On the Complexity of Best Arm Identification in Multi-Armed Bandit Models

Aurélien Garivier

Institut de Mathématiques de Toulouse

Information Theory, Learning and Big Data
Simons Institute, Berkeley, March 2015

Roadmap

The (stochastic) Multi-Armed Bandit Model

Environment K arms with parameters $\theta = (\theta_1, \dots, \theta_K)$ such that for any possible choice of arm $a_t \in \{1, \dots, K\}$ at time t, one receives the reward

$$X_t = X_{a_t,t}$$

where, for any $1 \leq a \leq K$ and $s \geq 1$, $X_{a,s} \sim \nu_a$, and the $(X_{a,s})_{a,s}$ are independent.

Reward distributions $\nu_a \in \mathcal{F}_a$ parametric family, or not: canonical exponential family, general bounded rewards

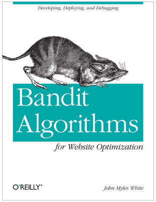
Example Bernoulli rewards: $\theta \in [0, 1]^K$, $\nu_a = \mathcal{B}(\theta_a)$

Strategy The agent's actions follow a dynamical strategy $\pi = (\pi_1, \pi_2, \dots)$ such that

$$A_t = \pi_t(X_1, \dots, X_{t-1})$$

Real challenges

- Randomized clinical trials
  - original motivation since the 1930's
  - dynamic strategies can save resources
- Recommender systems:
  - advertisement
  - website optimization
  - news, blog posts, ...



- Computer experiments
  - large systems can be simulated in order to optimize some criterion over a set of parameters
  - but the simulation cost may be high, so that only few choices are possible for the parameters
- Games and planning (tree-structured options)

Performance Evaluation: Cumulated Regret

Cumulated Reward: $S_T = \sum_{t=1}^{T} X_t$

Goal: Choose π so as to maximize

$$\mathbb{E}[S_T] = \sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{E}\Big[\mathbb{E}[X_t \mathbb{1}\{A_t = a\} | X_1, \dots, X_{t-1}]\Big]$$

$$= \sum_{a=1}^{K} \mu_a \mathbb{E}[N_a^\pi(T)]$$

where $N_a^\pi(T) = \sum_{t \leq T} \mathbb{1}\{A_t = a\}$ is the number of draws of arm a up to time T, and $\mu_a = E(\nu_a)$.

Regret Minimization: maximizing $\mathbb{E}[S_T] \iff$ minimizing

$$R_T = T\mu^* - \mathbb{E}[S_T] = \sum_{a:\mu_a < \mu^*} (\mu^* - \mu_a) \mathbb{E}[N_a^\pi(T)]$$

where $\mu^* \in \max\{\mu_a : 1 \leq a \leq K\}$

Upper Confidence Bound Strategies

UCB [Lai&Robins '85; Agrawal '95; Auer&al '02]

- Construct an upper confidence bound for the expected reward of each arm:

$$\underbrace{\frac{S_a(t)}{N_a(t)}}_{\text{estimated reward}} + \underbrace{\sqrt{\frac{\log(t)}{2N_a(t)}}}_{\text{exploration bonus}}$$

- Choose the arm with the highest UCB

- It is an *index strategy* [Gittins '79]
- Its behavior is easily interpretable and intuitively appealing
- Listen to Robert Nowak's talk tomorrow!

Optimality?

Generalization of [Lai&Robbins '85]

Theorem [Burnetas and Katehakis, '96]

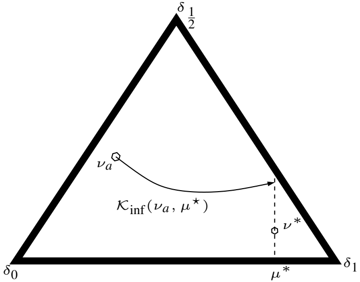
If π is a uniformly efficient strategy, then for any $\theta \in [0,1]^K$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[N_a(T)\right]}{\log(T)} \geq \frac{1}{K_{inf}(\nu_a, \mu^*)}$$

where

$$K_{inf}(\nu_a, \mu^*) = \inf\Big\{K(\nu_a, \nu') : \nu' \in \mathcal{F}_a, E(\nu') \geq \mu^*\Big\}$$

Idea: change of distribution



Reaching Optimality: Empirical Likelihood

The KL-UCB Algorithm, AoS 2013
joint work with O. Cappé, O-A. Maillard, R. Munos, G. Stoltz

Parameters: An operator $\Pi_{\mathcal{F}} : \mathcal{M}_1(\mathcal{S}) \to \mathcal{F}$; a non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

Initialization: Pull each arm of $\{1, \ldots, K\}$ once

for $t = K$ to $T - 1$ **do**

compute for each arm a the quantity

$$U_a(t) = \sup\left\{E(\nu) : \quad \nu \in \mathcal{F} \quad \text{and} \quad KL\Big(\Pi_{\mathcal{F}}\big(\hat{\nu}_a(t)\big), \nu\Big) \leq \frac{f(t)}{N_a(t)}\right\}$$

pick an arm $\quad A_{t+1} \in \underset{a \in \{1,\ldots,K\}}{\arg\max} \; U_a(t)$

end for

Regret bound

Theorem: Assume that $\mathcal{F}$ is the set of finitely supported probability distributions over $\mathcal{S} = [0, 1]$, that $\mu_a > 0$ for all arms a and that $\mu^\star < 1$. There exists a constant $M(\nu_a, \mu^\star) > 0$ only depending on ν_a and $\mu^\star$ such that, with the choice $f(t) = \log(t) + \log\big(\log(t)\big)$ for $t \geq 2$, for all $T \geq 3$:

$$\begin{aligned}
\mathbb{E}\big[N_a(T)\big] \quad \leq \quad & \frac{\log(T)}{K_{inf}\big(\nu_a, \mu^\star\big)} + \frac{36}{(\mu^\star)^4}\big(\log(T)\big)^{4/5} \log\big(\log(T)\big) \\
& + \left(\frac{72}{(\mu^\star)^4} + \frac{2\mu^\star}{(1-\mu^\star)\, K_{inf}\big(\nu_a, \mu^\star\big)^2}\right)\big(\log(T)\big)^{4/5} \\
& + \frac{(1-\mu^\star)^2\, M(\nu_a, \mu^\star)}{2(\mu^\star)^2}\big(\log(T)\big)^{2/5} \\
& + \frac{\log\big(\log(T)\big)}{K_{inf}\big(\nu_a, \mu^\star\big)} + \frac{2\mu^\star}{(1-\mu^\star)\, K_{inf}\big(\nu_a, \mu^\star\big)^2} + 4\,.
\end{aligned}$$

Regret bound

Theorem: Assume that $\mathcal{F}$ is the set of finitely supported probability distributions over $\mathcal{S} = [0,1]$, that $\mu_a > 0$ for all arms a and that $\mu^\star < 1$. There exists a constant $M(\nu_a, \mu^\star) > 0$ only depending on ν_a and $\mu^\star$ such that, with the choice $f(t) = \log(t) + \log\big(\log(t)\big)$ for $t \geq 2$, for all $T \geq 3$:

$$
\begin{aligned}
\mathbb{E}\big[N_a(T)\big] \quad \leq \quad & \frac{\log(T)}{K_{inf}\big(\nu_a, \mu^\star\big)} + \frac{36}{(\mu^\star)^4}\big(\log(T)\big)^{4/5} \log\big(\log(T)\big) \\
& + \left(\frac{72}{(\mu^\star)^4} + \frac{2\mu^\star}{(1-\mu^\star)\, K_{inf}\big(\nu_a, \mu^\star\big)^2}\right)\big(\log(T)\big)^{4/5} \\
& + \frac{(1-\mu^\star)^2\, M(\nu_a, \mu^\star)}{2(\mu^\star)^2}\big(\log(T)\big)^{2/5} \\
& + \frac{\log\big(\log(T)\big)}{K_{inf}\big(\nu_a, \mu^\star\big)} + \frac{2\mu^\star}{(1-\mu^\star)\, K_{inf}\big(\nu_a, \mu^\star\big)^2} + 4\,.
\end{aligned}
$$

Roadmap

1 Simple Multi-Armed Bandit Model

2 Complexity of Best Arm Identification
- Lower bounds on the complexities
- Gaussian Feedback
- Binary Feedback

Best Arm Identification Strategies

A two-armed bandit model is

- a pair $\nu = (\nu_1, \nu_2)$ of probability distributions ('arms') with respective means μ_1 and μ_2
- $a^* = \text{argmax}_a \, \mu_a$ is the (unknown) best arm

Strategy =

- a *sampling rule* $(A_t)_{t \in \mathbb{N}}$ where $A_t \in \{1, 2\}$ is the arm chosen at time t (based on past observations) a sample $Z_t \sim \nu_{A_t}$ is observed
- a *stopping rule* τ indicating when he stops sampling the arms
- a *recommendation rule* $\hat{a}_\tau \in \{1, 2\}$ indicating which arm he thinks is best (at the end of the interaction)

In classical A/B Testing, the sampling rule A_t is uniform on $\{1, 2\}$ and the stopping rule $\tau = t$ is fixed in advance.

Best Arm Identification

Joint work with Emilie Kaufmann and Olivier Cappé (Telecom ParisTech)

Goal: design a strategy $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ such that:

Fixed-budget setting	Fixed-confidence setting
$\tau = t$	$\mathbb{P}_\nu(\hat{a}_\tau \neq a^*) \leq \delta$
$p_t(\nu) := \mathbb{P}_\nu(\hat{a}_t \neq a^*)$ as small as possible	$\mathbb{E}_\nu[\tau]$ as small as possible

See also: [Mannor&Tsitsiklis '04], [Even-Dar&al. '06], [Audibert&al.'10], [Bubeck&al. '11,'13], [Kalyanakrishnan&al. '12], [Karnin&al. '13], [Jamieson&al. '14]...

Two possible goals

Goal: design a strategy $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ such that:

Fixed-budget setting	**Fixed-confidence setting**
$\tau = t$	$\mathbb{P}_\nu(\hat{a}_\tau \neq a^*) \leq \delta$
$p_t(\nu) := \mathbb{P}_\nu(\hat{a}_t \neq a^*)$ as small as possible	$\mathbb{E}_\nu[\tau]$ as small as possible

In the particular case of uniform sampling :

Fixed-budget setting	**Fixed-confidence setting**
classical test of $(\mu_1 > \mu_2)$ against $(\mu_1 < \mu_2)$ based on t samples	sequential test of $(\mu_1 > \mu_2)$ against $(\mu_1 < \mu_2)$ with probability of error uniformly bounded by δ

[Siegmund 85]: sequential tests can save samples !

The complexities of best-arm identification

For a class $\mathcal{M}$ bandit models, algorithm $\mathcal{A} = ((A_t), \tau, \hat{a}_\tau)$ is...

Fixed-budget setting	Fixed-confidence setting
consistent on $\mathcal{M}$ if $\forall \nu \in \mathcal{M}, p_t(\nu) = \mathbb{P}_\nu(\hat{a}_t \neq a^*) \underset{t\to\infty}{\longrightarrow} 0$	δ-PAC on $\mathcal{M}$ if $\forall \nu \in \mathcal{M},\ \mathbb{P}_\nu(\hat{a}_\tau \neq a^*) \leq \delta$

From the literature

$p_t(\nu) \simeq \exp\left(-\frac{t}{CH(\nu)}\right)$ [Audibert&al.'10],[Bubeck&al'11] [Bubeck&al'13],...	$\mathbb{E}_\nu[\tau] \simeq C'H'(\nu)\log(1/\delta)$ [Mannor&Tsitsiklis '04],[Even-Dar&al. '06] [Kalanakrishnan&al'12],...

$\Longrightarrow$ **two complexities**

$\kappa_{\mathrm{B}}(\nu) = \inf\limits_{\mathcal{A}\text{ cons.}} \left(\limsup\limits_{t\to\infty} -\frac{1}{t}\log p_t(\nu)\right)^{-1}$ for a probability of error $\leq \delta$, budget $t \simeq \kappa_B(\nu)\log(1/\delta)$	$\kappa_{\mathrm{C}}(\nu) = \inf\limits_{\mathcal{A}\ \delta-\text{PAC}} \limsup\limits_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)}$ for a probability of error $\leq \delta$, $\mathbb{E}_\nu[\tau] \simeq \kappa_C(\nu)\log(1/\delta)$

Changes of distribution

Theorem: *how to use (and hide) the change of distribution*

Let ν and ν' be two bandit models with K arms such that for all a, the distributions ν_a and ν'_a are mutually absolutely continuous. For any almost-surely finite stopping time σ with respect to $(\mathcal{F}_t)$,

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\sigma)] \, \mathrm{KL}(\nu_a, \nu'_a) \geq \sup_{\mathcal{E} \in \mathcal{F}_\sigma} \mathrm{kl}\big(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})\big),$$

where $\mathrm{kl}(x, y) = x \log(x/y) + (1-x) \log\big((1-x)/(1-y)\big)$.

Useful remark:

$$\forall \delta \in [0,1], \quad \mathrm{kl}\big(\delta, 1-\delta\big) \geq \log \frac{1}{2.4\,\delta} \ ,$$

General lower bounds

Theorem 1

Let $\mathcal{M}$ be a class of two armed bandit models that are continuously parametrized by their means. Let $\nu = (\nu_1, \nu_2) \in \mathcal{M}$.

Fixed-budget setting	**Fixed-confidence setting**
any consistent algorithm satisfies	any δ-PAC algorithm satisfies
$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq \mathrm{K}^*(\nu_1,\nu_2)$	$\mathbb{E}_\nu[\tau] \geq \frac{1}{\mathrm{K}_*(\nu_1,\nu_2)} \log\left(\frac{1}{2.4\delta}\right)$
with $\mathrm{K}^*(\nu_1,\nu_2)$ $= \mathrm{KL}(\nu^*,\nu_1) = \mathrm{KL}(\nu^*,\nu_2)$	with $\mathrm{K}_*(\nu_1,\nu_2)$ $= \mathrm{KL}(\nu_1,\nu_*) = \mathrm{KL}(\nu_2,\nu_*)$
Thus, $\kappa_B(\nu) \geq \frac{1}{\mathrm{K}^*(\nu_1,\nu_2)}$	Thus, $\kappa_C(\nu) \geq \frac{1}{\mathrm{K}_*(\nu_1,\nu_2)}$

Gaussian Rewards: Fixed-Budget Setting

For fixed (known) values σ_1, σ_2, we consider Gaussian bandit models

$$\mathcal{M} = \left\{\nu = \left(\mathcal{N}\left(\mu_1, \sigma_1^2\right), \mathcal{N}\left(\mu_2, \sigma_2^2\right)\right) : (\mu_1, \mu_2) \in \mathbb{R}^2, \mu_1 \neq \mu_2\right\}$$

- Theorem 1:

$$\kappa_B(\nu) \geq \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$

- A strategy allocating $t_1 = \left\lceil \frac{\sigma_1}{\sigma_1+\sigma_2} t \right\rceil$ samples to arm 1 and $t_2 = t - t_1$ samples to arm 1, and recommending the empirical best satisfies

$$\liminf_{t\to\infty} -\frac{1}{t} \log p_t(\nu) \geq \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1 + \sigma_2)^2}$$

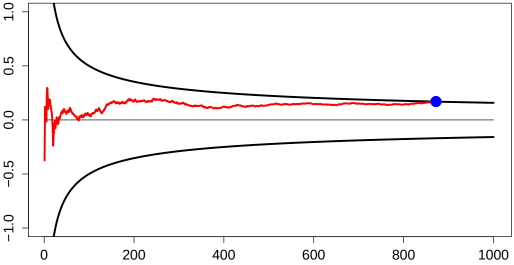
$$\boxed{\kappa_B(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}}$$

Gaussian Rewards: Fixed-confidence setting

The α-Elimination algorithm with exploration rate $\beta(t,\delta)$

- ➜ chooses A_t in order to keep a proportion $N_1(t)/t \simeq \alpha$
- ➜ if $\hat{\mu}_a(t)$ is the empirical mean of rewards obtained from a up to time t, $\sigma_t^2(\alpha) = \sigma_1^2/\lceil \alpha t \rceil + \sigma_2^2/(t - \lceil \alpha t \rceil)$,

$$\tau = \inf\left\{ t \in \mathbb{N} : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)} \right\}$$



- ➜ recommends the empirical best arm $\hat{a}_\tau = \text{argmax}_a \hat{\mu}_a(\tau)$

Gaussian Rewards: Fixed-confidence setting

- From Theorem 1:

$$\mathbb{E}_\nu[\tau] \geq \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\log\left(\frac{1}{2.4\delta}\right)$$

- $\frac{\sigma_1}{\sigma_1+\sigma_2}$-Elimination with $\beta(t,\delta) = \log\frac{t}{\delta} + 2\log\log(6t)$ is δ-PAC and

$$\forall \epsilon > 0, \quad \mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\log\left(\frac{1}{2.4\delta}\right) + \underset{\delta\to 0}{o_\epsilon}\left(\log\frac{1}{\delta}\right)$$

$$\boxed{\kappa_C(\nu) = \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}}$$

Gaussian Rewards: Conclusion

For any two fixed values of σ_1 and σ_2,

$$\boxed{\kappa_B(\nu) = \kappa_C(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}}$$

If the variances are equal, $\sigma_1 = \sigma_2 = \sigma$,

$$\kappa_B(\nu) = \kappa_C(\nu) = \frac{8\sigma^2}{(\mu_1 - \mu_2)^2}$$

- uniform sampling is optimal only when $\sigma_1 = \sigma_2$
- $1/2$-Elimination is δ-PAC for a smaller exploration rate $\beta(t, \delta) \simeq \log(\log(t)/\delta)$

Binary Rewards: Lower Bounds

$$\mathcal{M} = \{\nu = (\mathcal{B}(\mu_1), \mathcal{B}(\mu_2)) : (\mu_1, \mu_2) \in]0;1[^2, \mu_1 \neq \mu_2\},$$

shorthand: $\mathrm{K}(\mu, \mu') = \mathrm{KL}\left(\mathcal{B}(\mu), \mathcal{B}(\mu')\right)$.

Fixed-budget setting	**Fixed-confidence setting**
any consistent algorithm satisfies $\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq \mathrm{K}^*(\mu_1, \mu_2)$ (Chernoff information)	any δ-PAC algorithm satisfies $\mathbb{E}_\nu[\tau] \geq \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)} \log\left(\frac{1}{2\delta}\right)$

$$\boxed{\mathrm{K}^*(\mu_1, \mu_2) > \mathrm{K}_*(\mu_1, \mu_2)}$$

Binary Rewards: Uniform Sampling

	For any consistent...	For any δ-PAC...
... algorithm	$p_t(\nu) \gtrsim e^{-\mathrm{K}^*(\mu_1,\mu_2)t}$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)}$
... algorithm using uniform sampling	$p_t(\nu) \gtrsim e^{-\frac{\mathrm{K}(\overline{\mu},\mu_1)+\mathrm{K}(\overline{\mu},\mu_2)}{2}t}$ with $\overline{\mu} = f(\mu_1, \mu_2)$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{2}{\mathrm{K}(\mu_1,\underline{\mu})+\mathrm{K}(\mu_2,\underline{\mu})}$ with $\underline{\mu} = \frac{\mu_1+\mu_2}{2}$

Remark: Quantities in the same column appear to be close from one another

$\Rightarrow$ **Binary rewards: uniform sampling close to optimal**

Binary Rewards: Uniform Sampling

	For any consistent...	For any δ-PAC...
... algorithm	$p_t(\nu) \simeq e^{-\mathrm{K}^*(\mu_1,\mu_2)t}$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)}$
... algorithm using uniform sampling	$p_t(\nu) \simeq e^{-\frac{\mathrm{K}(\overline{\mu},\mu_1)+\mathrm{K}(\overline{\mu},\mu_2)}{2}t}$ with $\overline{\mu} = f(\mu_1,\mu_2)$	$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \gtrsim \frac{2}{\mathrm{K}(\mu_1,\underline{\mu})+\mathrm{K}(\mu_2,\underline{\mu})}$ with $\underline{\mu} = \frac{\mu_1+\mu_2}{2}$

Remark: Quantities in the same column appear to be close from one another

$\Rightarrow$ **Binary rewards: uniform sampling close to optimal**

Binary Rewards: Fixed-Budget Setting

In fact,

$$\boxed{\kappa_B(\nu) = \frac{1}{\mathrm{K}^*(\mu_1, \mu_2)}}$$

The algorithm using uniform sampling and recommending the empirical best arm is very close to optimal

Binary Rewards: Fixed-Confidence Setting

δ-PAC algorithms using uniform sampling satisfy

$$\frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \geq \frac{1}{I_*(\nu)} \quad \text{with} \quad I_*(\nu) = \frac{\mathrm{K}\left(\mu_1, \frac{\mu_1+\mu_2}{2}\right) + \mathrm{K}\left(\mu_2, \frac{\mu_1+\mu_2}{2}\right)}{2}.$$

The algorithm using uniform sampling and

$$\tau = \inf\left\{ t \in 2\mathbb{N}^* : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \log \frac{\log(t)+1}{\delta} \right\}$$

is δ-PAC but not optimal: $\frac{\mathbb{E}[\tau]}{\log(1/\delta)} \simeq \frac{2}{(\mu_1-\mu_2)^2} > \frac{1}{I_*(\nu)}$.

A better stopping rule NOT based on the difference of empirical means

$$\tau = \inf\left\{ t \in 2\mathbb{N}^* : t\, I_*(\hat{\mu}_1(t), \hat{\mu}_2(t)) > \log \frac{\log(t)+1}{\delta} \right\}$$

Binary Rewards: Conclusion

Regarding the complexities:

- $\kappa_B(\nu) = \frac{1}{\mathrm{K}^*(\mu_1,\mu_2)}$
- $\kappa_C(\nu) \geq \frac{1}{\mathrm{K}_*(\mu_1,\mu_2)} > \frac{1}{\mathrm{K}^*(\mu_1,\mu_2)}$

Thus

$$\boxed{\kappa_C(\nu) > \kappa_B(\nu)}$$

Regarding the algorithms

- There is not much to gain by departing from uniform sampling
- In the fixed-confidence setting, a sequential test based on the difference of the empirical means is no longer optimal

Conclusion

- the complexities $\kappa_B(\nu)$ and $\kappa_C(\nu)$ are not always equal (and feature some different informational quantities)
- strategies using random stopping do not necessarily lead to a saving in terms of the number of sample used
- for Bernoulli distributions and Gaussian with similar variances, strategies using uniform sampling are (almost) optimal
- Generalization to m best arms identification among K arms

Elements of Bibliography (see references therein!)

1. **[Lai&Robins '85]** T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1) :4-22, 1985.
2. **[Agrawal '95]** R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. Advances in Applied Probability, 27(4) :1054-1078, 1995.
3. **[Auer&al '02]** P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2) :235-256, 2002.
4. **[Even-Dar&al '06]** Action elimination and stopping conditions for multi-armed bandit and reinforcement leraning problems, JMLR 7:1079-1105, 2006.
5. **[Audibert&al '09]** J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19), 2009
6. **[Filippi &al '10]** S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In Allerton Conf. on Communication, Control, and Computing, Monticello, US, 2010.
7. **[Cappé&al '13]** O. Cappé, A. Garivier, O-A. Maillard, R Munos, G. Stoltz. Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation . Annals of Statistics (41:3) Jun. 2013 pp.1516-1541.
8. **[Abbasi-Yadkori&al '11]** Yasin Abbasi-Yadkori, Dávid Pál, Csaba Szepesvári: Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems. CoRR abs/1102.2670: (2011)
9. **[Bubeck&Cesa-Bianchi '12]** S. Bubeck and N. Cesa-Bianchi, Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Foundations and Trends in Machine Learning 5(1): 1-122 (2012)
10. **[Cappé&al '13]** O. Cappé, A. Garivier, O-A. Maillard, R Munos, G. Stoltz. Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation . Annals of Statistics (41:3) Jun. 2013 pp.1516-1541.
11. **[Jamieson&al '14]** K. Jamieson, M. Malloy, R. Nowak and S. Bubeck. lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. COLT 2014 :423-439
12. **[Kaufmann&al '15]** E. Kaufmann, O. Cappé, A. Garivier, On the Complexity of Best Arm Identification in Multi-Armed Bandit Models, ArXiv:1407.4443