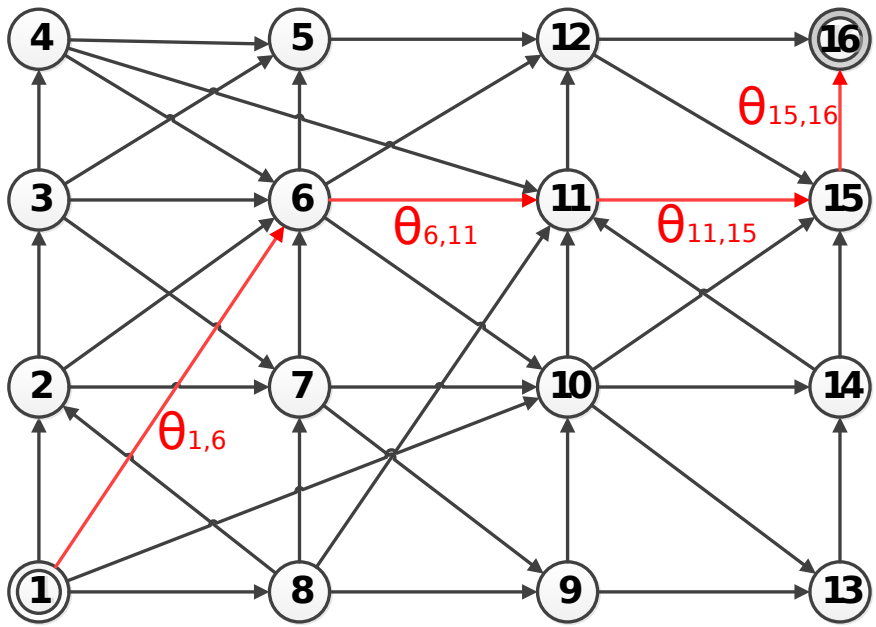
- Rates: $r_1, r_2, \ldots, r_K$
- Success probabilities: $\theta_1, \theta_2, \ldots, \theta_K$
- Throughputs: $\mu_1, \mu_2, \ldots, \mu_K$

Structure: unimodality + $\theta_1 > \theta_2 > \cdots > \theta_K$.



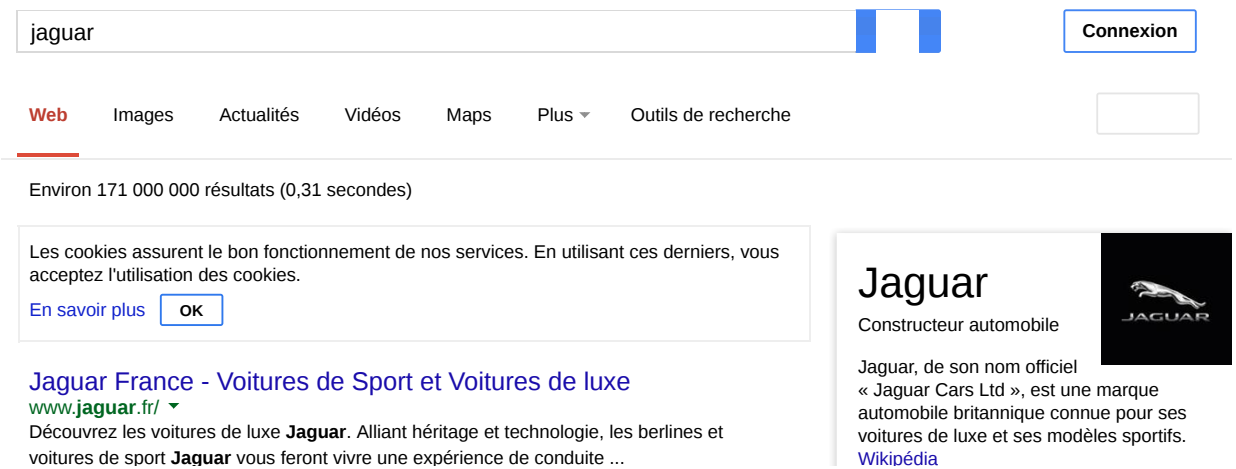
Example 2: Shortest path routing

- Choose a path minimizing expected delay
- Stochastic delays: $X_i(n) \sim \text{Geometric}(\theta_i)$
- Path $M \in \{0,1\}^d$, expected delay $\sum_{i=1}^{d} M_i/\theta_i$.
- Semi-bandit feedback: $X_i(n)$, for $\{i : M_i(n) = 1\}$



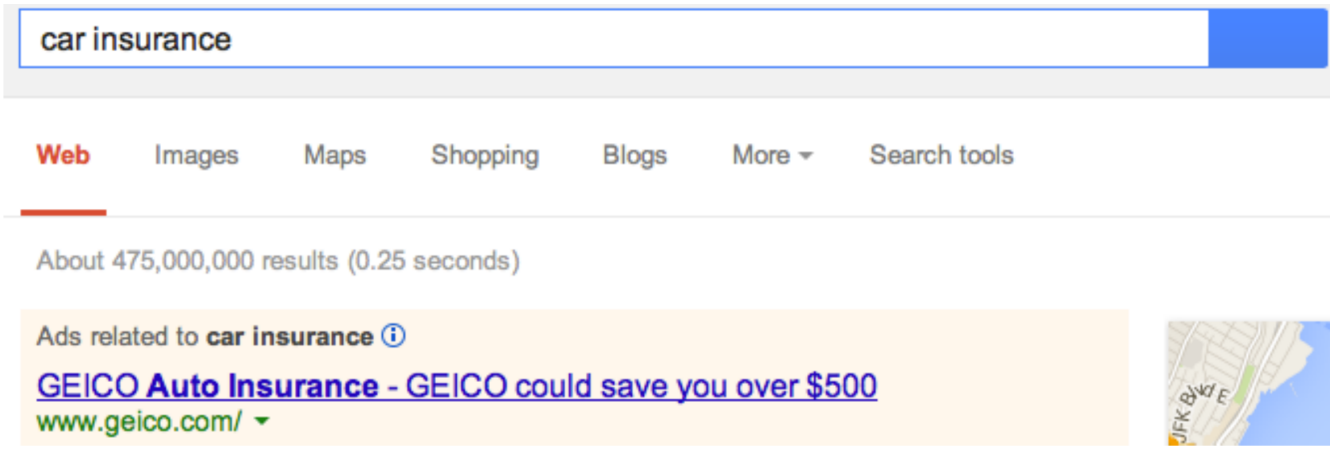
Example 3: Learning to Rank (search engines)

- Given a query, N relevant items, L display slots
- A user is shown L items, scrolls down and selects the first relevant item
- One must show the most relevant items in the first slots.
- θ_n probability of clicking on item n (independence between items is assumed)
- Reward $r(\ell)$ if user clicks on the ℓ-th item, and 0 if the user does not click



Example 4: Ad-display optimization

- Users are shown ads relevant to their queries
- Announcers $x \in \{1, ..., K\}$, with μ_x click-through-rate and budget per unit of time c_x
- Bandit with budgets: each arm has a budget of plays
- Displayed announcer is charged per impression/click



Outline

Optimism in the face of uncertainty

- Replace arm values by upper confidence bounds
- "Index" $b_x(n)$ such that $b_x(n) \geq \theta_x$ with high probability
- Select the arm with highest index $x_n \in \arg\max_{x \in \mathcal{X}} b_x(n)$
- Analysis idea:

$$\mathbb{E}[t_x(T)] \leq \underbrace{\sum_{n=1}^{T} \mathbb{P}[b_{x^\star}(n) \leq \theta^\star]}_{o(\log(T))} + \underbrace{\sum_{n=1}^{T} \mathbb{P}[x_n = x, b_x(n) \geq \theta^\star]}_{\text{dominant term}}.$$

Almost all algorithms in the literature are optimistic (sic!)

Information theory and statistics

- Distribtions P, Q with densities p and q w.r.t a measure m
- Kullback-Leibler divergence:

$$D(P||Q) = \int_x p(x) \log\left(\frac{p(x)}{q(x)}\right) m(dx),$$

- Pinsker's inequality:

$$\sqrt{\frac{D(P||Q)}{2}} \geq TV(P, Q) = \frac{1}{2}\int_x |p(x) - q(x)| m(dx).$$

- If $P, Q \sim \text{Ber}(p), \text{Ber}(q)$:

$$D(P||Q) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

- Also (Pinkser + inequality $\log(x) \leq x - 1$):

$$2(p-q)^2 \leq D(P||Q) \leq \frac{(p-q)^2}{q(1-q)}$$

The KL-divergence is ubiquitous in bandit problems

Empirical divergence and Sanov's inequality

- P a discrete distribution with support $\mathcal{P}$
- $\hat{P}_n$ empirical distribution of an i.i.d sample of size n from P,
- Sanov's inequality:

$$\mathbb{P}[D(\hat{P}_n||P) \geq \delta] \leq \binom{n + |\mathcal{P}| - 1}{|\mathcal{P}| + 1} e^{-n\delta}$$

- Suggests confidence regions (risk α) of the type:

$$\{P : n\, D(\hat{P}_n||P) \leq \log(1/\alpha)\}$$

Hypothesis testing and sample complexity

- How many samples are needed to distinguish P from Q ?
- Observe $X = (X_1, ..., X_n)$ i.i.d with distribution P^n or Q^n
- Additivity of the KL divergence: $D(P^n||Q^n) = nD(P||Q)$
- Test $\phi(X) \in \{0, 1\}$, risk $\alpha > 0$:

$$\mathbb{E}_P[\phi(X)] + \mathbb{E}_Q[1 - \phi(X)] \leq \alpha$$

- Tsybakov's inequality:

$$(1/2)e^{-\min\{D(P^n||Q^n), D(Q^n||P^n)\}} \leq \mathbb{E}_P[\phi(X)] + \mathbb{E}_Q[1 - \phi(X)]$$

- Minimal number of samples:

$$n \geq \frac{\log(1/\alpha) - \log(2)}{\min\{D(P||Q), D(Q||P)\}}$$

Regret Lower Bounds: general technique

- Decision x, two parameters θ, λ, with $x^\star(\lambda) = x \neq x^\star(\theta)$.
- Consider consider an algorithm with $R^\pi(T) = \log(T)$ for all parameters (unformly good):

$$\mathbb{E}_\theta[t_x(T)] = O(\log(T)) \quad , \quad \mathbb{E}_\lambda[t_x(T)] = T - O(\log(T)).$$

- Markov inequality:

$$\mathbb{P}_\theta[t_x(T) \geq T/2] + \mathbb{P}_\lambda[t_x(T) < T/2] \leq O(T^{-1}\log(T)).$$

- $\mathbf{1}\{t_x(T) \leq T/2\}$ is a hypothesis test, risk $O(T^{-1}\log(T))$
- Hence (Neyman-Pearson / Tsybakov):

$$\underbrace{\sum_x \mathbb{E}_\theta[t_x(T)] I(\theta_x, \lambda_x)}_{\text{KL divergence of the observations}} \geq \log(T) - O(\log(\log(T))).$$

Concentration inequalities: Chernoff bounds

- Building indexes requires tight concentration inequalities
- Chernoff bounds: upper bound the MGF
- $X = (X_1, ..., X_n)$ independent, with mean μ, $S_n = \sum_{n'=1}^{n} X_{n'}$
- G such that $\log(\mathbb{E}[e^{\lambda(X_n - \mu)}]) \leq G(\lambda)$, $\lambda \geq 0$
- Generic technique:

$$\begin{aligned}
\mathbb{P}[S_n - n\mu \geq \delta] &= \mathbb{P}[e^{\lambda(S_n - n\mu)} \geq e^{\lambda\delta}] \\
&\leq e^{-\lambda\delta}\mathbb{E}[e^{\lambda(S_n - n\mu)}] \text{ (Markov)} \\
&= \exp(nG(\lambda) - \lambda\delta) \text{ (independence)} \\
&\leq \exp\left(-n \max_{\lambda \geq 0}\{\lambda\delta n^{-1} - G(\lambda)\}\right).
\end{aligned}$$

Concentration inequalities: Chernoff and Hoeffding's inequality

- Bounded variables: if $X_n \in [a, b]$ a.s then $\mathbb{E}[e^{\lambda(X_n-\mu)}] \leq e^{\lambda^2(b-a)^2/8}$ (Hoeffding lemma)
- Hoeffding's inequality:

$$\mathbb{P}[S_n - n\mu \geq \delta] \leq \exp\left(-\frac{2\delta^2}{n(b-a)^2}\right)$$

- Subgaussian variables: $\mathbb{E}[e^{\lambda(X_n-\mu)}] \leq e^{\sigma^2\lambda^2/2}$, similar
- Bernoulli variables: $\mathbb{E}[e^{\lambda(X_n-\mu)}] = \mu e^{\lambda(1-\mu)} - (1-\mu)e^{-\lambda\mu}$
- Chernoff's inequality:

$$\mathbb{P}[S_n - n\mu \geq \delta] \leq \exp(-nI(\mu + \delta/n, \mu))$$

- Pinsker's inequality: Chernoff is stronger than Hoeffding.

Concentration inequalities: variable sample size and peeling

- In bandit problems, the sample size is random and depends on the samples themselves
- Intervals $\mathcal{N}_k = \{n_k, ..., n_{k+1}\}$, $\mathcal{N} = \cup_{k=1}^{K} \mathcal{N}$
- Idea: $Z_n = e^{\lambda(S_n - n\mu)}$ is a positive sub-martingale:

$$\begin{aligned}
\mathbb{P}[\max_{n \in \mathcal{N}_k}(S_n - \mu n) \geq \delta] &= \mathbb{P}[\max_{n \in \mathcal{N}_k} Z_n \geq e^{\lambda\delta})] \\
&\leq e^{-\lambda\delta}\mathbb{E}[Z_{n_{k+1}}] \text{ (Doob's inequality)} \\
&= \exp(-\lambda\delta + n_{k+1} G(\lambda)) \\
&\leq \exp\left(-n_{k+1} \max_{\lambda \geq 0}\{\lambda\delta n_{k+1}^{-1} - G(\lambda)\}\right).
\end{aligned}$$

- Peeling trick (Neveu): union bound over k, $n_k = (1+\alpha)^k$.

Concentration inequalities: self normalized versions

- Self-normalized versions of classical inequalities
- Garivier's inequality:

$$\mathbb{P}\left[\max_{1\leq n\leq T} nI(S_n/n,\mu)\geq\delta\right]\leq 2e\lceil\log(T)\delta\rceil e^{-\delta}$$

- From Pinkser's inequality (self-normalized Hoeffding):

$$\mathbb{P}\left[\max_{1\leq n\leq T}\sqrt{n}|S_n/n-\mu|\geq\delta\right]\leq 4e\lceil\log(T)\delta^2\rceil e^{-2\delta^2}$$

- Multi-dimensional version, $Y_{n_k}^k = n_k I(S_{n_k}/n_k,\mu)$

$$\mathbb{P}\left[\max_{(n_1,\ldots,n_K)\in[1,T]^K}\sum_{k=1}^{K} Y_{n_k}^k\geq\delta\right]\leq C_K(\log(T)\delta)^K e^{-\delta}$$

Outline

The Lai-Robbins bound

- Actions $\mathcal{X} = \{1, ..., K\}$
- Rewards $\theta = (\theta_1, ..., \theta_K) \in [0, 1]^K$
- Uniformly good algorithm: $R(T) = O(\log(T))$, $\forall \theta$

Theorem (Lai '85)

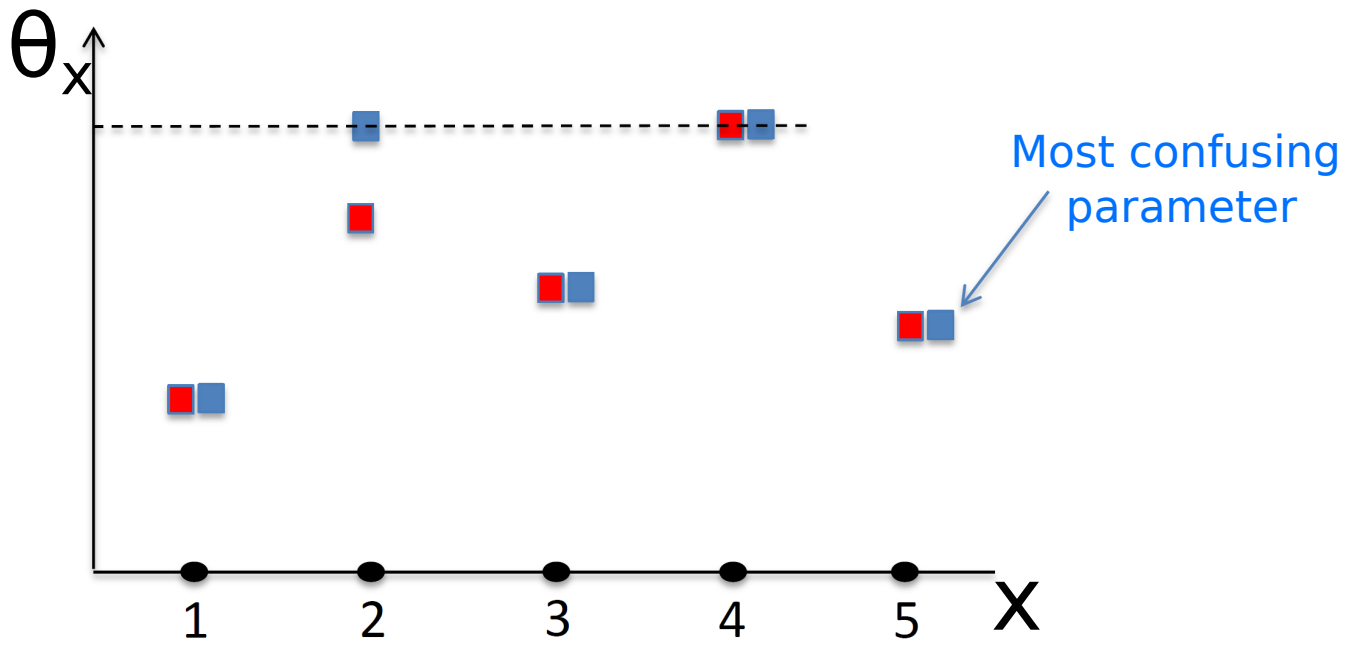
For any uniformly good algorithm, and x s.t $\theta_x < \theta^\star$ we have:

$$\liminf_{T \to \infty} \frac{\mathbb{E}[t_x(T)]}{\log(T)} \geq \frac{1}{I(\mu_x, \mu^\star)}$$

- For $x \neq x^\star$, apply the generic technique with:

$$\lambda = (\theta_1, ..., \theta_{x-1}, \theta^\star + \epsilon, \theta_{x+1}, ..., \theta_K)$$

The Lai-Robbins bound



Optimistic Algorithms

- Select the arm with highest index $x_n \in \arg\max_{x \in \mathcal{X}} b_x(n)$
- UCB algorithm (Hoeffding's ineqality):

$$b_x(n) = \underbrace{\hat{\theta}_x(n)}_{\text{empirical mean}} + \underbrace{\sqrt{\frac{2\log(n)}{t_x(n)}}}_{\text{exploration bonus}} .$$

- KL-UCB algorithm (using Garivier's inequality):

$$b_x(n) = \max\{q \leq 1 : \underbrace{t_x(n) I(\hat{\theta}_x(n), q)}_{\text{likelihood ratio}} \leq \underbrace{f(n)}_{\log(\text{confidence level}^{-1})} \}.$$

with $f(n) = \log(n) + 3\log(\log(n))$.

Regret of optimistic Algorithms

Theorem (Auer'02)

Under algorithm UCB, for all x s.t $\theta_x < \theta^\star$*:*

$$\mathbb{E}[t_x(T)] \leq \frac{8\log(T)}{(\theta_x - \theta^\star)^2} + \frac{\pi^2}{6}.$$

Theorem (Garivier'11)

Under algorithm KL-UCB, for all x s.t $\theta_x < \theta^\star$ *and for all* $\delta < \theta^\star - \theta_x$*:*

$$\mathbb{E}[t_x(T)] \leq \frac{\log(T)}{I(\theta_x + \delta, \theta^\star)} + C\log(\log(T)) + \delta^{-2}.$$

Regret of KL-UCB: sketch of proof

Decompose:

$$\mathbb{E}[t_x(T)] \leq \mathbb{E}[|A|] + \mathbb{E}[|B|] + \mathbb{E}[|C|],$$
$$A = \{n \leq T : b_{x^\star}(n) \leq \theta^\star\},$$
$$B = \{n \leq T : n \notin A, x_n = x, |\hat{\theta}_x(n) - \theta_x| \geq \delta\},$$
$$C = \{n \leq T : n \notin A, x_n = x, |\hat{\theta}_x(n) - \theta_x| \leq \delta, \quad t_x(T) \leq f(T)/I(\theta_x + \delta, \theta^\star)\}.$$

Union bound:

$$\mathbb{E}[|A|] \leq C \log(\log(T)), \qquad \text{(Index property)}$$
$$\mathbb{E}[|B|] \leq \delta^{-2}, \qquad \text{(Hoeffding + Union bound)}$$
$$|C| \leq f(T)/I(\theta_x + \delta, \theta^\star) \qquad \text{(Counting)}$$

Randomized algorithms: Thompson Sampling

- Prior distribution $\theta \sim P$
- Time n, select x with probability $\mathbb{P}[\theta_x = \max_{x'} \theta_{x'} | x_0, r_0, ..., x_n, r_n]$

Bernoulli with uniform priors:

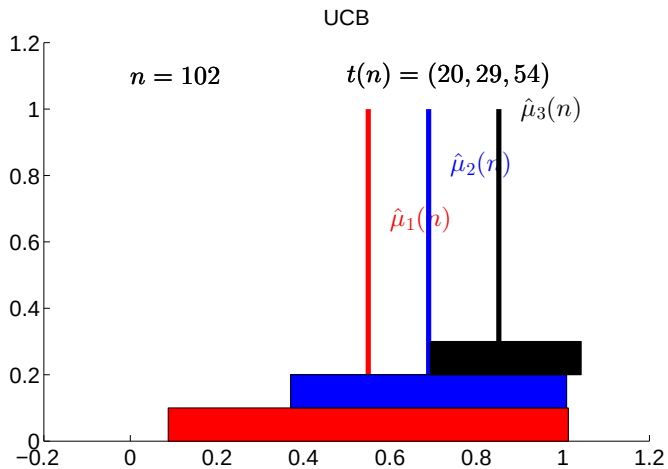
- $Z_x(n) \sim \text{Beta}(t_x(n)\hat{\theta}_x(n) + 1, t_x(n)(1 - \hat{\theta}_x(n)) + 1)$
- $x_{n+1} \in \arg\max_x Z_x(n)$

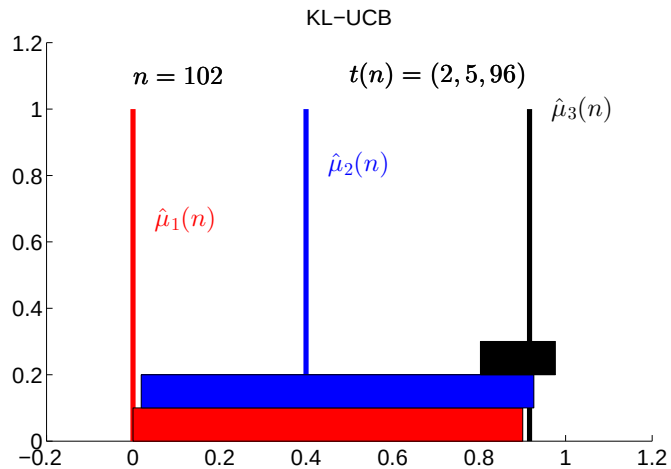
Theorem (Kaufmann'12 , Agrawal'12)

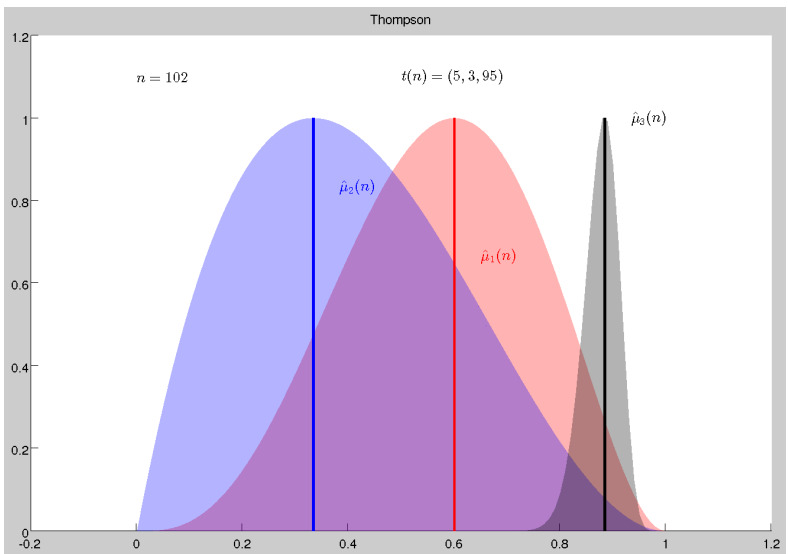
Thompson sampling is asymptotically optimal. If $\theta_x < \theta^\star$ then:

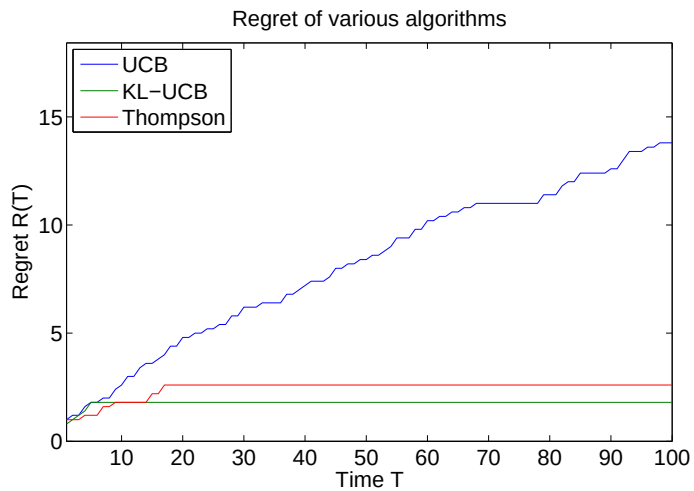
$$\limsup_{T \to \infty} \frac{\mathbb{E}[t_x(T)]}{\log(T)} \leq \frac{1}{I(\theta_x, \theta^\star)}.$$

Illustration of algorithms (one sample path)









The EXP3 algorithm

- Adversarial setting: arm selection must be randomized
- At time n, select x_n with distribution $p(n) = (p_1(n), ..., p_K(n))$
- Reward estimate for x (unbiased):

$$R_x(n) = \sum_{n'=1}^{n} \tilde{r}_x(n') = \sum_{n'=1}^{n} \frac{r_n}{p_x(n)} \mathbf{1}\{x_n = x\}$$

- Action distribution, $p_x(n) \propto \exp(\eta R_x(n))$ with $\eta > 0$ fixed.
- Favor actions with good historical rewards + explore a bit: $p(n)$ is a soft approximation to the max function
- For small η, EXP3 is the replicator dynamics (!)

Regret of EXP3

Theorem

Under EXP3 with $\eta = \sqrt{\frac{2\log(K)}{KT}}$*, the regret is upper bounded by:*

$$R^{\pi}(T) \leq \sqrt{2TK\log(K)}$$

- Larger exponent in T, but smaller in K
- Suggests two regimes for (K, T): Stochastic regime vs. Adversarial regime
- Matching lower bound: consider a stochastic adversary with close arms

Bibliography

Discrete bandits

- Thompson, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, 1933
- Robbins, Some aspects of the sequential design of experiments, 1952
- Lai and Robbins. Asymptotically efficient adaptive allocation rules, 1985
- Lai. Adaptive treatment allocation and the multi-armed bandit problem, 1987
- Gittins, Bandit Processes and Dynamic Allocation Indices, 1989
- Auer, Cesa-Bianchi and Fischer, Finite time analysis of the multiarmed bandit problem. 2002.
- Garivier and Moulines, On upper-confidence bound policies for non-stationary bandit problems, 2008
- Slivkins and Upfal, Adapting to a changing environment: the brownian restless bandits, 2008

Bibliography

- Garivier and Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond, 2011
- Honda and Takemura, An Asymptotically Optimal Bandit Algorithm for Bounded Support Models, 2010

Discrete bandits with correlated arms

- Anantharam, Varaiya, and Walrand, Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays, 1987
- Graves and Lai Asymptotically efficient adaptive choice of control laws in controlled Markov chains, 1997
- György, Linder, Lugosi and Ottucsák, The on-line shortest path problem under partial monitoring, 2007
- Yu and Mannor, Unimodal bandits, 2011
- Cesa-Bianchi and Lugosi, Combinatorial bandits, 2012.

Bibliography

- Gai, Krishnamachari, and Jain, Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations, 2012
- Chen, Wang and Yuan. Combinatorial multi-armed bandit: General framework and applications, 2013
- Combes and Proutiere, Unimodal bandits: Regret lower bounds and optimal algorithms, 2014
- Magureanu, Combes, and Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms, 2014.

Thompson Sampling

- Chapelle and Li, An Empirical Evaluation of Thompson Sampling, 2011
- Korda, Kaufmann and Munos, Thompson Sampling: an asymptotically optimal finite-time analysis, 2012
- Korda, Kaufmann and Munos, Thompson Sampling for one-dimensional exponential family bandits, 2013.

Bibliography

- Agrawal and Goyal, Further optimal regret bounds for Thompson Sampling, 2013.
- Agrawal and Goyal, Thompson Sampling for contextual bandits with linear payoffs, June 2013.

Discrete adversarial bandits

- Auer, Cesa-Bianchi, Freund and Schapire, The non-stochastic multi-armed bandit, 2002

Continuous Bandits (Lipschitz)

- R. Agrawal, The continuum-armed bandit problem, 1995
- Auer, Ortner, and Szepesvári, Improved rates for the stochastic continuum-armed bandit problem, 2007
- Bubeck, Munos, Stoltz, and Szepesvári, Online optimization in x-armed bandits, 2008

Bibliography

- Kleinberg. Nearly tight bounds for the continuum-armed bandit problem, 2004
- Kleinberg, Slivkins, and Upfal, Multi-armed bandits in metric spaces, 2008
- Bubeck, Stoltz and Yu, Lipschitz bandits without the Lipschitz constant, 2011

Continuous Bandits (strongly convex)

- Cope, Regret and convergence bounds for a class of continuum-armed bandit problems, 2009
- Flaxman, Kalai, and McMahan, Online convex optimization in the bandit setting: gradient descent without a gradient, 2005
- Shamir, On the complexity of bandit and derivative-free stochastic convex optimization, 2013
- Agarwal, Foster, Hsu, Kakade, and Rakhlin, Stochastic convex optimization with bandit feedback, 2013.

Bibliography

Continuous Bandits (linear)

- Dani, Hayes, and Kakade, Stochastic linear optimization under bandit feedback, 2008
- Rusmevichientong and Tsitsiklis, Linearly Parameterized Bandits, 2010
- Abbasi-Yadkori, Pal, Szepesvári, Improved Algorithms for Linear Stochastic Bandits, 2011

Best Arm identification

- Mannor and Tsitsiklis, The sample complexity of exploration in the multi-armed bandit problem, 2004
- Even-Dar, Mannor, Mansour, Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems, 2006
- Audibert, Bubeck, and Munos, Best arm identification in multi-armed bandits, 2010.

Bibliography

- Kalyanakrishnan, Tewari, Auer, and Stone, Pac subset selection in stochastic multi-armed bandits, 2012
- Kaufmann and Kalyanakrishnan, Information complexity in bandit subset selection, 2013
- Kaufmann, Garivier and Cappé, On the Complexity of A/B Testing, 2014

Infinite Bandits

- Berry, Chen, Zame, Heath, and Shepp, Bandit problems with infinitely many arms, 1997.
- Wang, Audibert, and Munos, Algorithms for infinitely many-armed bandits, 2008.
- Bonald and Proutiere, Two-Target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards, 2013