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

Jean-Yves Audibert\textsuperscript{1,2} and Rémi Munos\textsuperscript{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 than 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.
Many-armed bandits  Linear bandits  \(\mathcal{X}\)-armed bandits  Bandits in trees  Other related topics

A few references on bandits since 2005...


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 digging 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

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

$\beta$ characterizes the probability of selecting near-optimal arms

Large $\beta \iff$ 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

$$\arg\max_{1 \leq k \leq K(t)} \left[ \hat{\mu}_{k,t} + \sqrt{\frac{2 \hat{V}_{k,t} E_t}{T_k(t)} + \frac{3 E_t}{T_k(t)}} \right]$$

with empirical rewards and confidence interval

with exploration sequence: $c \log(\log t) \leq E_t \leq \log t$. 

$K(t)$ played arms
Arms not played yet
Regret analysis of UCB-AIR

Upper bound on the regret of UCB-AIR:

\[ \mathbb{E} R_n = \begin{cases} 
\tilde{O}(\sqrt{n}) & \text{if } \beta < 1 \text{ and } \mu^* < 1 \\
\tilde{O}(n^{1+\beta}) & \text{if } \mu^* = 1 \text{ or } \beta \geq 1 
\end{cases} \]

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 $\beta$ is unknown? (i.e. adaptive strategy that estimates $\beta$ while minimizing regret).
Structured set of actions or rewards

The mapping $\text{Arms} \rightarrow \text{Reward}$ possesses some known structure:

- Linear
- Lipschitz
- Tree structure

Reward samples from observed arms provides information about unseen arms.
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 $x^* = \arg\max_{x \in \mathcal{X}} \langle x, \alpha^* \rangle$ 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 $\alpha^*$ is enough. So the regret will scale with $D$ and not with the number of arms (which may be infinite).
Choose $x_t \in \mathcal{X}$ and get:

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

(provides information about $\alpha^*$ 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 $\alpha \in E_t$. 

**Geometric intuition**
Confidence Ball algorithms

[Dani, Hayes, Kakade, 2008]

**UCB idea:** Define a least-squares estimate $\hat{\alpha}_t$ of $\alpha^*$:

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

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

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

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

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

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} \max_{\alpha \in E_t} \langle x, \alpha \rangle.$$
Regret analysis

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

**Upper bounds:**

- *Problem independent:* With probability $1 - \delta$,
  \[
  R_n = O(D\sqrt{n}(\log n/\delta)^{3/2})
  \]

- *Problem dependent:* With probability $1 - \delta$,
  \[
  R_n = O\left(\frac{D^2}{\Delta} (\log n/\delta)^3\right),
  \]
  where $\Delta$ 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 \( \ell_1 \)-ellipsoid instead of \( \ell_2 \), 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 \( \mu \) is a real-valued function (such as logistic regression function, in order to deal with binary rewards). GLM-UCB selects the arm:

  \[
  \argmax_{x \in \mathcal{X}} \left( \mu(\langle x, \hat{\alpha}_t, \rangle) + \rho(t)\|x\|_{2,A_t^{-1}} \right),
  \]

  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:** \( \alpha^* \) 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 \( f \).
Outline of this Section:

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

**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

\[ f(x_t) \]

Lipschitz property → 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 $x$?
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,

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

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

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

Tradeoff between size of the confidence interval and diameter.

By considering several domains we can derive a tighter upper bound.
A hierarchical decomposition

Use a tree of partitions at all scales:

\[
B_i(t) \overset{\text{def}}{=} \min \left\{ \hat{\mu}_i(t) + \sqrt{\frac{2 \log(t)}{T_i(t)}} + \text{diam}(i), \max_{j \in C(i)} B_j(t) \right\}
\]
Let $\mathcal{X}$ be a space equipped with a semi-metric $\ell(x, y)$. Let $f(x)$ be a function such that:

$$f(x^*) - f(x) \leq \ell(x, x^*),$$
[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 $l_t$ of $\mathcal{T}_t$ by maximizing the B-values,
- $\mathcal{T}_{t+1} = \mathcal{T}_t \cup \{l_t\}$
- Select $x_t \in X_{l_t}$
- Observe reward $r_t$ and update the B-values:

$$B_i(t) \overset{\text{def}}{=} \min_{\hat{\mu}_i(t)} \left[ \hat{\mu}_i(t) + \sqrt{\frac{2 \log(t)}{T_i(t)}} + \text{diam}(i), \max_{j \in \mathcal{C}(i)} B_j(t) \right]$$
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 $\varepsilon$-optimal states

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

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

Then

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

(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 $\varepsilon$ to cover $X_\varepsilon$. Thus $d = 0$ and the regret is $\tilde{O}(\sqrt{n})$. 
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 \( \varepsilon \) to cover \( X_\varepsilon \). Thus \( d = D/2 \) and \( R_n = \tilde{O}(n^{D+2\over D+4}) \).
- For \( \ell(x, y) = ||x - y||^2 \), it takes \( O(\varepsilon^0) \) \( \ell \)-balls of radius \( \varepsilon \) to cover \( X_\varepsilon \). Thus \( d = 0 \) and \( R_n = \tilde{O}(\sqrt{n}) \).
Known smoothness around the maximum

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

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

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

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

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 $\tilde{O}(\sqrt{n})$

- **If the smoothness is underestimated**, for example $f$ is $\alpha$-smooth but we only use $\ell(x, y) = ||x - y||^\beta$, with $\beta < \alpha$, then the near-optimality dimension is $d = D\left(\frac{1}{\beta} - \frac{1}{\alpha}\right)$ and the regret is $\tilde{O}(n^{(d+1)/(d+2)})$

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

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

- Define bound for each path $x$:

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

- Select path $x_{t+1} = \arg\max_x B_t(x)$

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 $O(\varepsilon^{-d}) \ell$-balls of size $\varepsilon$.

**Regret of OLOP:** (Open Loop Optimistic Planning) after $n$ calls to the generative model,

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

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

Exponent of the regret

HOO, Zooming $\frac{d+1}{d+2}$

UCB-AIR $\frac{\beta}{\beta+1}$

OLOP $\frac{d-1}{d}$

$\log K$ $\frac{\log K}{\log 1/\gamma}$

$\frac{\log K}{\log 1/\gamma}$ $1$ $\log 1/\gamma$ $d$ $0$ $
\beta$
Conclusion on stochastic bandits

- Success of “Optimism in the face of uncertainty” principle
- Use reward structure as much as possible
- Better concentration inequalities $\implies$ 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].


- **Markov decision processes** [Burnetas, Katehakis, 1997], [Auer, Ortner, 2007], [Jaksch, Ortner, Auer, 2010], [Bartlett, Tewari, 2009].


- **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 $f_t : \mathcal{X} \rightarrow \mathbb{R}$, 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]

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].

Regret bounds and minimax policies under partial monitoring.

Minimax policies for combinatorial prediction games.

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.

[ACBFS95] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire.
Gambling in a rigged casino: The adversarial multi-armed bandit problem.

[ACBFS03] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire.
The nonstochastic multiarmed bandit problem.

The continuum-armed bandit problem.
Bibliography II

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.

[AM05] Peter Auer and Ron Meir, editors.

[AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári.
Exploration-exploitation trade-off using variance estimates in multi-armed bandits.

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

Improved rates for the stochastic continuum-armed bandit problem.

[Aue02] Peter Auer.
Using confidence bounds for exploitation-exploration trade-offs.

Bandit problems with infinitely many arms.
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.

Agnostic online learning.
In 22th annual conference on learning theory, 2009.

[BG07] Nader H. Bshouty and Claudio Gentile, editors.
Springer.

Adaptive online gradient descent.
In Platt et al. [PKSR07], pages 65–72.

Optimal adaptive policies for sequential allocation problems.

Optimal adaptive policies for markov decision processes.

[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.
Many-armed bandits  Linear bandits  $\alpha$-armed bandits  Bandits in trees  Other related topics

Bibliography IV


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


Many-armed bandits  Linear bandits  $\lambda$-armed bandits  Bandits in trees  Other related topics

Bibliography V


Many-armed bandits  Linear bandits  $\mathcal{X}$-armed bandits  Bandits in trees  Other related topics

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.


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


Many-armed bandits  Linear bandits  \(\mathcal{X}\)-armed bandits  Bandits in trees  Other related topics

Bibliography VII


Bibliography VIII


Bibliography IX

[KS06]  L. Kocsis and Cs. Szepesvari.
Bandit based Monte-carlo planning.


[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.

Multi-armed bandit problems in metric spaces.

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

Contextual multi-armed bandits.

Asymptotically efficient adaptive allocation rules.

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

      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,

[LWST+10] John D. Lafferty, Chris K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta,
          editors.

      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

      Apprentissage séquentiel: Bandits, Statistique et Renforcement.

      Online learning in adversarial lipschitz environments.
      In Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in

      Adaptive bandits: Towards the best history-dependent strategy.
      In To appear in Proceedings of the 14th international conference on Artificial Intelligence and
Bibliography XI

Finite-time analysis of multi-armed bandits problems with kullback-leibler divergences.

The sample complexity of exploration in the multi-armed bandit problem.


Random walk approach to regret minimization.
In Lafferty et al. [LWST+10], pages 1777–1785.

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

Bandits for taxonomies: A model-based approach.

Multi-armed bandit problems with dependent arms.

[PCA07b] Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal.
Multi-armed bandit problems with dependent arms.
In Ghahramani [Gha07].
Many-armed bandits  Linear bandits  $\chi$-armed bandits  Bandits in trees  Other related topics

Bibliography XII


[Pol08] Jan Poland.  
Nonstochastic bandits: Countable decision set, unbounded costs and reactive environments.  

[Rob52] Herbert Robbins.  
Some aspects of the sequential design of experiments.  

Online learning: Beyond regret.  

[RT10] Paat Rusmevichientong and John N. Tsitsiklis.  
Linearly parameterized bandits.  

[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.  

Contextual bandits with similarity information.  

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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>[SZ08]</td>
<td>Rocco A. Servedio and Tong Zhang, editors.</td>
<td>volume 80 of COLT '08, Helsinki, Finland, jul 2008. Omnipress.</td>
</tr>
</tbody>
</table>
Bibliography XIV

