Introduction to Bandits: Algorithms and Theory

Jean-Yves Audibert[1,2] & Rémi Munos[3]

1. Université Paris-Est, LIGM, Imagine,
2. CNRS/École Normale Supérieure/INRIA, LIENS, Sierra
3. INRIA Sequential Learning team, France

ICML 2011, Bellevue (WA), USA

Outline

- Bandit problems and applications
- Bandits with small set of actions
  - Stochastic setting
  - Adversarial setting
- Bandits with large set of actions
  - unstructured set
  - structured set
    - linear bandits
    - Lipschitz bandits
    - tree bandits
  - Extensions

Bandit game

Parameters available to the forecaster:
the number of arms (or actions) K and the number of rounds n
Unknown to the forecaster: the way the gain vectors $g_t = (g_{1,t}, \ldots, g_{K,t}) \in [0,1]^K$ are generated

For each round $t = 1, 2, \ldots, n$

1. the forecaster chooses an arm $I_t \in \{1, \ldots, K\}$
2. the forecaster receives the gain $g_{I_t,t}$
3. only $g_{I_t,t}$ is revealed to the forecaster

Cumulative regret goal: maximize the cumulative gains obtained. More precisely, minimize

$$R_n = \left(\max_{i=1,\ldots,K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} \right) - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t}$$

where $\mathbb{E}$ comes from both a possible stochastic generation of the gain vector and a possible randomization in the choice of I_t

Stochastic and adversial environments

- **Stochastic environment:** the gain vector g_t is sampled from an unknown product distribution $\nu_1 \otimes \ldots \otimes \nu_K$ on $[0,1]^K$, that is $g_{i,t} \sim \nu_i$.
- **Adversarial environment:** the gain vector g_t is chosen by an adversary (which, at time t, knows all the past, but not I_t)

Numerous variants

- different environments: adversarial, "stochastic", non-stationary
- different targets: cumulative regret, simple regret, tracking the best expert
- Continuous or discrete set of actions
- extension with additional rules: varying set of arms, pay-per-observation, ...

Various applications

- Clinical trials (Thompson, 1933)
- Ads placement on webpages
- Nash equilibria (traffic or communication networks, agent simulation, tic-tac-toe phantom, ...)
- Game-playing computers (Go, urban rivals, ...)
- Packet routing, itinerary selection
- ...

Outline

- Bandit problems and applications
- Bandits with small set of actions
  - Stochastic setting
  - Adversarial setting
- Bandits with large set of actions
  - unstructured set
  - structured set
    - linear bandits
    - Lipschitz bandits
    - tree bandits
  - Extensions

Stochastic bandit game (Robbins, 1952)

Parameters available to the forecaster: K and n

Parameters unknown to the forecaster: the reward distributions $\nu_1, \ldots, \nu_K$ of the arms (with respective means $\mu_1, \ldots, \mu_K$)

For each round $t = 1, 2, \ldots, n$

1. the forecaster chooses an arm $I_t \in \{1, \ldots, K\}$
2. the environment draws the gain vector $g_t = (g_{1,t}, \ldots, g_{K,t})$ according to $\nu_1 \otimes \cdots \otimes \nu_K$
3. the forecaster receives the gain $g_{I_t,t}$

Notation: $i^* = \arg\max_{i=1,\ldots,K} \mu_i \qquad \mu^* = \max_{i=1,\ldots,K} \mu_i$

$\Delta_i = \mu^* - \mu_i, \qquad T_i(n) = \sum_{t=1}^n \mathbb{1}_{I_t = i}$

Cumulative regret: $\hat{R}_n = \sum_{t=1}^n g_{i^*,t} - \sum_{t=1}^n g_{I_t,t}$

Goal: minimize the expected cumulative regret

$$R_n = \mathbb{E}\hat{R}_n = n\mu^* - \mathbb{E}\sum_{t=1}^n g_{I_t,t} = n\mu^* - \mathbb{E}\sum_{i=1}^K T_i(n)\mu_i = \sum_{i=1}^K \Delta_i \mathbb{E} T_i(n)$$

A simple policy: ε-greedy

For simplicity, all rewards are in $[0,1]$

- Playing the arm with highest empirical mean does not work
- ε-greedy: at time t,
  - with probability $1-\varepsilon_t$, play the arm with highest empirical mean
  - with probability ε_t, play a random arm
- Theoretical guarantee: (Auer, Cesa-Bianchi, Fischer, 2002)
  - Let $\Delta = \min_{i:\Delta_i>0} \Delta_i$ and consider $\varepsilon_t = \min(\frac{6K}{\Delta^2 t}, 1)$
  - When $t \geq \frac{6K}{\Delta^2}$, the probability of choosing a suboptimal arm i is bounded by $\frac{C}{\Delta^2 t}$ for some constant $C > 0$
  - As a consequence, $\mathbb{E}[T_i(n)] \leq \frac{C}{\Delta^2} \log n$ and $R_n \leq \sum_{i:\Delta_i>0} \frac{C\Delta_i}{\Delta^2} \log n$
    $\longrightarrow$ logarithmic regret
- drawbacks:
  - naive exploration for $K > 2$: no distinction of sub-optimal arms
  - requires knowledge of Δ
  - outperformed by UCB policy in practice

Optimism in face of uncertainty

- At time t, from past observations and some probabilistic argument, you have an upper confidence bound (UCB) on the expected rewards.

- Simple implementation:

  play the arm having the largest UCB !

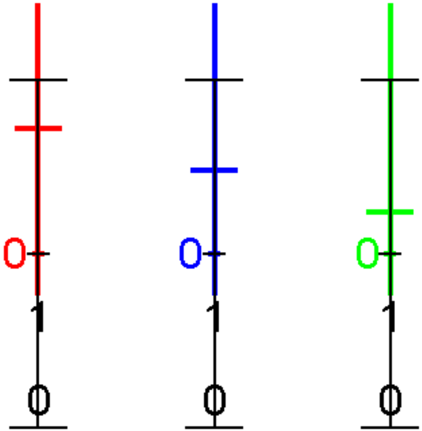
Why does it make sense?

- Could we stay a long time drawing a wrong arm?

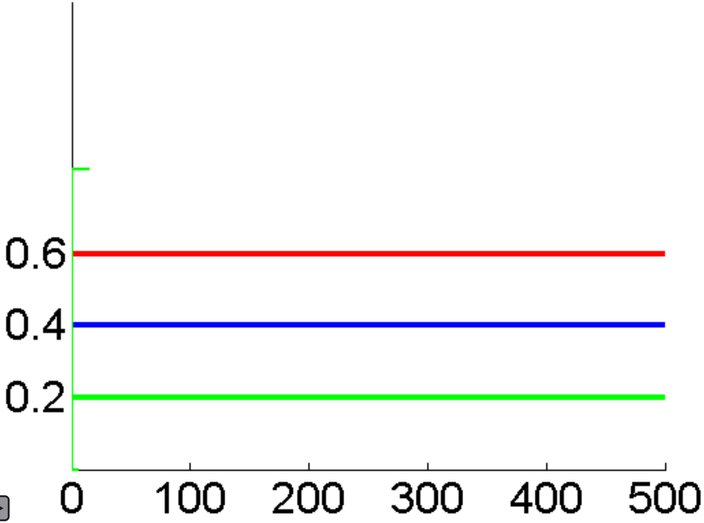
  No, since:

  - The more we draw a wrong arm i the closer the UCB gets to the expected reward μ_i,
  - $\mu_i < \mu^* \leq$ UCB on μ^*

Illustration of UCB policy



Confidence intervals vs sampling times



Hoeffding-based UCB (Auer, Cesa-Bianchi, Fischer, 2002)

- Hoeffding's inequality: Let $X, X_1, \ldots, X_m$ be i.i.d. r.v. taking their values in $[0,1]$. For any $\varepsilon > 0$, with probability at least $1-\varepsilon$, we have

$$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{\log(\varepsilon^{-1})}{2m}}$$

- UCB1 policy: at time t, play

$$I_t \in \operatorname*{arg\,max}_{i\in\{1,\ldots,K\}} \left\{ \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log t}{T_i(t-1)}} \right\},$$

where $\hat{\mu}_{i,t-1} = \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s}$

- Regret bound:

$$R_n \leq \sum_{i\neq i^*} \min\left(\frac{10}{\Delta_i}\log n, n\Delta_i\right)$$

Hoeffding-based UCB (Auer, Cesa-Bianchi, Fischer, 2002)

- Hoeffding's inequality: Let $X_1, \dots, X_m$ be i.i.d. r.v. taking their values in $[0,1]$. For any $\varepsilon > 0$, with probability at least $1-\varepsilon$, we have
$$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{\log(\varepsilon^{-1})}{2m}}$$
- UCB1 policy: At time t, play
$$I_t \in \operatorname*{arg\,max}_{i \in \{1,\dots,K\}} \left\{ \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log t}{T_i(t-1)}} \right\},$$
where $\hat{\mu}_{i,t-1} = \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s}$
- UCB1 is an anytime policy (it does not need to know n to be implemented)

Hoeffding-based UCB (Auer, Cesa-Bianchi, Fischer, 2002)

- Hoeffding's inequality: Let $X_1, \ldots, X_m$ be i.i.d. r.v. taking their values in $[0,1]$. For any $\varepsilon > 0$, with probability at least $1-\varepsilon$, we have

$$\mathbb{E}X \le \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{\log(\varepsilon^{-1})}{2m}}$$

- UCB1 policy: At time t, play

$$I_t \in \operatorname*{arg\,max}_{i \in \{1,\ldots,K\}} \left\{ \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log t}{T_i(t-1)}} \right\},$$

where $\hat{\mu}_{i,t-1} = \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} X_{i,s}$

- UCB1 corresponds to $2\log t = \frac{\log(\varepsilon^{-1})}{2}$, hence $\varepsilon = 1/t^4$
- Critical confidence level $\varepsilon = 1/t$ (Lai & Robbins, 1985; Agrawal, 1995; Burnetas & Katehakis, 1996; Audibert, Munos, Szepesvári, 2009; Honda & Takemura, 2010)

Better confidence bounds imply smaller regret

- Hoeffding's inequality $\frac{1}{t}$-confidence bound

$$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{\log(t)}{2m}}$$

- Bernstein's inequality $\frac{1}{t}$-confidence bound

$$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{2\log(t)\mathbb{V}\mathrm{ar}\,X}{m}} + \frac{\log(t)}{3m}$$

- Empirical Bernstein's inequality $\frac{3}{t}$-confidence bound

$$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{2\log(t)\widehat{\mathbb{V}\mathrm{ar}\,X}}{m}} + \frac{8\log(t)}{3m}$$

(Audibert, Munos, Szepesvári, 2009; Maurer, 2009; Audibert, 2010)

- Asymptotic confidence bound leads to catastrophy:

$$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{\widehat{\mathbb{V}\mathrm{ar}\,X}}{m}}x \quad \text{with } x \text{ s.t. } \int_x^{+\infty} \frac{e^{-u^2/2}}{\sqrt{2\pi}}du = \frac{1}{t}$$

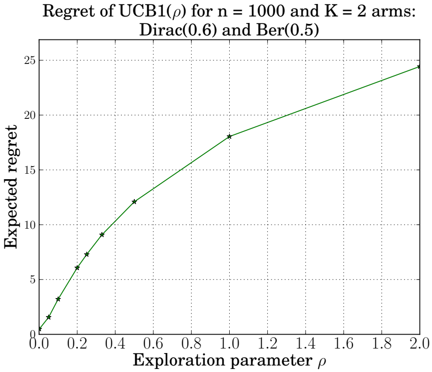
Better confidence bounds imply smaller regret

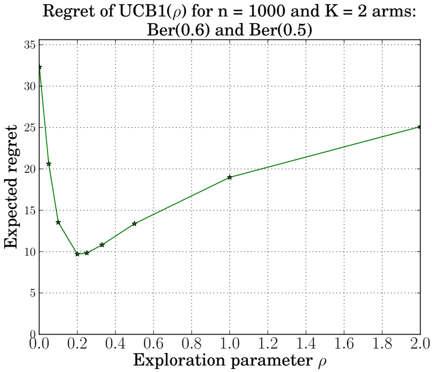
Hoeffding-based UCB	empirical Bernstein-based UCB
$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{\log(\varepsilon^{-1})}{2m}}$	$\mathbb{E}X \leq \frac{1}{m}\sum_{s=1}^{m} X_s + \sqrt{\frac{2\log(\varepsilon^{-1})\widehat{\mathbb{V}\mathrm{ar}\, X}}{m}} + \frac{8\log(\varepsilon^{-1})}{3m}$
$R_n \leq \sum_{i \neq i^*} \min\left(\frac{c}{\Delta_i}\log n, n\Delta_i\right)$	$R_n \leq \sum_{i \neq i^*} \min\left(c\left(\frac{\mathbb{V}\mathrm{ar}\,\nu_i}{\Delta_i} + 1\right)\log n, n\Delta_i\right)$

Tuning the exploration: simple vs difficult bandit problems

- UCB1(ρ) policy: At time t, play

$$I_t \in \arg\max_{i \in \{1,\dots,K\}} \left\{ \hat{\mu}_{i,t-1} + \sqrt{\frac{\rho \log t}{T_i(t-1)}} \right\},$$



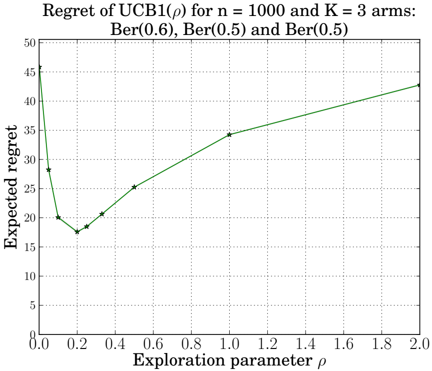


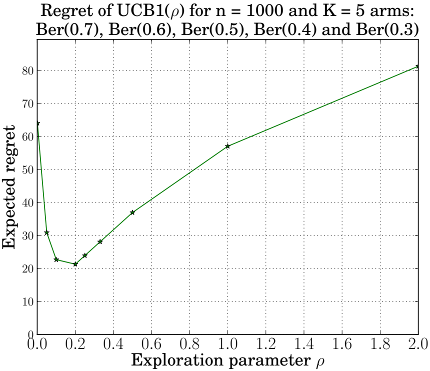
Tuning the exploration parameter: from theory to practice

- Theory:
  - for $\rho < 0.5$, UCB1(ρ) has polynomial regret
  - for $\rho > 0.5$, UCB1(ρ) has logarithmic regret

(Audibert, Munos, Szepesvári, 2009; Bubeck, 2010)

- Practice: $\rho = 0.2$ seems to be the best default value for $n < 10^8$





Deviations of UCB1 regret

- UCB1 policy: At time t, play

$$I_t \in \arg\max_{i\in\{1,\dots,K\}} \left\{ \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log t}{T_i(t-1)}} \right\}$$

- Inequality of the form $\mathbb{P}(\hat{R}_n > \mathbb{E}\hat{R}_n + \gamma) \leq ce^{-c\gamma}$ does not hold!
- If the smallest reward observable from the optimal arm is smaller than the mean reward of the second optimal arm, then the regret of UCB1 satisfies: for any $C > 0$, there exists $C' > 0$ such that for any $n \geq 2$

$$\mathbb{P}(\hat{R}_n > \mathbb{E}\hat{R}_n + C\log n) > \frac{1}{C'(\log n)^{C'}}$$

(Audibert, Munos, Szepesvári, 2009)

Anytime UCB policies has a heavy-tailed regret

- For some difficult bandit problems, the regret of UCB1 satisfies: for any $C > 0$, there exists $C' > 0$ such that for any $n \geq 2$

$$\mathbb{P}(\hat{R}_n > \mathbb{E}\hat{R}_n + C \log n) > \frac{1}{C'(\log n)^{C'}} \qquad (\star)$$

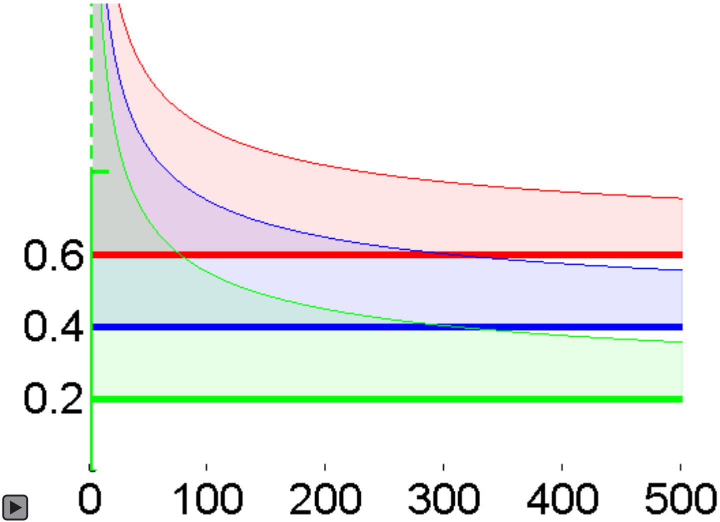
- UCB-H$_{\text{orizon}}$ policy: At time t, play

$$I_t \in \operatorname*{arg\,max}_{i \in \{1,\dots,K\}} \left\{ \hat{\mu}_{i,t-1} + \sqrt{\frac{2 \log n}{T_i(t-1)}} \right\},$$

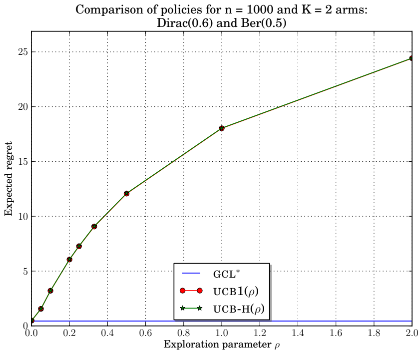
(Audibert, Munos, Szepesvári, 2009)

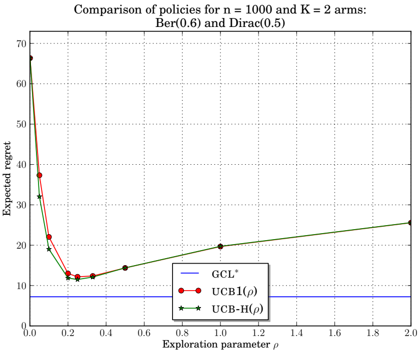
- UCB-H satisfies $\mathbb{P}(\hat{R}_n > \mathbb{E}\hat{R}_n + C \log n) \leq \frac{C}{n}$ for some $C > 0$
- $(\star)$ = unavoidable for anytime policies (Salomon, Audibert, 2011)

Comparison of UCB1 (solid lines) and UCB-H (dotted lines)



Comparison of UCB1(ρ) and UCB-H(ρ) in expectation



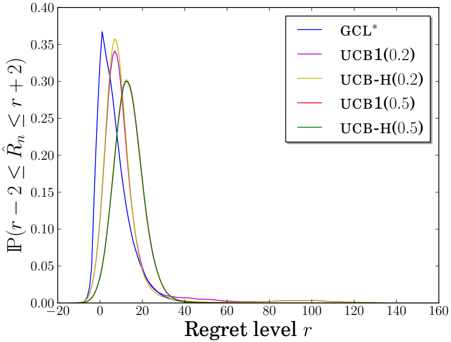


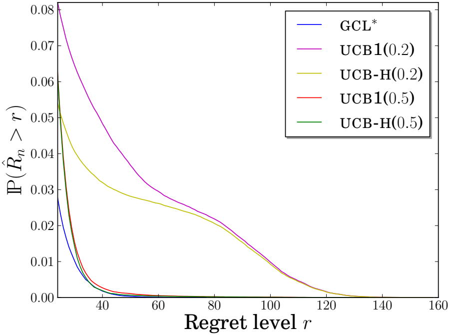
Left: Dirac(0.6) vs Bernoulli(0.5)

Right: Bernoulli(0.6) vs Dirac(0.5)

Comparison of UCB1(ρ) and UCB-H(ρ) in deviations

- For $n = 1000$ and $K = 2$ arms: Bernoulli(0.6) and Dirac(0.5)





Left: smoothed probability mass function. Right: tail distribution of the regret.

Knowing the horizon: theory and practice

- Theory: use UCB-H to avoid heavy tails of the regret

- Practice: Theory is right. Besides, thanks to this robustness, the expected regret of UCB-H(ρ) consistently outperforms the expected regret of UCB1(ρ). However:
  - the gain is small.
  - a better way to have small regret tails is to take larger ρ

Knowing μ^*

- Hoeffding-based GCL* policy: play each arm once, then play

$$I_t \in \underset{i\in\{1,\dots,K\}}{\operatorname{argmin}}\, T_i(t-1)\,(\mu^* - \hat{\mu}_{i,t-1})_+^2$$

(Salomon, Audibert, 2011)

- Underlying ideas:
  - compare p-values of the K tests: $H_0 = \{\mu_i = \mu^*\}$, $i \in \{1,\dots,K\}$
  - the p-values are estimated using Hoeffding's inequality

$$\mathbb{P}_{H_0}\big(\hat{\mu}_{i,t-1} \le \hat{\mu}_{i,t-1}^{(obs)}\big) \lessapprox \exp\Big(-2T_i(t-1)\Big(\mu^* - \hat{\mu}_{i,t-1}^{(obs)}\Big)_+^2\Big)$$

  - play the arm for which we have the Greatest Confidence Level that it is the optimal arm.
- Advantages:
  - logarithmic expected regret
  - anytime policy
  - regret with a subexponential right-tail
  - parameter-free policy !
  - outperforms any other Hoeffding-based algorithm !

From Chernoff's inequality to KL-based algorithms

- Let $\mathcal{K}(p, q)$ be the Kullback-Leibler divergence between Bernoulli distributions of respective parameter p and q
- Let $X_1, \ldots, X_T$ be i.i.d. r.v. of mean μ, and taking their values in $[0, 1]$. Let $\bar{X} = \frac{1}{T}\sum_{i=1}^{T} X_i$. For any $\gamma > 0$

  $$\mathbb{P}(\bar{X} \leq \mu - \gamma) \leq \exp\big(-T\,\mathcal{K}(\mu-\gamma, \mu)\big).$$

  In particular, we have

  $$\mathbb{P}(\bar{X} \leq \bar{X}^{(obs)}) \leq \exp\Big(-T\,\mathcal{K}\big(\min(\bar{X}^{(obs)}, \mu), \mu\big)\Big).$$

- If μ^* is known, using the same idea of comparing the *p*-values of the tests $H_0 = \{\mu_i = \mu^*\}$, $i \in \{1, \ldots, K\}$, we get the Chernoff-based GCL* policy: play each arm once, then play

  $$I_t \in \operatorname*{argmin}_{i \in \{1,\ldots,K\}} T_i(t-1)\,\mathcal{K}\big(\min(\hat{\mu}_{i,t-1}, \mu^*), \mu^*\big)$$

Back to unknown μ^*

- When μ^* is unknown, the principle

  playing the arm for which we have the greatest confidence level that it is the optimal arm

  is replaced by

  being optimistic in face of uncertainty:

  - an arm i is represented by the highest mean of a distribution ν for which the hypothesis $H_0 = \{\nu_i = \nu\}$ has a p-value greater than $\frac{1}{t^\beta}$ (critical $\beta = 1$, as usual)
  - the arm with the highest index (=UCB) is played

KL-based algorithms when μ^* is unknown

- Approximating the p-value using Sanov's theorem is tightly linked to the DMED policy, which satisfies

$$\limsup_{n \to +\infty} \frac{R_n}{\log n} \le \frac{\Delta_i}{\inf_{\nu : \mathbb{E}_{X \sim \nu} X \ge \mu^*} \mathcal{K}(\nu_i, \nu)}$$

(Burnetas & Katehakis, 1996; Honda & Takemura, 2010)

It matches the lower bound

$$\liminf_{n \to +\infty} \frac{R_n}{\log n} \ge \frac{\Delta_i}{\inf_{\nu : \mathbb{E}_{X \sim \nu} X \ge \mu^*} \mathcal{K}(\nu_i, \nu)}$$

(Lai & Robbins, 1985; Burnetas & Katehakis, 1996)

- Approximating the p-value using non-asymptotic version of Sanov's theorem leads to the KL-UCB (Cappé & Garivier, COLT 2011) and the $\mathcal{K}$-strategy (Maillard, Munos, Stoltz, COLT 2011)

Outline

- Bandit problems and applications
- Bandits with small set of actions
  - Stochastic setting
  - Adversarial setting
- Bandits with large set of actions
  - unstructured set
  - structured set
    - linear bandits
    - Lipschitz bandits
    - tree bandits
  - Extensions

Adversarial bandit

Parameters: the number of arms K and the number of rounds n

For each round $t = 1, 2, \dots, n$

1. the forecaster chooses an arm $I_t \in \{1, \dots, K\}$, possibly with the help of an external randomization
2. the adversary chooses a gain vector $g_t = (g_{1,t}, \dots, g_{K,t}) \in [0,1]^K$
3. the forecaster receives and observes only the gain $g_{I_t,t}$

Goal: Maximize the cumulative gains obtained. We consider the regret:

$$R_n = \left(\max_{i=1,\dots,K} \mathbb{E} \sum_{t=1}^{n} g_{i,t} \right) - \mathbb{E} \sum_{t=1}^{n} g_{I_t,t},$$

- In full information, step 3. is replaced by the forecaster receives $g_{I_t,t}$ and observes the full gain vector g_t
- In both settings, the forecaster should use an external randomization to have $o(n)$ regret.

Adversarial setting in full information: an optimal policy

- Cumulative reward on $[1, t-1]$: $G_{i,t-1} = \sum_{s=1}^{t-1} g_{i,s}$
- Follow-the-leader: $I_t \in \arg\max_{i \in \{1,\dots,K\}} G_{i,t-1}$ is a bad policy
- An "optimal" policy is obtained by considering

$$p_{i,t} = \mathbb{P}(I_t = i) = \frac{e^{\eta G_{i,t-1}}}{\sum_{k=1}^{K} e^{\eta G_{k,t-1}}}$$

- For this policy, $R_n \leq \frac{n\eta}{8} + \frac{\log K}{\eta}$
- in particular, for $\eta = \sqrt{\frac{8 \log K}{n}}$, we have $R_n \leq \sqrt{\frac{n \log K}{2}}$

(Littlestone, Warmuth, 1994; Long, 1996; Bylanger, 1997; Cesa-Bianchi, 1999)

Proof of the regret bound

$$p_{i,t} = \mathbb{P}(I_t = i) = \frac{e^{\eta G_{i,t-1}}}{\sum_{k=1}^{K} e^{\eta G_{k,t-1}}}$$

$$\begin{aligned}
&\mathbb{E}\sum_t g_{I_t,t} \\
=&\mathbb{E}\sum_t \sum_i p_{i,t} g_{i,t} \\
=&\mathbb{E}\sum_t \left(-\frac{1}{\eta} \log \sum_i p_{i,t} e^{\eta(g_{i,t} - \sum_j p_{j,t} g_{j,t})} + \frac{1}{\eta} \log \sum_i p_{i,t} e^{\eta g_{i,t}} \right) \\
=&\mathbb{E}\sum_t \left(-\frac{1}{\eta} \log \mathbb{E} e^{\eta(V_t - \mathbb{E} V_t)} + \frac{1}{\eta} \log \frac{\sum_i e^{\eta G_{i,t}}}{\sum_i e^{\eta G_{i,t-1}}} \right) \qquad \mathbb{P}(V_t = g_{i,t}) = p_{i,t} \\
\geq&\mathbb{E}\left(-\sum_t \frac{\eta}{8} \right) + \frac{1}{\eta} \mathbb{E} \log \frac{\sum_j e^{\eta G_{j,n}}}{\sum_j e^{\eta G_{j,0}}} \qquad \text{Hoeffding's inequality} \\
\geq& -\frac{n\eta}{8} + \frac{1}{\eta} \mathbb{E} \log \frac{e^{\eta \max_j G_{j,n}}}{K} = -\frac{n\eta}{8} - \frac{\log K}{\eta} + \mathbb{E} \max_j G_{j,n}
\end{aligned}$$

Adapting the exponentially weighted forecaster

- In bandit setting, $G_{i,t-1}$, $i = 1, \dots, K$ are not observed

  Trick = estimate them

- Precisely, $G_{i,t-1}$ is estimated by $\tilde{G}_{i,t-1} = \sum_{s=1}^{t-1} \tilde{g}_{i,s}$ with

$$\tilde{g}_{i,s} = 1 - \frac{1 - g_{I_s,s}}{p_{I_s,s}} \mathbb{1}_{I_s=i}.$$

Note that $\mathbb{E}_{I_s \sim p_s} \tilde{g}_{i,s} = 1 - \sum_{k=1}^{K} p_{k,s} \frac{1 - g_{k,s}}{p_{k,s}} \mathbb{1}_{k=i} = g_{i,s}$

$$p_{i,t} = \mathbb{P}(I_t = i) = \frac{e^{\eta \tilde{G}_{i,t-1}}}{\sum_{k=1}^{K} e^{\eta \tilde{G}_{k,t-1}}}$$

- For this policy, $R_n \leq \frac{nK\eta}{2} + \frac{\log K}{\eta}$
  - In particular, for $\eta = \sqrt{\frac{2 \log K}{nK}}$, we have $R_n \leq \sqrt{2nK \log K}$

(Auer, Cesa-Bianchi, Freund, Schapire, 1995)

Implicitly Normalized Forecaster (Audibert, Bubeck, 2010)

Let $\psi : \mathbb{R}_-^* \to \mathbb{R}_+^*$ increasing, convex, twice continuously differentiable, and s.t. $[\frac{1}{K}, 1] \subset \psi(\mathbb{R}_-^*)$

Let p_1 be the uniform distribution over $\{1, \ldots, K\}$

For each round $t = 1, 2, \ldots,$

- $I_t \sim p_t$
- Compute $p_{t+1} = (p_{1,t+1}, \ldots, p_{K,t+1})$ where

$$p_{i,t+1} = \psi(\tilde{G}_{i,t} - C_t)$$

where C_t is the unique real number s.t. $\sum_{i=1}^K p_{i,t+1} = 1$

Minimax policy

- $\psi(x) = \exp(\eta x)$ with $\eta > 0$; this corresponds exactly to the exponentially weighted forecaster
- $\psi(x) = (-\eta x)^{-q}$ with $q > 1$ and $\eta > 0$; this is a new policy which is minimax optimal: for $q = 2$ and $\eta = \sqrt{2n}$, we have

$$R_n \leq 2\sqrt{2nK}$$

(Audibert, Bubeck, 2010; Audibert, Bubeck, Lugosi, 2011)

while for any strategy, we have

$$\sup R_n \geq \frac{1}{20}\sqrt{nK}$$

(Auer, Cesa-Bianchi, Freund, Schapire, 1995)

Outline

- Bandit problems and applications
- Bandits with small set of actions
  - Stochastic setting
  - Adversarial setting
- Bandits with large set of actions
  - unstructured set
  - structured set
    - linear bandits
    - Lipschitz bandits
    - tree bandits
  - Extensions