Revisiting the Exploration-Exploitation Tradeoff in Bandit Models

Emilie Kaufmann

joint work with Aurélien Garivier (IMT, Toulouse)
and Tor Lattimore (University of Alberta)

Workshop on Optimization and Decision-Making in Uncertainty,
Simons Institute, Berkeley, September 21st, 2016

The multi-armed bandit model

K arms $=$ K probability distributions (ν_a has mean μ_a)

ν_1 ν_2 ν_3 ν_4 ν_5

At round t, an agent:

- chooses an arm A_t
- observes a sample $X_t \sim \nu_{A_t}$

using a sequential sampling strategy (A_t):

$$A_{t+1} = F_t(A_1, X_1, \dots, A_t, X_t).$$

Generic goal: learn the best arm, $a^* = \text{argmax}_a \, \mu_a$ of mean $\mu^* = \max_a \mu_a$

Regret minimization in a bandit model

Samples = **rewards**, (A_t) is adjusted to

- maximize the (expected) sum of rewards,

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

- or equivalently minimize the *regret*:

$$R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \sum_{a=1}^{K}(\mu^* - \mu_a)\mathbb{E}[N_a(T)]$$

$N_a(T)$: number of draws of arm a up to time T

$\Rightarrow$ **Exploration/Exploitation tradeoff**

Algorithms: naive ideas

- **Idea 1 :** Choose each arm T/K times

$\Rightarrow$ EXPLORATION

- **Idea 2 :** Always choose the best arm so far

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

...Linear regret

Algorithms: naive ideas

- **Idea 1 :** Choose each arm T/K times

$\Rightarrow$ EXPLORATION

- **Idea 2 :** Always choose the best arm so far

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

...Linear regret

- **A better idea:**
  First explore the arms uniformly,
  then commit to the empirical best until the end

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

...Still sub-optimal

A motivation: should we minimize regret?

$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

For the t-th patient in a clinical study,

- chooses a treatment A_t
- observes a response $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Goal: maximize the number of patient healed during the study

A motivation: should we minimize regret?

$\mathcal{B}(\mu_1)$

$\mathcal{B}(\mu_2)$

$\mathcal{B}(\mu_3)$

$\mathcal{B}(\mu_4)$

$\mathcal{B}(\mu_5)$

For the t-th patient in a clinical study,

- chooses a treatment A_t
- observes a response $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Goal: maximize the number of patient healed during the study

Alternative goal: allocate the treatments so as to identify as quickly as possible the best treatment
(no focus on curing patients during the study)

Two different objectives

	Regret minimization	Best arm identification
Bandit algorithm	sampling rule (A_t)	sampling rule (A_t) stopping rule τ recommendation rule $\hat{a}_\tau$
Input	horizon T	risk parameter δ
Objective	minimize $R_T = \mu^* T - \mathbb{E}\left[\sum_{t=1}^T X_t\right]$	ensure $\mathbb{P}(\hat{a}_\tau = a^*) \geq 1 - \delta$ and minimize $\mathbb{E}[\tau]$
	Exploration/Exploitation	**pure Exploration**

This talk:

- ➜ (distribution-dependent) optimal algorithm for both objectives
- ➜ best performance of an Explore-Then-Comit strategy?

We focus on distributions parameterized by their means

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$$

(Bernoulli, Gaussian)

Outline

Optimal algorithms for regret minimization

$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$. $N_a(t)$: number of draws of arm a up to time t

$$\mathrm{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = \sum_{a=1}^{K} (\mu^* - \mu_a) \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]$$

Notation: Kullback-Leibler divergence

$$d(\mu, \mu') := \mathrm{KL}\left(\nu_\mu, \nu_{\mu'}\right)$$

$$\text{(Gaussian): } d(\mu, \mu') = \frac{(\mu - \mu')^2}{2\sigma^2}$$

- [Lai and Robbins, 1985]: for uniformly efficient algorithms,

$$\mu_a < \mu^* \Rightarrow \liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \geq \frac{1}{d(\mu_a, \mu^*)}$$

A bandit algorithm is **asymptotically optimal** if, for every $\boldsymbol{\mu}$,

$$\mu_a < \mu^* \Rightarrow \limsup_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}$$

Optimal algorithms for regret minimization

$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$. $N_a(t)$: number of draws of arm a up to time t

$$\mathrm{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = \sum_{a=1}^{K} (\mu^* - \mu_a) \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]$$

Notation: Kullback-Leibler divergence

$$d(\mu, \mu') := \mathrm{KL}\left(\nu_\mu, \nu_{\mu'}\right)$$

$$\text{(Bernoulli): } d(\mu, \mu') = \mu \log \frac{\mu}{\mu'} + (1-\mu) \log \frac{1-\mu}{1-\mu'}$$

- [Lai and Robbins, 1985]: for uniformly efficient algorithms,

$$\mu_a < \mu^* \Rightarrow \liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \geq \frac{1}{d(\mu_a, \mu^*)}$$

A bandit algorithm is **asymptotically optimal** if, for every $\boldsymbol{\mu}$,

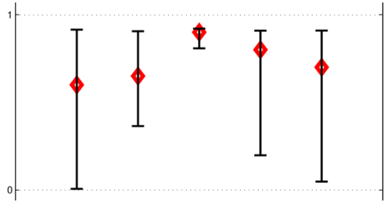
$$\mu_a < \mu^* \Rightarrow \limsup_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}$$

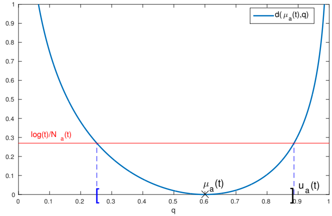
Mixing Exploration and Exploitation: the UCB approach

- A UCB-type (or *optimistic*) algorithm chooses at round t

$$A_{t+1} = \underset{a=1\dots K}{\text{argmax}} \ \textcolor{red}{\mathrm{UCB}_a(t)}.$$

where $\mathrm{UCB}_a(t)$ is an Upper Confidence Bound on μ_a.





The KL-UCB index

$$\mathrm{UCB}_a(t) := \max\left\{ q : \textcolor{orange}{d}\left(\hat{\mu}_a(t), q\right) \leq \frac{\log(t)}{N_a(t)} \right\},$$

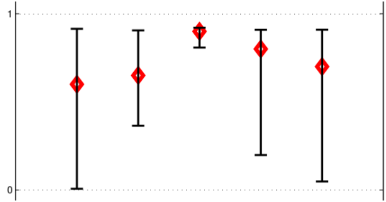
satisfies $\mathbb{P}(\mu_a \leq \mathrm{UCB}_a(t)) \gtrsim 1 - t^{-1}$.

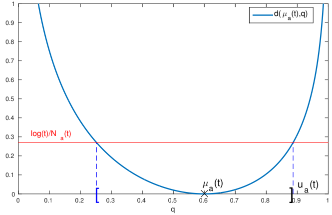
Mixing Exploration and Exploitation: KL-UCB

- A UCB-type (or *optimistic*) algorithm chooses at round t

$$A_{t+1} = \underset{a=1...K}{\text{argmax}}\ \mathrm{UCB}_a(t).$$

where $\mathrm{UCB}_a(t)$ is an Upper Confidence Bound on μ_a.





The KL-UCB index [Cappé et al. 13]: KL-UCB satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1}{d(\mu_a, \mu^*)} \log T + O(\sqrt{\log(T)}).$$

Outline

1. Optimal algorithms for Regret Minimization
2. Optimal algorithms for Best Arm Identification
3. Explore-Then-Commit strategies

A sample complexity lower bound

A Best Arm Identification algorithm $(A_t, \tau, \hat{a}_\tau)$ is δ-PAC if

$$\forall \boldsymbol{\mu}, \ \mathbb{P}_{\boldsymbol{\mu}}(\hat{a}_\tau = a^*(\boldsymbol{\mu})) \geq 1 - \delta.$$

Theorem [Garivier and K. 2016]

For any δ-PAC algorithm,

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

where

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

$\Sigma_K = \{w \in [0,1]^K : \sum_{i=1}^K w_i = 1\}$, $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$

Moreover, the vector of optimal proportions, $\left(\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \simeq w_a^*(\boldsymbol{\mu})\right)$

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

is well-defined, and we propose an efficient way to compute it.

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

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

Lemma

Under the Tracking sampling rule,

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

An asymptotically optimal algorithm

Theorem [K. and Garivier, 2016]

The Track-and-Stop strategy, that uses

- the Tracking sampling rule
- a stopping rule based on GLRT tests:
  $\tau_\delta = \inf\left\{t \in \mathbb{N} : Z(t) > \log \frac{2Kt}{\delta}\right\}$, with
  $$Z(t) := t \times \sup_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\mu}(t))} \left[\sum_{a=1}^{K} \frac{N_a(t)}{t} d(\hat{\mu}_a(t), \lambda_a)\right]$$
- and recommends $\hat{a}_\tau = \underset{a=1\dots K}{\operatorname{argmax}}\ \hat{\mu}_a(\tau)$

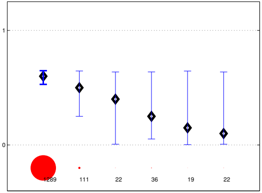
is δ-PAC for every $\delta \in]0,1[$ and satisfies

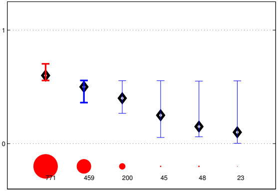
$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T^*(\boldsymbol{\mu}).$$

Regret minimization versus Best Arm Identification

Algorithms for regret minimization and BAI are very different!

- playing mostly the best arm vs. optimal proportions





- different "complexity terms" (featuring KL-divergence)

$$R_T(\boldsymbol{\mu}) \simeq \left(\sum_{a \neq a^*} \frac{\mu^* - \mu_a}{d(\mu_a, \mu^*)} \right) \log(T) \qquad \mathbb{E}_{\boldsymbol{\mu}}[\tau] \simeq T^*(\boldsymbol{\mu}) \log(1/\delta)$$

Outline

Gaussian two-armed bandits

$\nu_1 = \mathcal{N}(\mu_1, 1)$ and $\nu_2 = \mathcal{N}(\mu_2, 1)$. $\boldsymbol{\mu} = (\mu_1, \mu_2)$.
Let $\Delta = |\mu_1 - \mu_2|$

Regret minimization

For any uniformly efficient algorithm $\mathcal{A} = (A_t)$,

$$\liminf_{T\to\infty} \frac{\mathrm{R}_{\boldsymbol{\mu}}(T, \mathcal{A})}{\log(T)} \geq \frac{2}{\Delta}$$

u.e.:
$\forall \boldsymbol{\mu}, \forall \alpha \in]0,1[, R_{\boldsymbol{\mu}}(T, \mathcal{A}) = o(T^\alpha)$

Best Arm Identification

For any δ-PAC algorithm $\mathcal{A} = (A_t, \tau, \hat{a}_\tau)$,

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq \frac{8}{\Delta^2}$$

(optimal algorithms use uniform sampling)

Explore-Then-Commit (ETC) strategies

ETC stragies: given a stopping rule τ and a commit rule $\hat{a}$,

$$A_t = \begin{cases} 1 & \text{if } t \leq \tau \text{ and } t \text{ is odd }, \\ 2 & \text{if } t \leq \tau \text{ and } t \text{ is even }, \\ \hat{a} & \text{otherwise }. \end{cases}$$

Assume $\mu_1 > \mu_2$.

$$\begin{aligned} \mathrm{R}_\mu\left(T, \mathcal{A}^{\mathrm{ETC}}\right) &= \Delta \mathbb{E}_{\boldsymbol{\mu}}[N_2(T)] \\ &= \Delta \mathbb{E}_{\boldsymbol{\mu}}\left[\frac{\tau \wedge T}{2} + (T - \tau)_+ \mathbb{1}_{(\hat{a}=2)}\right] \\ &\leq \frac{\Delta}{2}\mathbb{E}_{\boldsymbol{\mu}}[\tau] + T\Delta\mathbb{P}_{\boldsymbol{\mu}}(\hat{a} = 2). \end{aligned}$$

Explore-Then-Commit (ETC) strategies

ETC stragies: given a stopping rule τ and a commit rule $\hat{a}$,

$$A_t = \begin{cases} 1 & \text{if } t \leq \tau \text{ and } t \text{ is odd} , \\ 2 & \text{if } t \leq \tau \text{ and } t \text{ is even} , \\ \hat{a} & \text{otherwise} . \end{cases}$$

Assume $\mu_1 > \mu_2$.

For $\mathcal{A} = (\tau, \hat{a})$ as in an optimal BAI algorithm with $\delta = \frac{1}{T}$

$$\begin{aligned} \mathrm{R}_\mu\left(T, \mathcal{A}^{\mathrm{ETC}}\right) &= \Delta \mathbb{E}_{\boldsymbol{\mu}}[N_2(T)] \\ &= \Delta \mathbb{E}_{\boldsymbol{\mu}}\left[\frac{\tau \wedge T}{2} + (T - \tau)_+ \mathbb{1}_{(\hat{a}=2)}\right] \\ &\leq \frac{\Delta}{2} \underbrace{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}_{(8/\Delta^2)\log(T)} + T\Delta \underbrace{\mathbb{P}_{\boldsymbol{\mu}}(\hat{a} = 2)}_{1/T} . \end{aligned}$$

Hence

$$\limsup \frac{\mathrm{R}_\mu\left(T, \mathcal{A}\right)}{\log T} \leq \frac{4}{\Delta}.$$

Is this the best we can do? Lower bounds.

Lemma

Let $\boldsymbol{\mu}, \boldsymbol{\lambda} : \ a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$
Let σ s.t. $N_2(T)$ is $\mathcal{F}_\sigma$-measurable. For any u.e. algorithm,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\sigma)]\frac{(\lambda_1-\mu_1)^2}{2} + \mathbb{E}_{\boldsymbol{\mu}}[N_2(\sigma)]\frac{(\lambda_2-\mu_2)^2}{2}}{\log(T)} \geq 1.$$

Proof. Introducing the log-likelihood ratio

$$L_t(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \log \frac{p_{\boldsymbol{\mu}}(X_1, \dots, X_t)}{p_{\boldsymbol{\lambda}}(X_1, \dots, X_t)},$$

one needs to prove that $\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[L_\sigma(\boldsymbol{\mu},\boldsymbol{\lambda})]}{\log(T)} \geq 1$.

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mu}}[L_\sigma(\boldsymbol{\mu}, \boldsymbol{\lambda})] &= \mathrm{KL}\left(\mathcal{L}_{\boldsymbol{\mu}}(X_1, \dots, X_\sigma), \mathcal{L}_{\boldsymbol{\lambda}}(X_1, \dots, X_\sigma)\right) \\ &\geq \mathrm{kl}(\mathbb{E}_{\boldsymbol{\mu}}[Z], \mathbb{E}_{\boldsymbol{\lambda}}[Z]) \ \ \text{for any } Z \in [0,1], \mathcal{F}_\sigma\text{-mesurable} \end{aligned}$$

[Garivier et al. 16]

$$\mathbb{E}_{\boldsymbol{\mu}}[L_\sigma(\boldsymbol{\mu}, \boldsymbol{\lambda})] \geq \mathrm{kl}\left(\mathbb{E}_{\boldsymbol{\mu}}\left[N_2(T)/T\right], \mathbb{E}_{\boldsymbol{\lambda}}\left[N_2(T)/T\right]\right) \sim \log(T) \qquad (u.e.)$$

Is this the best we can do? Lower bounds.

Lemma

Let $\boldsymbol{\mu}, \boldsymbol{\lambda}$: $a^*(\boldsymbol{\mu}) \neq a^*(\boldsymbol{\lambda})$.
Let σ s.t. $N_2(T)$ is $\mathcal{F}_\sigma$-measurable. For any u.e. algorithm,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_1(\sigma)]\frac{(\lambda_1-\mu_1)^2}{2} + \mathbb{E}_{\boldsymbol{\mu}}[N_2(\sigma)]\frac{(\lambda_2-\mu_2)^2}{2}}{\log(T)} \geq 1.$$

Assume $\mu_1 > \mu_2$:

- **Lai and Robbins' bound:**

$\lambda_1 = \mu_1,\ \lambda_2 = \mu_1 + \epsilon$
$\sigma = T$

$$\Rightarrow \quad \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_2(T)]\frac{(\Delta+\epsilon)^2}{2}}{\log(T)} \geq 1.$$

- **For ETC strategies:**

$\lambda_1 = \frac{\mu_1+\mu_2-\epsilon}{2},\ \lambda_2 = \frac{\mu_1+\mu_2+\epsilon}{2}$
$\sigma = \tau \wedge T$

$$\Rightarrow \quad \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau\wedge T]\frac{(\Delta+\epsilon)^2}{8}}{\log(T)} \geq 1$$

$$\Rightarrow \quad \liminf_{T\to\infty} \frac{\mathcal{R}_{\boldsymbol{\mu}}(T,\mathcal{A})}{\log(T)} \geq \frac{4}{\Delta}.$$

An interesting matching algorithm

Theorem

Any uniformly efficient ETC strategy satisfies

$$\liminf_{T\to\infty} \frac{\mathrm{R}_{\boldsymbol{\mu}}(T,\mathcal{A})}{\log(T)} \geq \frac{4}{\Delta}.$$

The ETC strategy based on the stopping rule

$$\tau = \inf\left\{ t = 2n : |\hat{\mu}_{1,n} - \hat{\mu}_{2,n}| > \sqrt{\frac{4\log\left(T/(2n)\right)}{n}} \right\}.$$

satisfies, for $T\Delta^2 > 4e^2$,

$$\begin{aligned} R_\mu(T,\mathcal{A}) &\leq \frac{4\log\left(\frac{T\Delta^2}{4}\right)}{\Delta} + \frac{334\sqrt{\log\left(\frac{T\Delta^2}{4}\right)}}{\Delta} + \frac{178}{\Delta} + \Delta, \\ R_\mu(T,\mathcal{A}) &\leq 32\sqrt{T} + \Delta. \end{aligned}$$

Conclusion

In Gaussian two-armed bandits, ETC strategies are sub-optimal by a factor two compared to UCB strategies

$\Rightarrow$ rather than A/B Test + always showing the best product, dynamically present products to customers all day long!

On-going work:

- how does Optimal BAI + Commit behave in general?

$$T^*(\mu)\left(\sum_{a=2}^{K} w_a^*(\mu)(\mu_1-\mu_a)\right) \quad \text{v.s} \quad \sum_{a=2}^{K}\frac{\mu_1-\mu_a}{d(\mu_a,\mu_1)}.$$

References

- A. Garivier, E. Kaufmann, *Optimal Best Arm Identification with Fixed Confidence*, COLT 2016
- A. Garivier, E. Kaufmann, T. Lattimore, *On Explore-Then-Commit strategies*, NIPS 2016