Bandits for Recommendation:

Theoretical Contributions with Applications in Mind

Aurélien Garivier

May 27^{th}, 2016

Institut de Mathématiques de Toulouse
LabeX CIMI
Université Paul Sabatier

Overview

1. Best-Arm Identification: the true complexity, and how to reach it

   joint work with Emilie Kaufmann, accepted at COLT'16
2. Why should we use sequential methods?

   joint work with Emilie Kaufmann and Tor Lattimore, submitted to NIPS'16
3. Regret minimization: what the Lai&Robbins lower bound does not say

   joint work with Pierre Mnard and Gilles Stoltz, submitted
4. (Bandit and Games: optimizing short tree exploration)

   joint work with Emilie Kaufmann and Wouter Koolen, accepted at COLT'16
5. (Fading bandits: already presented by J. Loudec)

   joint work with J. Loudec, L. Rossi, M. Chevallier and J. Mothe, accepted at CAP'16

Best-Arm Identification: the True Complexity, and How to Reach it

Best-Arm Identification: the True Complexity, and How to Reach it

Goal : identify the best arm, a^*, as fast/accurately as possible.

$\Rightarrow$ **optimal exploration**

The agent's strategy is made of:

- a sequential sampling strategy (A_t)
- a stopping rule τ (stopping time)
- a recommendation rule $\hat{a}_\tau$

Possible goals:

Fixed-budget setting	Fixed-confidence setting
$\tau = T$	minimize $\mathbb{E}[\tau]$
minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	$\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$

Motivation: Market research, A/B Testing, clinical trials...

A New Lower Bound

Theorem

For any δ-PAC algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^*(\boldsymbol{\mu}) \log\left(\frac{1}{2.4\delta}\right),$$

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)\right).$$

Moreover, the vector

$$w^*(\boldsymbol{\mu}) = \operatorname*{argmax}_{w \in \Sigma_K} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)\right)$$

contains the optimal proportions of arm draws.

Sampling Rule: Tracking the Optimal Proportions

$\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t), \dots, \hat{\mu}_K(t))$: vector of empirical means

- Introducing

$$U_t = \{a : N_a(t) < \sqrt{t}\},$$

the arm sampled at round $t+1$ is

$$A_{t+1} \in \begin{cases} \underset{a \in U_t}{\operatorname{argmin}}\ N_a(t) \text{ if } U_t \neq \emptyset & (\textit{forced exploration}) \\ \underset{1 \leq a \leq K}{\operatorname{argmax}} \left[t\ w_a^*(\hat{\boldsymbol{\mu}}(t)) - N_a(t)\right] & (\textit{tracking}) \end{cases}$$

Lemma

Under the Tracking sampling rule,

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\lim_{t\to\infty} \frac{N_a(t)}{t} = w_a^*(\boldsymbol{\mu})\right) = 1.$$

Chernoff's Stopping Rule: SGLRT

High values of the Generalized Likelihood Ratio

$$Z_{a,b}(t) := \log \frac{\max_{\{\boldsymbol{\lambda}:\lambda_a \geq \lambda_b\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})}{\max_{\{\boldsymbol{\lambda}:\lambda_a \leq \lambda_b\}} \ell(X_1, \dots, X_t; \boldsymbol{\lambda})},$$

reject the hypothesis that $(\mu_a < \mu_b)$.

We stop when one arm is assessed to be significantly larger than all other arms, according to a SGLR Test:

$$\begin{aligned} \tau_\delta &= \inf\{t \in \mathbb{N} : \exists a \in \{1, \dots, K\}, \forall b \neq a, Z_{a,b}(t) > \beta(t, \delta)\} \\ &= \inf\left\{t \in \mathbb{N} : \max_{a \in \{1,\dots,K\}} \min_{b \neq a} Z_{a,b}(t) > \beta(t, \delta)\right\} \end{aligned}$$

An asymptotically optimal algorithm

Theorem

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- the Chernoff stopping rule with $\beta(t,\delta) = \log\left(\frac{2(K-1)t}{\delta}\right)$
- and recommends $\hat{a}_\tau = \underset{a=1\dots K}{\mathrm{argmax}}\ \hat{\mu}_a(\tau)$

is δ-PAC for every $\delta \in]0,1[$ and satisfies

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T^*(\boldsymbol{\mu}).$$

Numerical experiments

Experiments on two Bernoulli bandit models:

- $\boldsymbol{\mu}_1 = [0.5\ 0.45\ 0.43\ 0.4]$, such that

$$w^*(\boldsymbol{\mu}_1) = [0.417\ 0.390\ 0.136\ 0.057]$$

- $\boldsymbol{\mu}_2 = [0.3\ 0.21\ 0.2\ 0.19\ 0.18]$, such that

$$w^*(\boldsymbol{\mu}_2) = [0.336\ 0.251\ 0.177\ 0.132\ 0.104]$$

In practice, set the threshold to $\beta(t,\delta) = \log\left(\frac{\log(t)+1}{\delta}\right)$.

	Track-and-Stop	Chernoff-Racing	KL-LUCB	KL-Racing
$\boldsymbol{\mu}_1$	4052	4516	8437	9590
$\boldsymbol{\mu}_2$	1406	3078	2716	3334

Table 1: Expected number of draws $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ for $\delta = 0.1$, averaged over $N = 3000$ experiments.

Why should we use sequential methods?

Why should we use sequential methods?

- Two Gaussian arms with variance 1
- Gap Δ known or unkown
- We know how to find the best arm "optimally"
- Can we perform exploration at the beginning?
- Are Explore-Then-Commit strategies optimal?

Fixed-Budget ETC: Algorithm

input: T and Δ
$n := \left\lceil 2W\big(T^2\Delta^4/(32\pi)\big)/\Delta^2 \right\rceil$
for $k \in \{1, \ldots, n\}$ **do**
    choose $A_{2k-1} = 1$ and $A_{2k} = 2$
end for
$\hat{a} := \operatorname{argmax}_i \hat{\mu}_{i,n}$
for $t \in \{2n+1, \ldots, T\}$ **do**
    choose $A_t = \hat{a}$
end for

Algorithm 1: FB-ETC algorithm

Fixed-Budget ETC: Regret Bound

Theorem

Let $\mu \in \mathcal{H}_\Delta$*, and let* $\overline{n} = \left\lceil \frac{2}{\Delta^2} W\left(\frac{T^2\Delta^4}{32\pi}\right)\right\rceil$*. Then*

$$R_\mu^{\overline{n}}(T) \le \frac{4}{\Delta}\log\left(\frac{T\Delta^2}{4.46}\right) - \frac{2}{\Delta}\log\log\left(\frac{T\Delta^2}{4\sqrt{2\pi}}\right) + \Delta$$

whenever $T\Delta^2 > 4\sqrt{2\pi e}$*, and* $R_\mu^{\overline{n}}(T) \le T\Delta/2 + \Delta$ *otherwise. In all cases,* $R_\mu^{\overline{n}}(T) \le 2.04\sqrt{T} + \Delta$*. Furthermore, for all* $\epsilon > 0$*,* $T \ge 1$ *and* $n \le 4(1-\epsilon)\log(T)/\Delta^2$*,*

$$R_\mu^n(T) \ge \left(1 - \frac{2}{n\Delta^2}\right)\left(1 - \frac{8\log(T)}{\Delta^2 T}\right)\frac{\Delta T^\epsilon}{2\sqrt{\pi\log(T)}}\,.$$

As $R_\mu^n(T) \ge n\Delta$*, this entails that* $\inf_{1\le n\le T} R_\mu^n(T) \sim 4\log(T)/\Delta$.

ETC, Known Gap: Algorithm

input: T and Δ
$A_1 = 1, A_2 = 2,\ s := 2$
while $(s/2)\Delta\,|\hat{\mu}_1(s) - \hat{\mu}_2(s)| < \log\left(T\Delta^2\right)$ **do**
    choose $A_{s+1} = 1$ and $A_{s+2} = 2$
    $s := s + 2$
end while
$\hat{a} := \operatorname{argmax}_i \hat{\mu}_i(s)$
for $t \in \{s+1, \ldots, T\}$ **do**
    choose $A_t = \hat{a}$
end for

Algorithm 2: SPRT ETC algorithm

ETC, Known Gap: Regret Bound

Theorem

If $T\Delta^2 \geq 1$, then the regret of the SPRT-ETC algorithm is upper-bounded as

$$R_\mu^{\text{SPRT-ETC}}(T) \leq \frac{\log(eT\Delta^2)}{\Delta} + \frac{4\sqrt{\log(T\Delta^2)} + 4}{\Delta} + \Delta\,.$$

Otherwise it is upper bounded by $T\Delta/2 + \Delta$, and for all T and Δ the regret is less than $10\sqrt{T/e} + \Delta$.

General Strategy, Known Gap: Algorithm

input: T and Δ
$\epsilon_T = \Delta \log^{-\frac{1}{8}}(e + T\Delta^2)/4$
for $t \in \{1, \dots, T\}$ **do**
    let $A_{t,\min} := \arg\min_{i \in 1,2} N_i(t-1)$ and $A_{t,\max} = 3 - A_{t,\min}$
    if $\hat{\mu}_{A_{t,\min}}(t-1) + \sqrt{\dfrac{2\log\left(\frac{T}{N_{A_{t,\min}}(t-1)}\right)}{N_{A_{t,\min}}(t-1)}} \geq \hat{\mu}_{A_{t,\max}}(t-1) + \Delta - 2\epsilon_T$ **then**
        choose $A_t = A_{t,\min}$
    else
        choose $A_t = A_{t,\max}$
    end if
end for

Algorithm 3: Δ-UCB

General Strategy, Known Gap: Regret Bound

Theorem

If $T(2\Delta - 3\epsilon_T)^2 \geq 2$ and $T\epsilon_T^2 \geq e^2$, the regret of the Δ-UCB algorithm is upper bounded as

$$R_\mu^{\Delta\text{-}UCB}(T) \leq \frac{\log\left(2T\Delta^2\right)}{2\Delta(1-3\epsilon_T/(2\Delta))^2} + \frac{\sqrt{\pi \log\left(2T\Delta^2\right)}}{2\Delta(1-3\epsilon_T/\Delta)^2} + \Delta\left[\frac{30e\sqrt{\log(\epsilon_T^2 T)}}{\epsilon_T^2} + \frac{80}{\epsilon_T^2} + \frac{2}{(2\Delta - 3\epsilon_T)^2}\right] + 5\Delta.$$

Moreover $\limsup_{T\to\infty} R_\mu^{\Delta\text{-}UCB}(T)/\log(T) \leq (2\Delta)^{-1}$ *and* $\forall \mu \in \mathcal{H}_\Delta,\ R_\mu^{\Delta\text{-}UCB}(T) \leq 328\sqrt{T} + 5\Delta.$

ETC, Unkown Gap: Algorithm

input: $T(\geq 3)$
$A_1 = 1, A_2 = 2,\ s := 2$
while $|\hat{\mu}_1(s) - \hat{\mu}_2(s)| < \sqrt{\frac{8\log(T/s)}{s}}$ **do**
  choose $A_{s+1} = 1$ and $A_{s+2} = 2$
  $s := s + 2$
end while
$\hat{a} := \mathrm{argmax}_i\, \hat{\mu}_i(s)$
for $t \in \{s+1, \dots, T\}$ **do**
  choose $A_t = \hat{a}$
end for

Algorithm 4: BAI-ETC algorithm

ETC, Unkown Gap: Algorithm: Regret Bound

Theorem

If $T\Delta^2 > 4e^2$, the regret of the BAI-ETC algorithm is upper bounded as

$$R_\mu^{BAI\text{-}ETC}(T) \leq \frac{4\log\left(\frac{T\Delta^2}{4}\right)}{\Delta} + \frac{334\sqrt{\log\left(\frac{T\Delta^2}{4}\right)}}{\Delta} + \frac{178}{\Delta} + \Delta.$$

It is upper bounded by $T\Delta$ otherwise, and by $32\sqrt{T} + \Delta$ in any case.

General Strategy, Unkown Gap: Algorithm

1: **input:** T
2: **for** $t \in \{1, \ldots, T\}$ **do**
3: $\quad A_t = \operatorname*{argmax}_{i \in \{1,2\}} \hat{\mu}_i(t-1) + \sqrt{\frac{2}{N_i(t-1)} \log\left(\frac{T}{N_i(t-1)}\right)}$
4: **end for**

Algorithm 5: UCB*

General Strategy, Unkown Gap: Regret Bound

Theorem

For all $\epsilon \in (0, \Delta)$, if $T(\Delta - \epsilon)^2 \geq 2$ and $T\epsilon^2 \geq e^2$, the regret of the UCB strategy is upper bounded as*

$$R_\mu^{UCB^*}(T) \leq \frac{2\log\left(\frac{T\Delta^2}{2}\right)}{\Delta\left(1 - \frac{\epsilon}{\Delta}\right)^2} + \frac{2\sqrt{\pi\log\left(\frac{T\Delta^2}{2}\right)}}{\Delta\left(1 - \frac{\epsilon}{\Delta}\right)^2} + \Delta\left(\frac{30e\sqrt{\log(\epsilon^2 T)} + 16e}{\epsilon^2}\right) + \frac{2}{\Delta\left(1 - \frac{\epsilon}{\Delta}\right)^2} + \Delta.$$

Moreover, $\limsup_{T\to\infty} R_\mu^\pi(T)/\log(T) = 2/\Delta$ *and for all* $\mu \in \mathcal{H}$, $R_\mu^\pi(T) \leq 33\sqrt{T} + \Delta$.

In a Nutshell

All those results come with a matching asymptotic lower bound

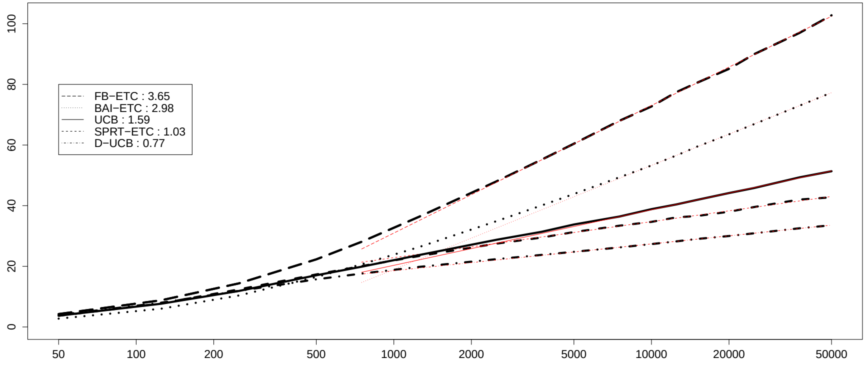
	Π_{ALL}	Π_{ETC}	Π_{DETC}
$\mathcal{H}$	2	4	NA
$\mathcal{H}_\Delta$	1/2	1	4

$\implies$ fully sequential methods are much better!

($\implies$ Lai&Robbins bound is not a lower bound)

Regret Minimization: What the Lai&Robbins Lower Bound Does Not Say

A Simple Experiment



Regret of the five strategies for a bandit problem with $\Delta = 1/5$ and different values of the horizon (4.10^5 Monte-Carlo replications). In the legend, the estimated slopes of $\Delta R^{\pi}(T)$ (in logarithmic scale) are indicated after the policy names.

Regret Minimization: What the Lai&Robbins Lower Bound Does Not Say

- New lower bound: For every $\mathcal{F}_T$ measurable rv in $[0,1]$,

$$\sum_{a=1}^{K} \mathbb{E}_\mu\big[N_a(T)\big]\mathrm{kl}(\mu_a, \mu'_a) \geq \mathrm{kl}\big(\mathbb{E}_\mu[Z], \mathbb{E}_{\mu'}[Z]\big)$$

- $\rightarrow$ non-asymptotic Lai&Robbins
- $\rightarrow$ short-horizon lower bounds
- In mind: multiple action bandits, combinatorial bandits: the $\log(T)/\Delta$ bound is not relevant!

Non-asymptotic Lai&Robbins

Theorem

For all super-consistent strategies ψ on well-behaved models $\mathcal{D}$, for all bandit problems ν in $\mathcal{D}$, for all suboptimal arms a,

$$\mathbb{E}_\nu\big[N_a(T)\big] \geq \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} - (a_T + b_T + c_T)\ln T - \frac{\ln 2}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,, \qquad (1)$$

for all $T \geq 2$ large enough so that

$$a_T = \frac{\omega(\nu_a, \mu^\star)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}(\ln T)^{-4}\,, \qquad b_T = C_{\psi,\mathcal{D}} H(\nu)\frac{\ln T}{T}\,, \qquad c_T = \frac{\ln\big(K\, C_{\psi,\mathcal{D}}(\ln T)^9\big)}{\ln T}\,,$$

are all smaller than 1.

Short-Horizon Lower Bound 1

Theorem

For all strategies ψ that are smarter than the uniform strategy, for all bandit problems ν, for all arms a, for all $T \geq 1$,

$$\mathbb{E}_\nu\big[N_a(T)\big] \geq \frac{T}{K}\Big(1 - \sqrt{2T\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\Big)\,.$$

In particular,

$$\forall\, T \leq \frac{1}{8\mathcal{K}_{\inf}(\nu_a, \mu^\star)}, \qquad \mathbb{E}_\nu\big[N_a(T)\big] \geq \frac{T}{2K}\,.$$

Short-Horizon Lower Bound 2

Theorem

For all strategies ψ *that are pairwise symmetric for optimal arms, for all bandit problems* ν*, for all suboptimal arms* a *and all optimal arms* $a^\star$*, for all* $T \geq 1$*,*

$$\text{either} \quad \mathbb{E}_\nu\big[N_a(T)\big] \geq \frac{T}{K}$$

or

$$\mathbb{E}_\nu\left[\frac{\max\{N_a(T),\,1\}}{\max\{N_{a^\star}(T),\,1\}}\right] \geq 1 - 2\sqrt{\frac{2T\,\mathrm{KL}(\nu_a,\nu_{a^\star})}{K}}\,.$$

Short-Horizon Lower Bound 3

Theorem

For all strategies ψ that are pairwise symmetric for optimal arms and monotonic, for all bandit problems ν,

$$\sum_{a \notin \mathcal{A}^\star(\nu)} \mathbb{E}_\nu\big[N_a(T)\big] \geq T\left(1 - \frac{A^\star_\nu}{K} - \frac{A^\star_\nu \sqrt{2T\,\mathcal{K}^{\max}_\nu}}{K} - \frac{2A^\star_\nu\, T \mathcal{K}^{\max}_\nu}{K}\right),$$

$$\text{where} \qquad \mathcal{K}^{\max}_\nu = \min_{w \in \mathcal{W}(\nu)} \max_{a^\star \in \mathcal{A}^\star(\nu)} \mathrm{KL}(\nu_w, \nu_{a^\star}).$$

In particular, the regret is lower bounded according to

$$R_{\nu,T} \geq \left(\min_{a \notin \mathcal{A}^\star(\nu)} \Delta_a\right) T\left(1 - \frac{A^\star_\nu}{K} - \frac{A^\star_\nu \sqrt{2T\,\mathcal{K}^{\max}_\nu}}{K} - \frac{2A^\star_\nu\, T \mathcal{K}^{\max}_\nu}{K}\right).$$