Introduction to Bandits: Algorithms and Theory
Part 2: Bandits with large sets of actions

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

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

ICML 2011, Bellevue (WA), USA

K-armed bandit, with $K = 4$

At each round t, select a tap. Optimize quality of n selected beers.

Bandit with a large number of arms

Goal: optimize the beer you drink before you get drunk...

Part 2: Bandits with a large set of actions

The number of arms is larger then the number of rounds.

- Unstructured set of actions:
  1. Many-armed bandits
- Structured set of actions:
  2. Linear bandits
  3. Lipschitz bandits
  4. Bandits in trees
- Extensions

We consider the “optimism in the face of uncertainty” principle in stochastic environments.

A few references on bandits 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]

and I surely missed many relevant references...

Unstructured set of actions: Examples

There is an infinite number of arms. The rewards received so far do not tell us anything about the value of unobserved arms.
Example: Enjoy Parisian restaurants.
Each day, select a restaurant:

- among the ones where you have already been
  - because it is good (Exploitation)
  - or not well known (Exploration)
- or choose a new one randomly (Discovery)

Other examples:

- Mining for valuable resources (such as gold or oil): exploit good wells, explore unknown wells, or start diging at a new location.
- Marketing (e.g. send catalogues to good customers, uncertain customers, or random people).

Many-armed bandits: Assumptions

We make a (probabilistic) assumption about the mean-value of any new arm.

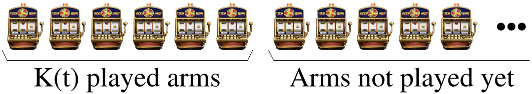
- **Usual assumption**: the distribution of the mean-reward of a new arm is known [Banks, Sundaram, 1992], [Berry, Chen, Zame, Heath, Shepp, 1997].
- **Weaker assumption:** [Wang, Audibert, Munos, 2008] We know $\beta > 0$ such that

$$\mathbb{P}(\mu(\text{new arm}) > \mu^* - \varepsilon) = \Theta(\varepsilon^\beta),$$

β characterizes the probability of selecting near-optimal arms

Large $\beta \implies$ small chance of pulling good arm, thus one needs to pull many arms. And vice-versa.

The UCB-AIR strategy



UCB with Arm Increasing Rule [Wang, Audibert, Munos, 2008]:

- $K(0) = 0$. At time $t+1$, pull a new arm if

$$K(t) < \begin{cases} t^{\frac{\beta}{2}} & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\ t^{\frac{\beta}{\beta+1}} & \text{if } \beta \geq 1 \text{ or } \mu^* = 1 \end{cases}$$

- Otherwise, apply UCB–V [Audibert, Munos, Szepesvári, 2009] on the $K(t)$ drawn arms, i.e., play

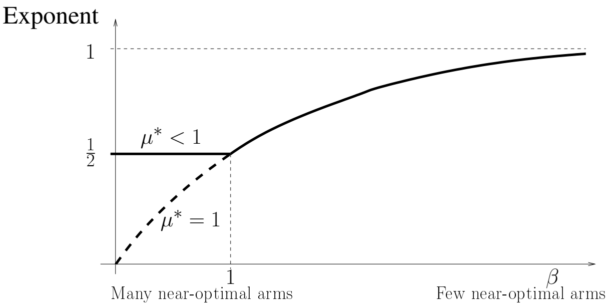
$$\underset{1\leq k\leq K(t)}{\operatorname{argmax}} \underbrace{\widehat{\mu}_{k,t}}_{\text{empirical rewards}} + \underbrace{\sqrt{\frac{2\widehat{V}_{k,t}\mathcal{E}_t}{T_k(t)}} + \frac{3\mathcal{E}_t}{T_k(t)}}_{\text{Confidence interval}},$$

with exploration sequence: $c\log(\log t) \leq \mathcal{E}_t \leq \log t$.

Regret analysis of UCB-AIR

Upper bound on the regret of UCB-AIR:

$$\mathbb{E}R_n = \begin{cases} \widetilde{O}(\sqrt{n}) & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\ \widetilde{O}\left(n^{\frac{\beta}{1+\beta}}\right) & \text{if } \mu^* = 1 \text{ or } \beta \geq 1 \end{cases}$$



Lower bound: $\forall \beta > 0, \mu^* \leq 1$, for any algorithm $\mathbb{E}R_n = \Omega(n^{\frac{\beta}{1+\beta}})$.

Remarks and possible extensions

Remarks

- When $\beta > 1$ or $\mu^* = 1$ the upper and lower bounds match (up to logarithmic factor).
- Exploration-Exploitation-Discovery tradeoff:
  - **Exploitation**: Pull a good arm
  - **Exploration**: Pull an uncertain arm
  - **Discovery**: Pull a new arm
- The exploration sequence $\mathcal{E}_t$ can be of order $\log \log t$ (instead of $\log t$): discovery replaces exploration
- **Open question**: similar performance when β is unknown? (i.e. adaptive strategy that estimates β while minimizing regret).

Structured set of actions or rewards

The mapping *Arms* $\rightarrow$ *Reward* possesses some known structure:

- Linear
- Lipschitz
- Tree structure

Reward samples from observed arms provides information about unseen arms.

Linear bandits

Outline of this section:

- Linear reward function
- UCB type of algorithms: Confidence Ellipsoid
- Extensions

References: [Auer, 2002], [Dani, Hayes, Kakade, 2008], [Abbasi-Yadkori, 2009], [Rusmevichientong, Tsitsiklis, 2010], [Filippi, Cappé, Garivier, Szepesvári, 2010].

Linear mean-reward function

The set of arms $\mathcal{X}$ is a subset of $\mathbb{R}^D$.
The mean-reward function is linear: $x \in \mathcal{X} \mapsto \langle x, \alpha^* \rangle$, where $\alpha^* \in \mathbb{R}^D$ is an unknown parameter.
At each time step t,

- Select $x_t \in \mathcal{X}$,
- Observe $y_t = \langle x_t, \alpha^* \rangle + \eta_t$, where $\mathbb{E}[\eta_t | x_t] = 0$.
  (we assume the noise is bounded or sub-Gaussian).

Let $x^* = \mathrm{argmax}_{x \in \mathcal{X}} \langle x, \alpha^* \rangle$ be the best arm in $\mathcal{X}$.
Define the regret:

$$R_n = n \langle x^*, \alpha^* \rangle - \sum_{t=1}^{n} y_t.$$

No need to estimate the mean-reward of all arms, estimating α^* is enough. So the regret will scale with D and not with the number of arms (which may be infinite).

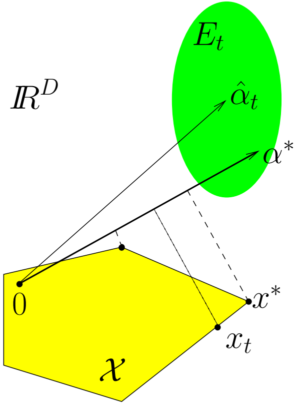
Geometric intuition

Choose $x_t \in \mathcal{X}$ and get:

$$y_t = \langle x_t, \alpha^* \rangle + \eta_t$$

(provides information about α^* along the direction x_t).

Idea: Build a high probability confidence set E_t s.t. $\alpha^* \in E_t$ w.h.p.



Play the arm $x \in \mathcal{X}$ that maximizes $\langle x, \alpha \rangle$ for some $\alpha \in E_t$.

Confidence Ball algorithms

[Dani, Hayes, Kakade, 2008]
UCB idea: Define a least-squares estimate $\widehat{\alpha}_t$ of α^*:

$$\widehat{\alpha}_t = A_t^{-1} \sum_{s=1}^{t} y_s x_s, \text{ where } A_t = \Big(\sum_{s=1}^{t} x_s x_s^T + A_0 \Big),$$

and a confidence ellipsoid E_t around $\widehat{\alpha}_t$:

$$E_t = \big\{ \alpha \in \mathbb{R}^D, \big\| \alpha - \widehat{\alpha}_t \big\|_{2,A_t} \leq \rho(t) \big\}, \text{ where } \rho(t) = c\sqrt{D} \log(t/\delta).$$

Property: w.p. $1-\delta$, $\alpha^* \in E_t$ for all $t \geq 1$.

Algorithm: At round $t+1$, select arm

$$x_{t+1} = \operatorname*{argmax}_{x \in \mathcal{X}} \max_{\alpha \in E_t} \langle x, \alpha \rangle .$$

Regret analysis

[Dani, Hayes, Kakade, 2008], [Rusmevichientong, Tsitsiklis, 2010]

Upper bounds:

- *Problem independent:* With probability $1-\delta$,

$$R_n = O\big(D\sqrt{n}(\log n/\delta)^{3/2}\big)$$

- *Problem dependent:* With probability $1-\delta$,

$$R_n = O\big(\frac{D^2}{\Delta}(\log n/\delta)^3\big),$$

  where Δ is the gap (mean reward difference between best and second best extremal points). Useful when $\mathcal{X}$ is finite or a polytope.

Lower bound: there exists a set $\mathcal{X}$ such that for any algorithm,

$$R_n = \Omega(D\sqrt{n}).$$

Possible extensions [1]

- One may consider ℓ_1-**ellipsoid** instead of ℓ_2, which yield a slightly poorer regret $\tilde{O}(D^{3/2}\sqrt{n})$ but which is more computationally efficient (computation of $\max_{\alpha \in E_t} \langle x, \alpha \rangle$ is $O(D)$).
- **Generalized Linear models** [Filippi, Cappé, Garivier, Szepesvári, 2010]:

$$y_t = \mu(\langle x_t, \alpha^*, \rangle) + \eta_t,$$

where μ is a real-valued function (such as logistic regression function, in order to deal with binary rewards). GLM-UCB selects the arm:

$$\operatorname*{argmax}_{x \in \mathcal{X}} \Big(\mu(\langle x, \widehat{\alpha}_t, \rangle) + \rho(t) \|x\|_{2, A_t^{-1}} \Big),$$

with enjoys similar performance guarantees.

Possible extensions [2]

- **Linear combination of features**:

$$y_t = \langle \varphi(x_t), \alpha^* \rangle + \eta_t,$$

  and apply previous analysis with the set of arms $\varphi(\mathcal{X})$.
- **Sparse linear bandits**: α^* is sparse. Derive algorithms that scale with $\|\alpha^*\|_0$ instead of D.
- **Open question**: is it possible to improve the upper- and lower- bounds in terms of a measure of the quantity of near-optimal states?

$$\mathcal{X}_\varepsilon = \{x \in \mathcal{X}, \langle x, \alpha^* \rangle \geq \langle x^*, \alpha^* \rangle - \varepsilon\}.$$

We now consider more general reward functions f.

$\mathcal{X}$-armed bandits

Outline of this Section:

- Gentle start: Optimization of a deterministic Lipschitz function
- Adding noise
- Relaxing Lipschitz assumption
- Hierarchical Optimistic Optimization (HOO)

References: [Agrawal, 1995], [Kleinberg, 2004], [Auer, Ortner, Szepesvári, 2007], [Kleinberg, Slivkins, Upfall, 2008], [Bubeck, Munos, Stoltz, Szepesvári, 2011].

Optimization of a deterministic Lipschitz function

Problem: Find online the maximum of $f : \mathcal{X} \to \mathbb{R}$, assumed to be Lipschitz: $|f(x) - f(y)| \leq \ell(x, y)$.

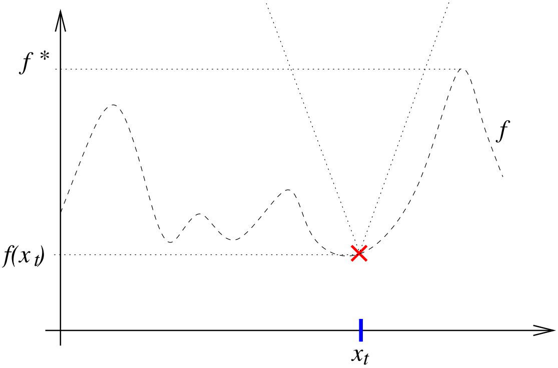
- At each time step t, select $x_t \in \mathcal{X}$
- Observe $f(x_t)$
- Goal: maximize the sum of rewards.

Define the cumulative regret

$$R_n = \sum_{t=1}^{n} \left[f^* - f(x_t) \right],$$

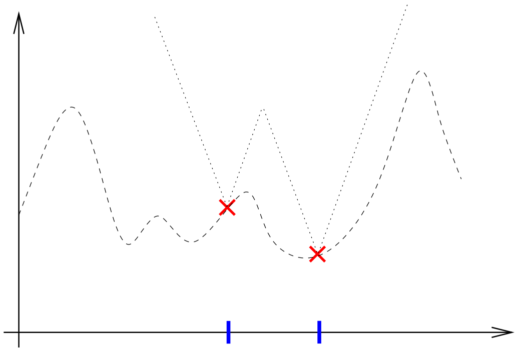
where $f^* = \sup_{x \in \mathcal{X}} f(x)$

Example in 1d



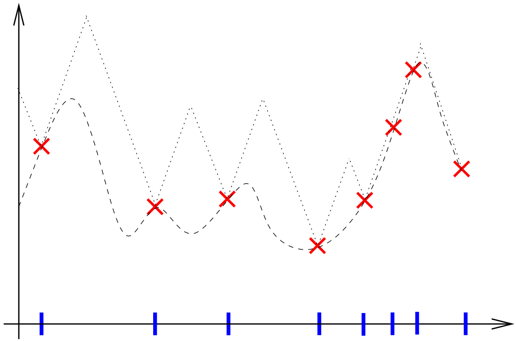
Lipschitz property $\rightarrow$ the evaluation of f at x_t provides a first upper-bound on f.

Example in 1d (continued)



New point $\rightarrow$ refined upper-bound on f.

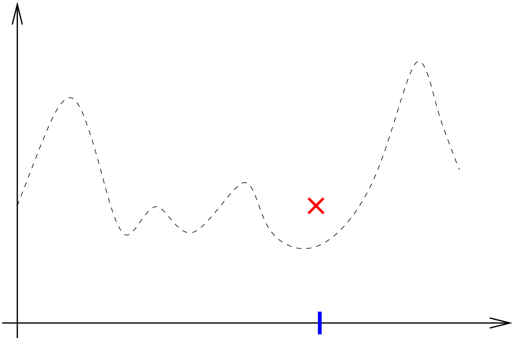
Example in 1d (continued)



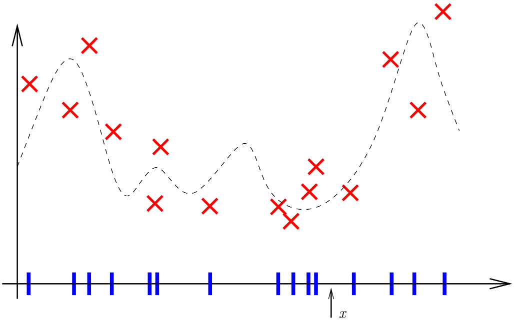
Question: where should one sample the next point?
Answer: select the point with highest upper bound!
"Optimism in the face of (partial observation) uncertainty"

Lipschitz optimization with noisy evaluations

f is still Lipschitz, but now, the evaluation of f at x_t returns a noisy evaluation r_t of $f(x_t)$, i.e. such that $\mathbb{E}[r_t|x_t] = f(x_t)$.

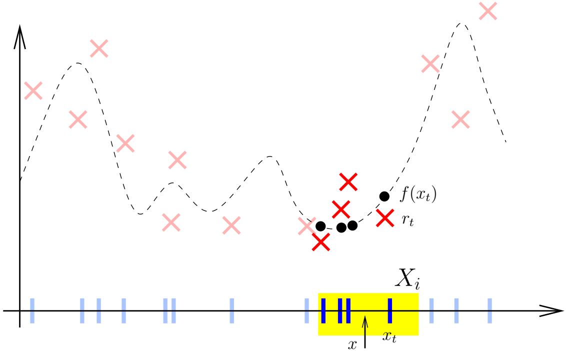


Where should one sample next?



How to define a high probability upper bound at any state x?

UCB in a given domain

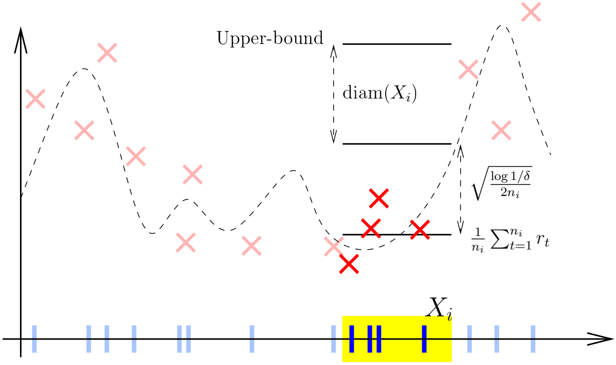


For a fixed domain $X_i \ni x$ containing n_i points $\{x_t\} \in X_i$, we have that $\sum_{t=1}^{n_i} r_t - f(x_t)$ is a Martingale. Thus by Azuma's inequality,

$$\frac{1}{n_i}\sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} \geq \frac{1}{n_i}\sum_{t=1}^{n_i} f(x_t) \geq f(x) - diam(X_i),$$

since f is Lipschitz (where $diam(X_i) = \sup_{x,y\in X_i} \ell(x,y)$).

High probability upper bound



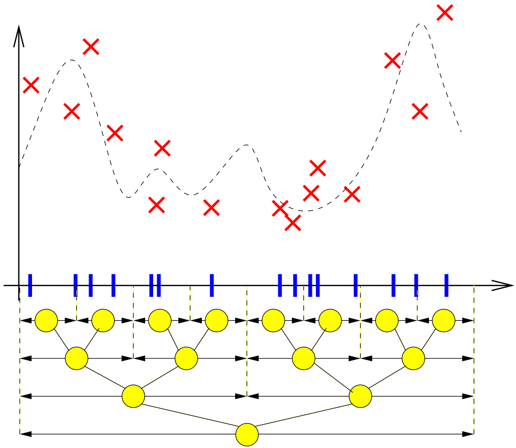
$$\text{w.p. } 1-\delta, \qquad \frac{1}{n_i}\sum_{t=1}^{n_i} r_t + \sqrt{\frac{\log 1/\delta}{2n_i}} + diam(X_i) \geq \sup_{x\in X_i} f(x).$$

Tradeoff between size of the confidence interval and diameter.

By considering several domains we can derive a tigher upper bound.

A hierarchical decomposition

Use a tree of partitions at all scales:

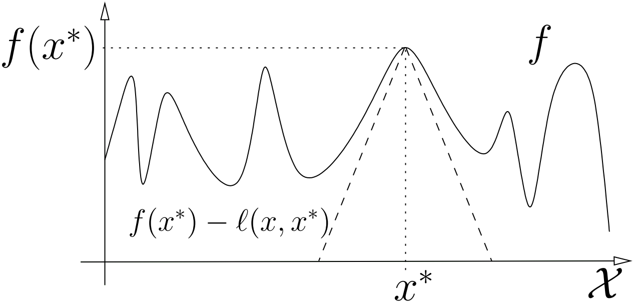


$$B_i(t) \stackrel{\mathrm{def}}{=} \min\left\{ \widehat{\mu}_i(t) + \sqrt{\frac{2\log(t)}{T_i(t)}} + \mathrm{diam}(i), \max_{j \in \mathcal{C}(i)} B_j(t) \right\}$$

$\mathcal{X}$-armed bandits

Let $\mathcal{X}$ be a space equipped with a semi-metric $\ell(x, y)$.
Let $f(x)$ be a function such that:

$$f(x^*) - f(x) \leq \ell(x, x^*),$$



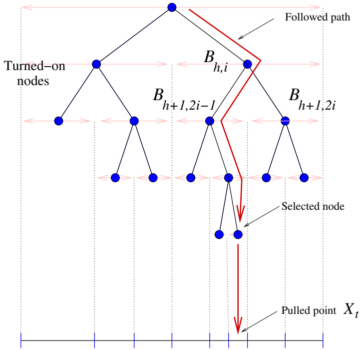
Hierarchical Optimistic Optimization (HOO)

[Bubeck, Munos, Stoltz, Szepesvári, 2008]: Consider a tree of partitions of $\mathcal{X}$, each node i corresponds to a subdomain X_i.

HOO Algorithm:

Let $\mathcal{T}_t$ be the set of expanded nodes at round t.

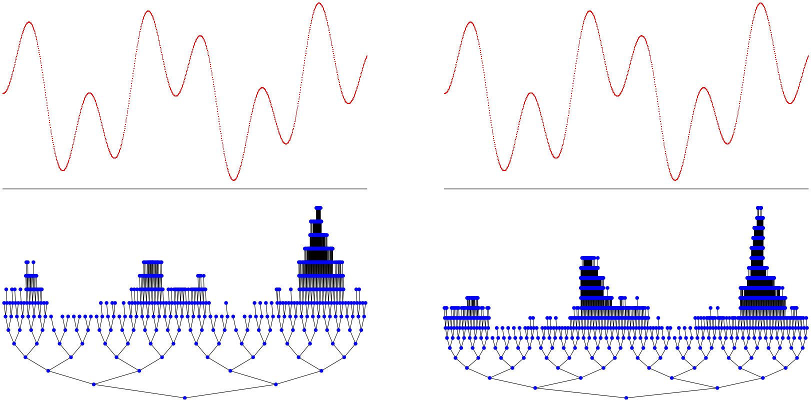
- $\mathcal{T}_1 = \{\text{root}\}$ (space $\mathcal{X}$)
- At t, select a leaf I_t of $\mathcal{T}_t$ by maximizing the B-values,
- $\mathcal{T}_{t+1} = \mathcal{T}_t \cup \{I_t\}$
- Select $x_t \in X_{I_t}$
- Observe reward r_t and update the B-values:



$$B_i(t) \stackrel{\text{def}}{=} \min\left[\widehat{\mu}_i(t) + \sqrt{\frac{2\log(t)}{T_i(t)}} + diam(i), \max_{j\in\mathcal{C}(i)} B_j(t)\right]$$

Example in 1d

$r_t \sim \mathcal{B}(f(x_t))$ a Bernoulli distribution with parameter $f(x_t)$



Resulting tree at time $n = 1000$ and at $n = 10000$.

Analysis of HOO

Let d be the **near-optimality dimension** of f in $\mathcal{X}$: i.e. such that the set of ε-optimal states

$$X_\varepsilon \stackrel{\text{def}}{=} \{x \in \mathcal{X}, f(x) \geq f^* - \varepsilon\}$$

can be covered by $O(\varepsilon^{-d})$ balls of radius ε.
Then

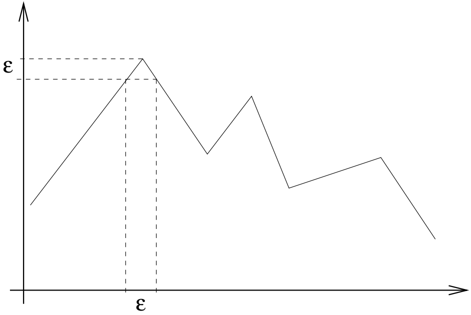
$$\mathbb{E}R_n = \widetilde{O}(n^{\frac{d+1}{d+2}}).$$

(Similar to Zooming algorithm of [Kleinberg, Slivkins, Upfall, 2008], but HOO requires a tree of partitions whereas Zooming requires a sampling oracle)

Example 1:

Assume the function is locally peaky around its maximum:

$$f(x^*) - f(x) = \Theta(||x^* - x||).$$

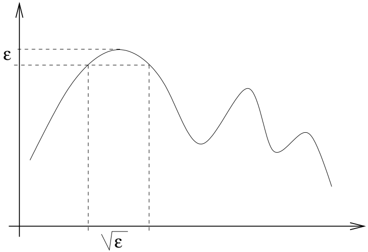


It takes $O(\varepsilon^0)$ balls of radius ε to cover X_ε. Thus $d = 0$ and the regret is $\widetilde{O}(\sqrt{n})$.

Example 2:

Assume the function is locally quadratic around its maximum:

$$f(x^*) - f(x) = \Theta(||x^* - x||^\alpha), \text{ with } \alpha = 2.$$



- For $\ell(x, y) = ||x - y||$, it takes $O(\varepsilon^{-D/2})$ balls of radius ε to cover X_ε. Thus $d = D/2$ and $R_n = \widetilde{O}(n^{\frac{D+2}{D+4}})$.
- For $\ell(x, y) = ||x - y||^2$, it takes $O(\varepsilon^0)$ ℓ-balls of radius ε to cover X_ε. Thus $d = 0$ and $R_n = \widetilde{O}(\sqrt{n})$.

Known smoothness around the maximum

Consider $\mathcal{X} = [0,1]^d$. Assume that f has a finite number of global maxima and is locally α-smooth around each maximum x^*, i.e.

$$f(x^*) - f(x) = \Theta(||x^* - x||^\alpha).$$

Then, by choosing $\ell(x,y) = ||x-y||^\alpha$, X_ε is covered by $O(1)$ ℓ-balls of "radius" ε. Thus the near-optimality dimension $d = 0$, and the regret of HOO is:

$$\mathbb{E}R_n = \widetilde{O}(\sqrt{n}),$$

The rate of growth is **independent of the ambient dimension** D.

Conclusions on $\mathcal{X}$-armed bandits

The near-optimality dimension may be seen as an excess order of smoothness of f (around its maxima) compared to what is known:

- **If the smoothness order of the function is known** then the regret of HOO is $\widetilde{O}(\sqrt{n})$
- **If the smoothness is underestimated**, for example f is α-smooth but we only use $\ell(x, y) = ||x - y||^\beta$, with $\beta < \alpha$, then the near-optimality dimension is $d = D(1/\beta - 1/\alpha)$ and the regret is $\widetilde{O}(n^{(d+1)/(d+2)})$
- **If the smoothness is overestimated**, the local-Lipschitz assumption is violated, thus there is no guarantee.
  For example UCT [Kocsis, Szepesvári, 2006] can be arbitrarily poor [Coquelin, Munos, 2007].

Bandits in trees

Outline of this Section:

- A more structured problem: finding a path in a tree
- An algorithm that does not fully use the reward structure
- An algorithm that does!

References: [Kocsis, Szepesvári, 2006], [Coquelin, Munos, 2007], [Bubeck, Munos, 2010]

A more structured problem

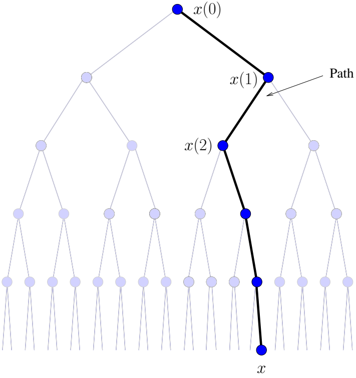
Finding an path in a tree: each arm is a path (in a graph or tree) and the *value of the path is the sum of rewards along the path*.

Example:

- Infinite horizon with γ-discounted rewards. K actions.
- Space of arms $\mathcal{X}$ = set of paths (infinite sequence of actions).
- Reward along a path x_t:

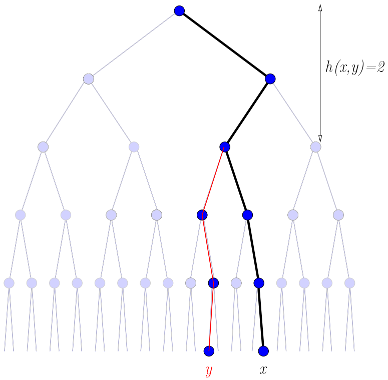
$$y_t = \sum_{i \geq 0} \gamma^i y_t(i),$$

  where $y_t(i) \sim \nu(x_t(i)) \in [0,1]$.
- Write $\mu(x(i)) = \mathbb{E}[\nu(x(i))]$, and $f(x) = \sum_{i \geq 0} \gamma^i \mu(x(i))$



Using HOO

- **Prop:** The mean-reward function $f(x) = \sum_{i\geq 1} \gamma^i \mu(x(i))$ is Lipschitz w.r.t. the metric: $\ell(x,y) = \frac{\gamma^{h(x,y)}}{1-\gamma}$.
- **Use HOO**: At round t, play the path x_t maximizing the B-value.
- Observe sample reward $y_t = \sum_{i\geq 1} \gamma^i y_t(i)$ of the path and use it to update the B-values.



Problem: HOO does not make full use of the tree structure: It uses the sample reward y_t of a path x_t but not the sample rewards $y_t(i)$ of all nodes $x_t(i)$ of the path x_t.

Optimistic sampling using the tree structure

OLOP algorithm [Bubeck, Munos, 2010]:

- At round t, play path x_t (up to depth $h = \frac{1}{2}\frac{\log n}{\log 1/\gamma}$)
- Observe sample rewards $y_t(i)$ of each node along the path x_t
- Compute empirical rewards for each node $x(i)$ of depth $i \leq h$

$$\widehat{\mu}_t(x(i)) = \frac{1}{T_{x(i)}(t)} \sum_{s=1}^{t} y_t(i)\mathbb{I}\{x(i) \in x_t\} \text{ where } T_{x(i)}(t) = \sum_{s=1}^{t} \mathbb{I}\{x(i) \in x_t\}$$

- Define bound for each path x:

$$B_t(x) = \min_{1\leq j\leq h} \Big[\sum_{i=1}^{j} \gamma^i \Big(\widehat{\mu}(x(i)) + \sqrt{\frac{2\log n}{T_{x(i)}(t)}}\Big) + \frac{\gamma^{j+1}}{1-\gamma}\Big]$$

- Select path $x_{t+1} = \mathrm{argmax}_x B_t(x)$

This algorithm fully uses the tree structure of the rewards.

Performance guarantee of OLOP

Consider the near-optimality dimension d, i.e., such that

$$\mathcal{X}_\varepsilon = \{x \in \mathcal{X}, f(x) \geq f^* - \varepsilon\}$$

is covered by $O(\varepsilon^{-d})$ ℓ-balls of size ε.
Regret of OLOP: (Open Loop Optimistic Planning) after n calls to the generative model,

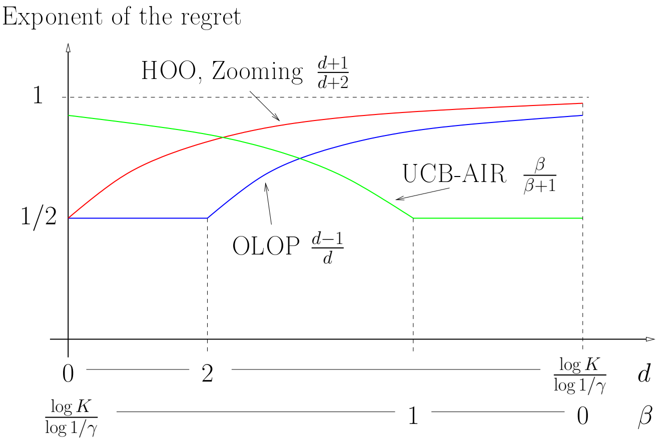
$$R_n = nf^* - \mathbb{E}\Big[\sum_{t=1}^{n} f(x_t)\Big] = \begin{cases} \widetilde{O}\Big(n^{\frac{d-1}{d}}\Big) & \text{if } d > 2 \\ \widetilde{O}(\sqrt{n}) & \text{if } d \leq 2 \end{cases}$$

Another measure of the set of near-optimal paths is $\beta \geq 0$:

$$\mathbb{P}(\text{Random path is } \varepsilon\text{-optimal}) = O(\varepsilon^\beta).$$

Note that we have $d = \frac{\log K}{\log 1/\gamma} - \beta \in [0, \frac{\log K}{\log 1/\gamma}]$.

Comparison: OLOP, HOO, Zooming, UCB-AIR



Conclusion on stochastic bandits

- Success of "Optimism in the face of uncertainty" principle
- Use reward structure as much as possible
- Better concentration inequalities $\Longrightarrow$ better bounds
- Regret bounds expressed in terms of a measure of near-optimal solutions

Other topics in stochastic bandits

A few pointers:

- **Contextual bandits** [Woodroofe, 1979], [Auer, 2002], [Wang, Kulkarni, Poor, 2005], [Pandey, Agarwal, Chakrabarti, Josifovski, 2007], [Langford, Zhang, 2007], [Hazan, Megiddo, 2007], [Rigollet, Zeevi, 2010], [Chu, Li, Reyzin, Schapire, 2011], [Slivkins, 2011].
- **Restless bandits** [Whittle, 1988], [Bertsimas, Nino-Mora, 1994], [Guha, Munagala, Shi, 2007], [Filippi, 2010].
- **Markov decision processes** [Burnetas, Katehakis, 1997], [Auer, Ortner, 2007], [Jaksch, Ortner, Auer, 2010], [Bartlett, Tewari, 2009].
- **Gaussian bandits** [Dorard, Glowacka, Shawe-Taylor, 2009], [Grünewälder, Audibert, Opper, Shawe-Taylor, 2010], [Srinivas, Krause, Kakade, Seeger, 2010]
- **Sleeping bandits** [Kleinberg, Niculescu-mizil, Sharma, 2008], [Kanade, McMahan, Bryan, 2009], **mortal bandits** [Chakrabarti, Kumar, Radlinski, Upfal, 2008], ...

Topics in adversarial bandits

At each round t,

- Simultaneously, the adversary selects a function $f_t : \mathcal{X} \mapsto \mathbb{R}$, and the player chooses $x_t \in \mathcal{X}$
- The reward $f_t(x_t)$ is revealed.

The performance of the player is compared to the best constant strategy:

$$R_n = \max_{x \in \mathcal{X}} \sum_{t=1}^{n} f_t(x) - \sum_{t=1}^{n} f_t(x_t).$$

Performance depends on

- Full versus bandit information
- Class of functions f_t
- Shape of the action space $\mathcal{X}$

[Cesa-Bianchi, Lugosi, 2006]

A few pointers

- **Linear bandits** [Dani, Hayes, Kakade, 2008], [Abernethy, Hazan, Rakhlin, 2008], [Awerbuch, Kleinberg, 2008],
- **Convex bandits** [Zinkevich, 2003], [Flaxman, Kalai, McMahan, 2005], [Hazan, Agarwal, Kale, 2006], [Bartlett, Hazan, Rakhlin, 2007], [Shalev-Shwartz, 2007], [Abernethy, Bartlett, Rakhlin, Tewari, 2008], [Narayanan, Rakhlin, 2010]
- **Lipschitz bandits** [Maillard, Munos, 2010]
- **Countable bandits** [Poland, 2008]
- **Combinatorial bandits** [Cesa-Bianchi, Lugosi, 2009], [Audibert, Bubeck, Lugosi, 2011]
- **Online learning** in stochastic/adversarial environments [Auer, Cesa-Bianchi, Freund, Schapire, 2002], [Kakade, Kalai, 2005], [Cesa-Bianchi et al. 2009], [Abernethy, Agarwal, Bartlett, Rakhlin, 2009], [Ben-David, Pal, Shalev-Shwartz, 2009], [Lazaric, Munos, 2009], [Rakhlin, Sridharan, Tewari, 2011].

Thank you

Material available on the Tutorial web page:

`https://sites.google.com/site/banditstutorial`

Bibliography I

[AB09] Jean-Yves Audibert and Sébastien Bubeck.
Minimax policies for adversarial and stochastic bandits.
In Dasgupta and Klivans [DK09].

[AB10] Jean-Yves Audibert and Sébastien Bubeck.
Regret bounds and minimax policies under partial monitoring.
Journal of Machine Learning Research, 11:2785–2836, December 2010.

[ABL11] Jean-Yves Audibert, Sébastien Bubeck, and Gabor Lugosi.
Minimax policies for combinatorial prediction games.
In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

[ABRT08] Jacob D. Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari.
Optimal strategies and minimax lower bounds for online convex games.
In Servedio and Zhang [SZ08].

[ACBF02] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer.
Finite time analysis of the multiarmed bandit problem.
Machine Learning, 47(2-3):235–256, 2002.

[ACBFS95] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire.
Gambling in a rigged casino: The adversarial multi-armed bandit problem.
In *Proceedings of the 36th annual symposium on Foundations of Computer Science*, FOCS '95, pages 322–331, Milwaukee, WI, USA, 1995. IEEE Computer Society Press.

[ACBFS03] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire.
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32:48–77, January 2003.

[Agr95] R. Agrawal.
The continuum-armed bandit problem.
SIAM Journal on Control and Optimization, 33:1926–1951, 1995.

Bibliography II

[AHR08] Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin.
Competing in the dark: An efficient algorithm for bandit linear optimization.
In Servedio and Zhang [SZ08], pages 263–274.

[AK08] Baruch Awerbuch and Robert D. Kleinberg.
Online linear optimization and adaptive routing.
Journal of Computer Systems and Science, 74:97–114, February 2008.

[AM05] Peter Auer and Ron Meir, editors.
volume 3559 of *COLT '05, Lecture Notes in Computer Science*, Bertinoro, Italy, jun 2005. Springer.

[AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári.
Exploration-exploitation trade-off using variance estimates in multi-armed bandits.
Theoretical Computer Science, 410:1876–1902, 2009.

[AO06] Peter Auer and Ronald Ortner.
Logarithmic online regret bounds for undiscounted reinforcement learning.
In Schölkopf et al. [SPH06], pages 49–56.

[AOS07] P. Auer, R. Ortner, and C. Szepesvári.
Improved rates for the stochastic continuum-armed bandit problem.
In *Proceedings of the 20th Conference on Learning Theory*, pages 454–468, 2007.

[Aue02] Peter Auer.
Using confidence bounds for exploitation-exploration trade-offs.
Journal of Machine Learning Research, 3:397–422, March 2002.

[BCZ+97] Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp.
Bandit problems with infinitely many arms.
Annals of Statistics, (25):2103–2116, 1997.

Bibliography III

[BDH+08] Peter L. Bartlett, Varsha Dani, Thomas P. Hayes, Sham M. Kakade, Alexander Rakhlin, and Ambuj Tewari.
High-probability regret bounds for bandit online linear optimization.
In Servedio and Zhang [SZ08], pages 335–342.

[BDPSS09] S. Ben-David, D. Pal, and S. Shalev-Shwartz.
Agnostic online learning.
In *22th annual conference on learning theory*, 2009.

[BG07] Nader H. Bshouty and Claudio Gentile, editors.
volume 4539 of *COLT '07, Lecture Notes in Computer Science*, San Diego, CA, USA, jun 2007. Springer.

[BHR07] Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin.
Adaptive online gradient descent.
In Platt et al. [PKSR07], pages 65–72.

[BK96] Apostolos N. Burnetas and Michaël N. Katehakis.
Optimal adaptive policies for sequential allocation problems.
Advances in Applied Mathematics, 17(2):122–142, 1996.

[BK97] Apostolos N. Burnetas and Michaël N. Katehakis.
Optimal adaptive policies for markov decision processes.
Mathematics of Operations Research, 22:222–255, February 1997.

[BLS06] José L. Balcázar, Philip M. Long, and Frank Stephan, editors.
volume 4264 of *ALT '06, Lecture Notes in Computer Science*, Barcelona, Spain, oct 2006. Springer.

[BM07] Avrim Blum and Yishay Mansour.
From external to internal regret.
Journal of Machine Learning Research, 8:1307–1324, December 2007.

Bibliography IV

[BM10] S. Bubeck and R. Munos.
Open loop optimistic planning.
In *Conference on Learning Theory*, 2010.

[BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz.
Pure exploration in multi-armed bandits problems.
In Gavaldà et al. [GLZZ09], pages 23–37.

[BMSS08] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári.
Online optimization of X-armed bandits.
In Koller et al. [KSBB08].

[Bro04] Carla E. Brodley, editor.
volume 69 of *ICML '04, ACM International Conference Proceeding Series*, Banff, Alberta, Canada, jul 2004. ACM.

[BSL+09] Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Chris K. I. Williams, and Aron Culotta, editors.
NIPS '09, Vancouver, British Columbia, Canada, dec 2009.

[BT09] Peter L. Bartlett and Ambuj Tewari.
Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps.
In *Proceedings of the 25th conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 35–42, Arlington, Virginia, United States, 2009. AUAI Press.

[BTO02] Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors.
NIPS '02, Vancouver, British Columbia, Canada, dec 2002. MIT Press.

[Bub10a] Sébastien Bubeck.
Bandits Games and Clustering Foundations.
PhD thesis, Université de Lille 1, 2010.

[Bub10b] Sébastien Bubeck.
Jeux de bandits et fondations du clustering.
PhD thesis, Université des Sciences et des Technologies de Lille 1, 2010.

Bibliography V

[CBL06] Nicolò Cesa-Bianchi and Gábor Lugosi.
Prediction, Learning, and Games.
Cambridge University Press, New York, NY, USA, 2006.

[CBL09] Nicolò Cesa-Bianchi and Gábor Lugosi.
Combinatorial bandits.
In Dasgupta and Klivans [DK09].

[CKRU08] Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal.
Mortal multi-armed bandits.
In Koller et al. [KSBB08], pages 273–280.

[CM06] William W. Cohen and Andrew Moore, editors.
volume 148 of *ICML '06, ACM International Conference Proceeding Series*, Pittsburgh, Pennsylvania, USA, jun 2006. ACM.

[CM07] P.-A. Coquelin and R. Munos.
Bandit algorithms for tree search.
In *Uncertainty in Artificial Intelligence*, 2007.

[CMR08] William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors.
volume 307 of *ICML '08, ACM International Conference Proceeding Series*, Helsinki, Finland, jun 2008. ACM.

[cWKP05] Chih chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor.
Bandit problems with side observations.
IEEE Transactions on Automatic Control, 50:338–355, 2005.

[DBL09] Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors.
volume 382 of *ICML '09, ACM International Conference Proceeding Series*, Montreal, Quebec, Canada, jun 2009. ACM.

Bibliography VI

[DHK08a] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade.
The price of bandit information for online optimization.
In Koller et al. [KSBB08], pages 345–352.

[DHK08b] Varsha Dani, Thomas P. Hayes, and Sham M. Kakade.
Stochastic linear optimization under bandit feedback.
In Servedio and Zhang [SZ08], pages 355–366.

[DK09] Sanjot Dasgupta and Adam Klivans, editors.
COLT '09, Montreal,Quebec, Canada, jun 2009.

[FCGS10] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvari.
Parametric bandits: The generalized linear case.
In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 586–594. 2010.

[FGTZ08] Yoav Freund, László Györfi, György Turán, and Thomas Zeugmann, editors.
volume 5254 of *ALT '08, Lecture Notes in Computer Science*, Budapest, Hungary, oct 2008. Springer.

[Fil10] Sarah Filippi.
Stratégies optimistes en apprentissage par renforcement.
PhD thesis, Télécom ParisTech, 2010.

[FJ10] Johannes Fürnkranz and Thorsten Joachims, editors.
ICML '10, Haifa, Israel, jun 2010. Omnipress.

[FKB05] Abraham D. Flaxman, Adam Tauman Kalai, and Hugh Brendan McMahan.
Online convex optimization in the bandit setting: gradient descent without a gradient.
In *Proceedings of the 16th annual ACM-SIAM Symposium On Discrete Algorithms*, SODA '05, pages 385–394. SIAM, 2005.

Bibliography VII

[GAOST10] Steffen Grünewälder, Jean-Yves Audibert, Manfred Opper, and John Shawe-Taylor.
Regret bounds for gaussian process bandit problems.
In *AISTATS'10*, 2010.

[GC11] Aurélien Garivier and Olivier Cappé.
The KL-UCB algorithm for bounded stochastic bandits and beyond.
In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

[Gha07] Zoubin Ghahramani, editor.
volume 227 of *ICML '07, ACM International Conference Proceeding Series*, Corvalis, Oregon, USA, jun 2007. ACM.

[GLZZ09] Ricard Gavaldà, Gábor Lugosi, Thomas Zeugmann, and Sandra Zilles, editors.
volume 5809 of *ALT '09, Lecture Notes in Computer Science*, Porto, Portugal, oct 2009. Springer.

[GMS07] Sudipto Guha, Kamesh Munagala, and Peng Shi.
Approximation algorithms for restless bandit problems.
CoRR, abs/0711.3861, 2007.

[GWG89] John C. Gittins, Richard Weber, and Kevin Glazebrook.
Multi-armed Bandit Allocation Indices.
Wiley, 1989.

[HAK06] Elad Hazan, Amit Agarwal, and Satyen Kale.
Logarithmic regret algorithms for online convex optimization.
In Lugosi and Simon [LS06], pages 499–513.

[HK08] Elad Hazan and Satyen Kale.
Extracting certainty from uncertainty: Regret bounded by variation in costs.
In Servedio and Zhang [SZ08], pages 57–68.

[HST07] Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors.
volume 4754 of *ALT '07, Lecture Notes in Computer Science*, Sendai, Japan, oct 2007. Springer.

Bibliography VIII

[HT10a] Junya Honda and Akimichi Takemura.
An asymptotically optimal bandit algorithm for bounded support models.
In Kalai and Mohri [KM10], pages 67–79.

[HT10b] Junya Honda and Akimichi Takemura.
An asymptotically optimal policy for finite support models in the multiarmed bandit problem.
arXiv:0905.2776, 2010.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer.
Near-optimal regret bounds for reinforcement learning.
Journal of Machine Learning Research, 99:1563–1600, August 2010.

[Kle04] Robert D. Kleinberg.
Nearly tight bounds for the continuum-armed bandit problem.
In *Proceedings of the 18th conference on advances in Neural Information Processing Systems* [NIP04].

[KM10] Adam Tauman Kalai and Mehryar Mohri, editors.
Omnipress, June 2010.

[KMB09] Varun Kanade, H. Brendan McMahan, and Brent Bryan.
Sleeping experts and bandits with stochastic action availability and adversarial rewards.
In *Proceedings of the 12th international conference on Artificial Intelligence and Statistics*, number 5 in AI&Stats '09, pages 272–279, 2009.

[KNMS08] Robert D. Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma.
Regret bounds for sleeping experts and bandits.
In Servedio and Zhang [SZ08], pages 425–436.

[KRS10] Satyen Kale, Lev Reyzin, and Robert E. Schapire.
Non-stochastic bandit slate problems.
In Lafferty et al. [LWST$^+$10], pages 1054–1062.

Bibliography IX

[KS06] L. Kocsis and Cs. Szepesvari.
Bandit based Monte-carlo planning.
In *Proceedings of the 15th European Conference on Machine Learning*, pages 282–293, 2006.

[KSBB08] Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors.
NIPS '08, Vancouver, British Columbia, Canada, dec 2008. MIT Press.

[KSST08] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari.
Efficient bandit algorithms for online multiclass prediction.
In Cohen et al. [CMR08], pages 440–447.

[KSU08] Robert D. Kleinberg, Alexander Slivkins, and Eli Upfal.
Multi-armed bandit problems in metric spaces.
In *Proceedings of the 40th ACM symposium on Theory Of Computing*, TOC '08, pages 681–690, 2008.

[LM09] Alessandro Lazaric and Rémi Munos.
Hybrid stochastic-adversarial online learning.
In Dasgupta and Klivans [DK09].

[LPP10] Tyler Lu, David Pál, and Martin Pál.
Contextual multi-armed bandits.
In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the 13th international conference on Artificial Intelligence and Statistics*, volume 9, pages 485–492, 2010.

[LR85] Tze Leung Lai and Herbert Robbins.
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6:4–22, 1985.

[LR03] Ehud Lehrer and Dinah Rosenberg.
A wide range no-regret theorem.
Game theory and information, EconWPA, 2003.

Bibliography X

[LS06] Gábor Lugosi and Hans-Ulrich Simon, editors.
volume 4005 of *COLT '06, Lecture Notes in Computer Science*, Pittsburgh, PA, USA, jun 2006. Springer.

[LW89] Nick Littlestone and Manfred K. Warmuth.
The weighted majority algorithm.
In *Proceedings of the 30th annual Symposium on Foundations of Computer Science*, pages 256–261, Washington, DC, USA, 1989. IEEE Computer Society.

[LWST+10] John D. Lafferty, Chris K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors.
NIPS '10, Vancouver, British Columbia, Canada, dec 2010.

[LZ08] J. Langford and T. Zhang.
The epoch-greedy algorithm for multi-armed bandits with side information.
In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. MIT Press, Cambridge, MA, 2008.

[Mai11] Odalric Ambrym Maillard.
Apprentissage séquentiel: Bandits, Statistique et Renforcement.
PhD thesis, Université des Sciences et des Technologies de Lille 1, 2011.

[MM10] Odalric-Ambrym Maillard and Rémi Munos.
Online learning in adversarial lipschitz environments.
In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD'10, pages 305–320, Berlin, Heidelberg, 2010. Springer-Verlag.

[MM11] Odalric-Ambrym Maillard and Rémi Munos.
Adaptive bandits: Towards the best history-dependent strategy.
In *To appear in Proceedings of the 14th international conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR W&CP*, 2011.

Bibliography XI

[MMS11] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz.
Finite-time analysis of multi-armed bandits problems with kullback-leibler divergences.
In *To appear in Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

[MT04] S. Mannor and J. N. Tsitsiklis.
The sample complexity of exploration in the multi-armed bandit problem.
Journal of Machine Learning Research, 5:623–648, 2004.

[NIP04] NIPS '04, Vancouver, British Columbia, Canada, dec 2004. MIT Press.

[NIP05] NIPS '05, Vancouver, British Columbia, Canada, dec 2005. MIT Press.

[NR10] Hariharan Narayanan and Alexander Rakhlin.
Random walk approach to regret minimization.
In Lafferty et al. [LWST+10], pages 1777–1785.

[Ort09] Ronald Ortner.
Online regret bounds for markov decision processes with deterministic transitions.
In Gavaldà et al. [GLZZ09], pages 123–137.

[PACJ07] S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski.
Bandits for taxonomies: A model-based approach.
In *Proceedings of the Seventh SIAM International Conference on Data Mining*, 2007.

[PCA07a] S. Pandey, D. Chakrabarti, and D. Agarwal.
Multi-armed bandit problems with dependent arms.
In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 721–728, New York, NY, USA, 2007. ACM.

[PCA07b] Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal.
Multi-armed bandit problems with dependent arms.
In Ghahramani [Gha07].

Bibliography XII

[PKSR07] John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors.
NIPS '07, Vancouver, British Columbia, Canada, dec 2007. MIT Press.

[Pol08] Jan Poland.
Nonstochastic bandits: Countable decision set, unbounded costs and reactive environments.
Theoretical Computuer Science, 397(1-3):77–93, jul 2008.

[Rob52] Herbert Robbins.
Some aspects of the sequential design of experiments.
Bulletin of the American Mathematics Society, 58:527–535, 1952.

[RST10] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari.
Online learning: Beyond regret.
ArXiv e-prints, nov 2010.

[RT10] Paat Rusmevichientong and John N. Tsitsiklis.
Linearly parameterized bandits.
Math. Oper. Res., 35:395–411, May 2010.

[RW05] Luc De Raedt and Stefan Wrobel, editors.
volume 119 of *ICML '05, ACM International Conference Proceeding Series*, Bonn, Germany, aug 2005. ACM.

[SKKS10] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger.
Gaussian process optimization in the bandit setting: No regret and experimental design.
In *ICML'10*, pages 1015–1022, 2010.

[Sli11] Aleksandrs Slivkins.
Contextual bandits with similarity information.
In *Proceedings of the 24th annual Conference On Learning Theory*, COLT '11, 2011.

[SPH06] Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors.
NIPS '06, Vancouver, British Columbia, Canada, dec 2006. MIT Press.

Bibliography XIII

[SS07] Shai Shalev-Shwartz.
Online Learning: Theory, Algorithms, and Applications.
PhD thesis, The Hebrew University of Jerusalem, July 2007.

[Sto05] Gilles Stoltz.
Incomplete Information and Internal Regret in Prediction of Individual Sequences.
Phd thesis, Université Paris-Sud, Orsay, France, May 2005.

[Sto11] Gilles Stoltz.
Contributions to the sequential prediction of arbitrary sequences: applications to the theory of repeated games and empirical studies of the performance of the aggregation of experts.
Habilitation à diriger des recherches, Université Paris-Sud, 2011.

[STS04] John Shawe-Taylor and Yoram Singer, editors.
volume 3120 of *COLT '04, Lecture Notes in Computer Science*, Banff, Canada, jul 2004. Springer.

[SU08] A. Slivkins and E. Upfal.
Adapting to a changing environment: the brownian restless bandits.
In Rocco A. Servedio and Tong Zhang, editors, *COLT*, pages 343–354. Omnipress, 2008.

[SZ08] Rocco A. Servedio and Tong Zhang, editors.
volume 80 of *COLT '08*, Helsinki, Finland, jul 2008. Omnipress.

[Tho33] William R. Thompson.
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25:285–294, 1933.

[TSS03] Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors.
NIPS '03, Vancouver, British Columbia, Canada, dec 2003. MIT Press.

[WAM08] Yizao Wang, Jean-Yves Audibert, and Rémi Munos.
Algorithms for infinitely many-armed bandits.
In Koller et al. [KSBB08], pages 1729–1736.

Bibliography XIV

[Whi80] Peter Whittle.
Multi-armed bandits and the gittins index.
Journal of the Royal Statistical Society. Series B (Methodological), 42(2):143–149, 1980.

[YBKJ09] Y. Yue, J. Broder, R. Kleinberg, and T. Joachims.
The k-armed dueling bandits problem.
In *22th annual conference on learning theory*, 2009.

[Zin03] Martin Zinkevich.
Online convex programming and generalized infinitesimal gradient ascent.
In *Proceedings of the 20th International Conference on Machine Learning*, ICML '03, pages 928–936, 2003.