Two optimization problems in a stochastic bandit model

Emilie Kaufmann

joint work with Olivier Cappé, Aurélien Garivier and Shivaram Kalyanakrishnan

Journées MAS 2014, Toulouse

Outline

Classical framework in stochastic optimization

$$f : \mathcal{X} \longrightarrow \mathbb{R} \qquad \max_{a \in \mathcal{X}} f(a) \ ?$$

Sequential observations: at time t, choose $a_t \in \mathcal{X}$, observe

$$x_t = f(a_t) + \epsilon_t$$

After T observations,

Minimize the optimization error

If $\tilde{a_T}$ is a guess of the argmax

$$\text{minimize} \ \ \mathbb{E}\left[f(\tilde{a_T}) - f(a^*)\right]$$

Minimize the regret

$$\text{minimize} \ \ \mathbb{E}\left[\sum_{t=1}^{T} (f(a^*) - f(a_t))\right]$$

A particular case: the bandit model

$$f : \{1,\dots,K\} \longrightarrow \mathbb{R} \qquad \max_{a=1,\dots,K} f(a) \ ?$$

Sequential observations: at time t, choose $A_t \in \{1,\dots,K\}$, observe $X_t \sim \nu_{A_t}$ where ν_a has mean $f(a)$

After T observations,

Minimize the probability of error

If $\tilde{A_T}$ is a guess of the argmax

$$\text{minimize } \mathbb{P}\left(\tilde{A_T} \neq A^*\right)$$

Minimize the regret

$$\text{minimize } \mathbb{E}\left[\sum_{t=1}^{T}(f(A^*) - f(A_t))\right]$$

Two bandit problems

A binary bandit model is a set of K arms, where

- arm a is a Bernoulli distribution with mean μ_a
- drawing arm a is observing a realization of $\mathcal{B}(\mu_a)$
- arms are assumed to be independent

In a bandit game, at round t, an agent

- chooses arm A_t based on past observations, according to his **sampling strategy**, or **bandit algorithm**
- observes a sample $X_t \sim \mathcal{B}(\mu_{A_t})$

Two possible objectives can be considered

- best arm identification
- regret minimization

Zoom on an application

A doctor can choose between K different treatments

- treatment number a: (unknown) probability of sucess μ_a
- (unknown) best treatment: $a^* = \text{argmax}_a\ \mu_a$
- If treatment a is given to patient t, he is cured with probability μ_a

The doctor:

- chooses treatment A_t to give to patient t
- observes whether the patient is healed : $X_t \sim \mathcal{B}(\mu_{A_t})$

His goal: ajust (A_t) so that to

Regret minimization	Best arm identification
maximize the number of patients healed during a study involving T patients	identify the best treatment with high probability (and always give this one later)

TELECOM ParisTech

Outline

Asymptotically optimal algorithms

$N_a(t)$ be the number of draws of arm a up to time t

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

- [Lai and Robbins,1985]: every consistent algorithm satisfies

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

- A bandit algorithm is **asymptotically optimal** if

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

where

$$d(x, y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y)).$$

A family of optimistic index policies

- For each arm a, compute a confidence interval on μ_a:

$$\mu_a \leq UCB_a(t) \quad w.h.p$$

- Act as if the best possible model was the true model (*optimism-in-face-of-uncertainty*):

$$A_t = \text{argmax}_a \ UCB_a(t)$$

Example UCB1 [Auer et al. 02] uses Hoeffding bounds:

$$UCB_a(t) = \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{2\log(t)}{N_a(t)}}.$$

$S_a(t)$: sum of the rewards collected from arm a up to time t.

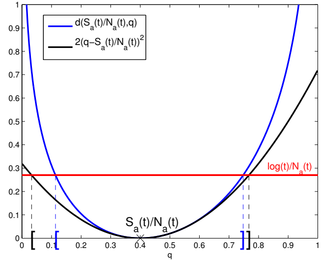
$$\mathbb{E}[N_a(T)] \leq \frac{8}{(\mu^* - \mu_a)^2} \log T + C.$$

KL-UCB: an asymptotically optimal algorithm

- KL-UCB [Cappé et al. 2013] uses the index:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(t) + c \log\log(t)}{N_a(t)} \right\}$$

with $d(p,q) = \text{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right)$.



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

Outline

m best arms identification

Assume $\mu_1 \geq \dots \geq \mu_m > \mu_{m+1} \geq \dots \mu_K$ (Bernoulli bandit model)

Parameters and notations

- m the number of arms to find
- $\delta \in]0, 1[$ a risk parameter
- $\mathcal{S}_m^* = \{1, \dots, m\}$ the set of m optimal arms

The forecaster

- chooses at time t one (or several) arms to draw
- decides to stop after a (possibly random) total number of samples from the arms τ
- recommends a set $\hat{\mathcal{S}}$ of m arms

His goal (in the *fixed-confidence setting*)

- $\mathbb{P}(\hat{\mathcal{S}} = \mathcal{S}_m^*) \geq 1 - \delta$ (the algorithm is δ-PAC)
- The sample complexity $\mathbb{E}[\tau]$ is small

Challenges for m best arm identification

The regret minimization problem is 'solved' in some sense:

- A lower bound on the regret of any good algorithm

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \geq \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{d(\mu_a, \mu_1)}$$

- Algorithms matching this bound, notably KL-UCB

Challenges for m best arm identification

The regret minimization problem is 'solved' in some sense:

- A lower bound on the regret of any good algorithm

$$\liminf_{T\to\infty} \frac{R_T}{\log(T)} \geq \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{d(\mu_a, \mu_1)}$$

- Algorithms matching this bound, notably KL-UCB

For m best arm identification, we would want to give:

- A lower bound on the sample complexity $\mathbb{E}[\tau]$ of any δ-PAC algorithm, featuring informational quantities
- δ-PAC algorithms matching this bound

A general lower bound

Theorem [K.,Cappé, Garivier 14]

Any algorithm that is δ-PAC on every binary bandit model such that $\mu_m > \mu_{m+1}$ satisfies, for $\delta \leq 0.15$,

$$\mathbb{E}[\tau] \geq \left(\sum_{t=1}^{m} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{t=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)} \right) \log \frac{1}{2\delta}$$

This result follows from *changes of distributions*:

Lemma

$\nu = (\nu_1, \nu_2, \dots, \nu_K)$, $\nu' = (\nu'_1, \nu'_2, \dots, \nu'_K)$ two bandit models, $A \in \mathcal{F}_\tau$,

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a] \mathrm{KL}(\nu_a, \nu'_a) \geq d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)).$$

An algorithm: KL-LUCB

Generic notation:

- confidence interval (C.I.) on the mean of arm a at round t:

$$\mathcal{I}_a(t) = [L_a(t), U_a(t)]$$

- $J(t)$ the set of m arms with highest empirical means

Our contribution: Introduce KL-based confidence intervals

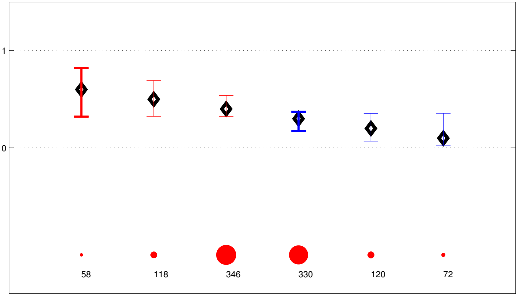
$$\begin{aligned} U_a(t) &= \max\left\{q \geq \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t, \delta)\right\} \\ L_a(t) &= \min\left\{q \leq \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t, \delta)\right\} \end{aligned}$$

for $\beta(t, \delta)$ some exploration rate.

An algorithm: KL-LUCB

At round t, the algorithm:

- draws two well-chosen arms: u_t and l_t (in bold)
- stops when C.I. for arms in $J(t)$ and $J(t)^c$ are separated



$$m = 3, K = 6$$

Set $J(t)$, arm l_t in bold Set $J(t)^c$, arm u_t in bold

Theoretical guarantees

Theorem [K.,Kalyanakrishnan 2013]

KL-LUCB using the exploration rate

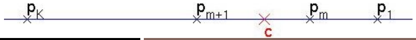
$$\beta(t,\delta) = \log\left(\frac{k_1 K t^{\alpha}}{\delta}\right),$$

with $\alpha > 1$ and $k_1 > 1 + \frac{1}{\alpha - 1}$ satisfies $\mathbb{P}(\hat{\mathcal{S}} = \mathcal{S}_m^*) \geq 1 - \delta$.
For $\alpha > 2$,

$$\mathbb{E}[\tau] \leq 4\alpha H^* \left[\log\left(\frac{k_1 K (H^*)^{\alpha}}{\delta}\right) + \log\log\left(\frac{k_1 K (H^*)^{\alpha}}{\delta}\right)\right] + C_{\alpha},$$

with

$$H^* = \min_{c \in [\mu_{m+1};\mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}.$$



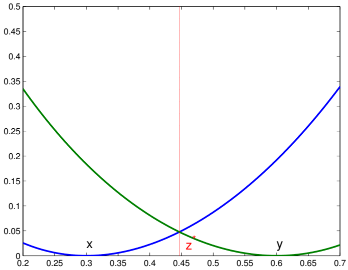
Theoretical guarantees

- **Another informational quantity: Chernoff information**

$$d^*(x,y) := d(z^*,x) = d(z^*,y),$$

where z^* is defined by the equality

$$d(z^*,x) = d(z^*,y).$$



Summary

Lower bound

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log \frac{1}{\delta}} \geq \sum_{t=1}^{m} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{t=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)}$$

Upper bound (for KL-UCB)

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log \frac{1}{\delta}} \leq 8 \min_{c\in[\mu_{m+1};\mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}$$

Refined results for two-armed bandits

A tighter lower bound [K.,Cappé, Garivier 14]

Any algorithm that is δ-PAC on every two-armed bandit model such that $\mu_1 > \mu_2$ satisfies, for $\delta \leq 0.15$,

$$\mathbb{E}[\tau] \geq \frac{1}{d_*(\mu_1, \mu_2)} \log \frac{1}{2\delta}$$

where $d_*(\mu_1, \mu_2) := d(\mu_1, z_*) = d(\mu_2, z^*)$, with z_* defined by

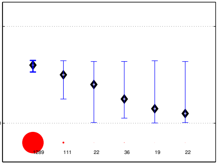
$$d(\mu_1, z^*) = d(\mu_2, z^*).$$

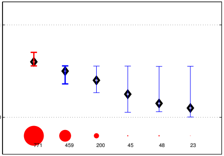
Matching algorithms?

- Uniform sampling is (almost) optimal
- A stopping rule τ based on the difference of empirical means is not optimal (and we propose a new one)

Conclusion

- KL-based confidence intervals are useful in both settings, though KL-UCB and KL-LUCB draw the arms differently





(m=1)

- Do the complexity of these two problems feature the same information-theoretic quantities?

$$\inf_{\substack{\text{consistent}\\ \text{algorithms}}} \limsup_{T\to\infty} \frac{R_T}{\log T} = \sum_{a=2}^{K} \frac{\mu_1 - \mu_a}{d(\mu_a, \mu_1)}$$

$$\inf_{\substack{\delta-PAC\\ \text{algorithms}}} \limsup_{\delta\to\infty} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \geq \sum_{a=1}^{K} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{a=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)}$$

References

- KL-UCB: Cappé, Garivier, Maillard, Munos, Stoltz, *Kulbback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation*, Annals of Statistics, 2013
- KL-LUCB: Kaufmann and Kalyanakrishanan, *Information Complexity in Bandit Subset Selection*, COLT 2013
- The complexity of best arm identification: Kaufmann, Cappé, Garivier, *On the Complexity of Best Arm Identification in Multi-Armed Bandit Models*, arXiv:1407.4443, 2014