A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences

Odalric-Ambrym Maillard[1], Rémi Munos[1] and Gilles Stoltz[2]

1: INRIA Lille Nord-Europe, 2: CNRS – ENS Paris – HEC Paris

July 11, 2011

Multi-Armed Bandit setting

Motivation

We first present known lower bounds and historical attempts to get matching upper bounds.

The stochastic Multi-Armed Bandit setting

Setting $\mathcal{D}$, a set of real-valued probability distributions.
ex: $\mathcal{D} = \mathcal{P}([0,1])$, distributions with support in $[0,1]$.

$\mathcal{A}$, a finite set of arms. Each arm $a \in \mathcal{A}$ is associated with an unknown probability distribution $\nu_a \in \mathcal{D}$ with mean μ_a.

Game The game is *sequential*: At each round $t \geqslant 1$,

- the player first picks an arm $A_t \in \mathcal{A}$ according to its dynamical policy $\pi = (\pi_1, \pi_2, \cdots)$:
$$A_t = \pi_t(X_1, \ldots, X_{t-1})$$
- then receives (and sees) a stochastic payoff $X_t \sim \nu_{A_t}$.

Measure of performance for a Multi-Armed Bandit

Optimal arm An optimal arm $a^\star \in \operatorname{argmax}\{\mu_a ; a \in \mathcal{A}\}$ is determined by its mean reward. We write $\mu^\star = \max\{\mu_a ; a \in \mathcal{A}\}$.

Regret The expected regret at round $T \geqslant 1$ for the dynamical policy $\pi = (\pi_1, \pi_2, \dots)$ is:

$$R_T \overset{\text{def}}{=} \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} \mu_{A_t}\right] = \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}\left[N_T^\pi(a)\right],$$

where $\Delta_a \overset{\text{def}}{=} \mu^\star - \mu_a$ and $N_T^\pi(a) \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{\{A_t = a\}}$.

Historical overview (I): First step

Definition A consistent π satisfies for any bandit, suboptimal arm a and any $\beta > 0$, $\mathbb{E}(N_T^\pi(a)) = o(T^\beta)$.

"π does not pull a bad arm too often"

Lower-bound

Theorem (Lai &Robbins, 1985)

If π is consistent, then for any bandit with some 1-dimensional parametric $\mathcal{D}$ we have

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geqslant \sum_{a:\Delta_a>0} \frac{\Delta_a}{\mathcal{K}(\nu_a, \nu^\star)},$$

where $\mathcal{K}$ is the Kullback-Leibler divergence. It includes e.g. Bernoulli, Poisson, Gaussian with known variance...

Upper-bound Explicit algorithm with a *matching* asymptotic upper-bound.

Historical overview (II): Extension

Lower-bound (Burnetas&Katehakis, 1996): they extend Lai&Robbins' asymptotic lower-bound to **arbitrary** $\mathcal{D} \subset \mathcal{P}([0,1])$, where $\mathcal{K}(\nu_a, \nu^\star)$ is replaced with $\mathcal{K}_{\inf}(\nu_a, \mu^\star)$:

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star) \overset{\text{def}}{=} \inf_{\nu \in \mathcal{D}:\, \nu \text{ has mean} > \mu^*} \mathcal{K}(\nu_a, \nu)\,.$$

Extension To achieve the optimal mean reward $\mu^\star$, we do not need $\mathcal{K}(\nu_a, \nu^\star)$: it measures how far ν_a is from $\nu^\star$; we just need to measure how far ν_a is from any distribution with mean higher than $\mu^\star$.

Upper-bound Explicit algorithm with matching upper-bound for distributions with finite support, and some finite-dimensional parametric classes.

Intuition: from $\mathcal{K}$ to $\mathcal{K}_{\inf}$

Important remark

$$\mathcal{K}_{\inf}(\nu_a, \mu^\star) \leqslant \mathcal{K}(\nu_a, \nu^\star)$$

This becomes an equality e.g. for Bernoulli distributions and more generally for 1-dimensional exponential families (see previous talk), but the inequality can be arbitrary loose in general.

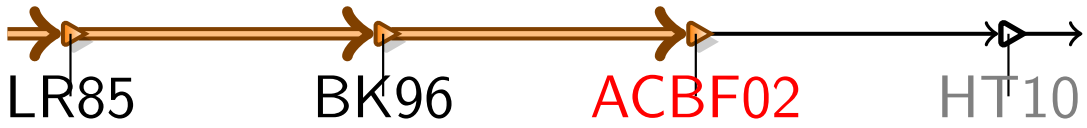
Example Let $\nu_a = U(\{1/4, 1\})$ and $\nu^\star = U(\{1/2, 1\})$ each with two atoms $\in [0, 1]$. Thus $\mu_a = \frac{5}{8}$ and $\mu^\star = \frac{3}{4}$.

- $\mathcal{K}(\nu_a, \nu^\star) = \infty$,
- $\mathcal{K}_{\inf}(\nu_a, \mu^\star) = \mathcal{K}_{\inf}(\nu_a, 3/4) = \frac{1}{2} \log \frac{9}{8} \sim 0.0589$.

For such a bandit, $\liminf_{T \to \infty} \dfrac{R_T}{\log T} \geqslant 2.122$.

Message The improvement by (Burnetas&Katehakis, 1996) is significant.

Historical overview (III): Non-asymptotic



Non-asymptotic bound

Theorem (Auer, Cesa-Bianchi & Fischer, 2002)

For an arbitrary class of distribution $\mathcal{D} \subset \mathcal{P}([0,1])$, the UCB2($\alpha$)-strategy satisfies for all $\alpha \in (0,1)$ and T:

$$R_T \leqslant \sum_{a \in \mathcal{A}} \frac{(1+\alpha)(1+4\alpha)\Delta_a}{2\Delta_a^2} \log T + C(\alpha, \Delta(a)),$$

where C is explicit and does not depend on T.

Asymptotic gap

The price for this non-asymptotic result is that UCB-like algorithms are no longer asymptotically optimal, since $\mathcal{K}_{\inf}(\nu_a, \mu^\star) \geqslant \mathcal{K}\Big(Bern(\mu_a),\, Bern(\mu^\star)\Big) \geqslant 2\Delta_a^2$.

Historical overview (III): General asymptotic

Asymptotic bound

Theorem (Honda &Takemura, 2010)

The DMED strategy achieves, for an arbitrary class of distribution $\mathcal{D} \subset \mathcal{P}([0,1])$ *the asymptotic bound:*

$$\limsup_{T\to\infty} \frac{R_T}{\log T} \leqslant \sum_{a:\Delta_a>0} \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

Numerical issue

They show that an efficient implementation is possible, that does not need to know the support of the distribution of the arms in advance.

Overview

This talk We derive a non-asymptotic analysis (like for UCB) for an algorithm that matches the asymptotic lower-bound involving $\mathcal{K}_{\inf}$ for some classes $\mathcal{D}$.

We consider the important class $\mathcal{D} = \mathcal{P}_F([0,1])$ of distributions with a finite, yet unknown, number of atoms. We explicit further the case of Bernoulli distributions.

Distributions with finite support

We assume that $\mathcal{D}$ is the set $\mathcal{P}_F\big([0,1]\big)$ that consists of distributions with finite support.

Algorithm: the $\mathcal{K}_{\inf}$-strategy.

Parameters: A non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

"Think of $f(t) \simeq \log(t)$"

Initialization: Pull each arm of $\mathcal{A}$ once

For rounds $t+1$, where $t \geqslant |\mathcal{A}|$,

– compute for each arm $a \in \mathcal{A}$ the quantity

$$B^+_{a,t} = \max\left\{\mu \in [0,1] : \quad \mathcal{K}_{\inf}\left(\widehat{\nu}_{a,N_t(a)},\, \mu\right) \leqslant \frac{f(t)}{N_t(a)}\right\},$$

where $$\widehat{\nu}_{a,N_t(a)} = \frac{1}{N_t(a)} \sum_{s \leqslant t:\, A_s = a} \delta_{X_s}\,;$$

– pull any arm $A_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} B^+_{a,t}$.

Non-asymptotic upper-bound for the $\mathcal{K}_{\inf}$-strategy

Theorem (M., Munos & Stoltz, 2011)

For $f(t) = \log t$, for all suboptimal arm a, for all $c_a > 0$ we have

$$\mathbb{E}\big[N_T(a)\big] \leqslant \frac{(1+c_a)\log T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + \underbrace{\frac{1}{\varepsilon^2}\log\left(\frac{1}{1-\mu^*+\varepsilon}\right)\sum_{k=1}^{T}(k+1)^{|\mathcal{S}^\star|}\, e^{-k\varepsilon^2}}_{o(\log(T))}$$

$$+ \underbrace{\frac{1}{1-e^{-\Theta_a(c_a,\varepsilon)}} + \frac{1}{(\Delta_a-\varepsilon)^2} + 1}_{O(1)},$$

where $\varepsilon = \frac{(1-\mu^\star)c_a}{1+c_a}\Delta_a^2$ *and* $\Theta_a(c_a,\varepsilon) \simeq \theta_a\left(\frac{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}{1+c_a} + \frac{\varepsilon}{1-\mu^\star}\right)$, *is controlled by* $\theta_a(\gamma) = \inf\Big\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\Big\}$.

One-slide summary.

- First, the $\mathcal{K}_{\inf}$-strategy achieves the asymptotic optimal behavior for the class of distributions with finite support, i.e.

$$\limsup_{T\to\infty} \frac{R_T}{\log T} \leqslant \sum_a \frac{\Delta_a}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

- The algorithm does not need to know the support of the distributions, as explained in (Honda &Takemura, 2010).
- The bound for the $\mathcal{K}_{\inf}$-strategy is non-asymptotic and involves:
  - The leading term $\log(T)$ with the optimal constant.
  - A second order term $o(\log(T))$ that depends on the size of the support $\mathcal{S}^\star$ of the optimal arm.
  - $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$, a feature that I will explain in a minute.

The explicit case of Bernoulli distributions (I)

In the case of Bernoulli distributions, we derive the following similar bound where we have explicited the second order and constant terms.

Theorem (M., Munos &Stoltz 2011)

When $\mu^\star \in (0,1)$, for all non-decreasing functions $f : \mathbb{N} \to \mathbb{R}_+$ such that $f(1) \geqslant 1$, the expected regret R_T of the $\mathcal{K}$-strategy (simplification of the $\mathcal{K}_{\inf}$-strategy for the Bernoulli case) satisfies

$$R_T \leqslant \inf_{(c_a)_{a\in\mathcal{A}}} \sum_{a\in\mathcal{A}} \Delta_a \left(\frac{(1+c_a)\, f(T)}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} + 4e \sum_{t=|\mathcal{A}|}^{T-1} \left\lceil f(t) \log t \right\rceil e^{-f(t)} + \frac{(1+c_a)^2}{8\, c_a^2 \Delta_a^2 \, \min\left\{\sigma_a^4,\, \sigma^{\star,4}\right\}} \mathbb{I}_{\{\mu_a \in (0,1)\}} + 3 \right).$$

where σ_a is the variance of arm a.
For $\mu^\star = 0$, $R_T = 0$. For $\mu^\star = 1$, $R_T \leqslant 2\big(|\mathcal{A}| - 1\big)$.

The explicit case of Bernoulli distributions (II)

In particular, with an appropriate choice of the $(c_a)_{a\in\mathcal{A}}$ (which are parameters of the analysis only, not of the algorithm), and of $f(t)$, we recover the constant 1 in factor of the leading term:

Corollary (M., Munos & Stoltz, 2011)

When $\mu^\star \in (0,1)$, for the choice $f(t) = \log(et\log^3(et))$, the expected regret R_T of the $\mathcal{K}$-strategy is upper bounded by

$$R_T \leqslant \sum_{a\in\mathcal{A}} \frac{\Delta_a}{\mathcal{K}\big(\beta(\mu_a),\, \beta(\mu^\star)\big)} \log(T) + O(\log(T)^{2/3})$$

where the second order term has an explicit and closed-form expression.

We do not know whether this second order term is optimal.

Concentration tools

- To control a sub-optimal arm a, we need to show that
  "The empirical distribution of a has not high mean"
  We show that $\left\{\nu' \in \mathcal{P}_F([0,1]) : \ \mathcal{K}_{\inf}\left(\nu', \mu^\star\right) \leqslant \gamma\right\}$ is convex.
  This enables us to apply a non-asymptotic Sanov's lemma:
  Lemma (Control of a sub-optimal arm)

$$\forall \gamma > 0 \ \ \mathbb{P}_{\nu_a}\left\{\mathcal{K}_{\inf}\left(\widehat{\nu}_{a,k},\, \mu^\star\right) \leqslant \gamma\right\} \leqslant e^{-k\,\theta_a(\gamma)}\,.$$

  This lemma explains both the leading and the constant term.

Concentration tools

- Now to control an optimal arm, we need to show that "The empirical distribution of $a^\star$ is close to $\nu^\star$"
  But $\left\{\nu' \in \mathcal{P}_F([0,1]) : \ \mathcal{K}_{\inf}\left(\nu', \mu^\star\right) > \gamma\right\}$ is not convex. Still, we can use the method of types:
  Lemma (Control of an optimal arm)
  If $\nu^\star$ has a finite support $\mathcal{S}^\star$, then for all $k \geqslant 1$, $\gamma > 0$,
  $$\mathbb{P}_{\nu^\star}\left\{\mathcal{K}\left(\widehat{\nu^\star}_k, \nu^\star\right) > \gamma\right\} \leqslant (k+1)^{|\mathcal{S}^\star|}\, e^{-k\gamma}\,.$$
  This lemma explains the second order term.

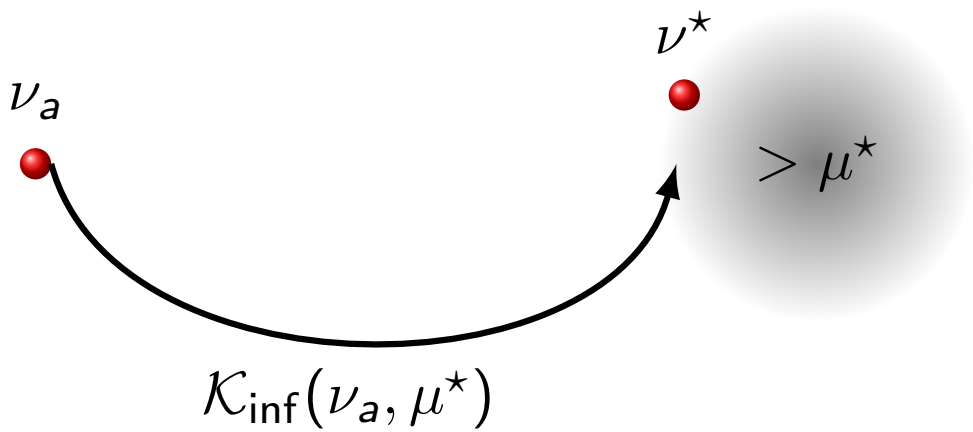
Intuition: Information complexity of sub-optimal arms

- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.

A geometric interpretation: in gray the set of distributions with mean higher than $\mu^\star$.

Intuition: Information complexity of sub-optimal arms

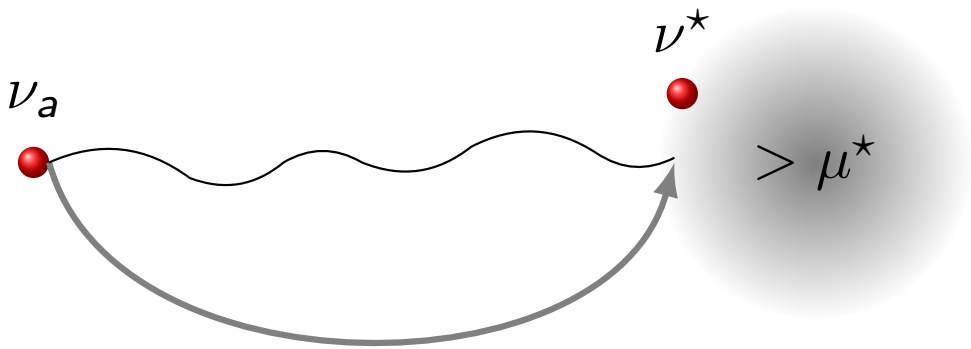
- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



Information gap between ν_a and distributions with high mean.

Intuition: Information complexity of sub-optimal arms

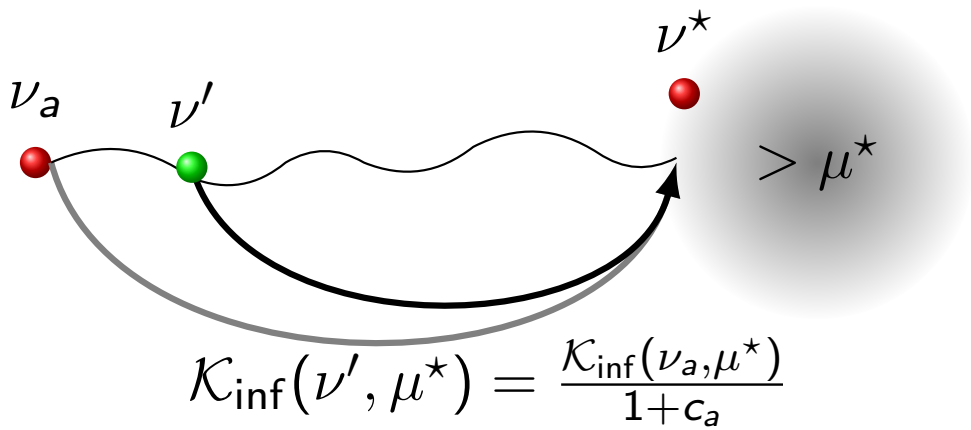
- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



Let us consider some geodesic starting from ν_a.

Intuition: Information complexity of sub-optimal arms

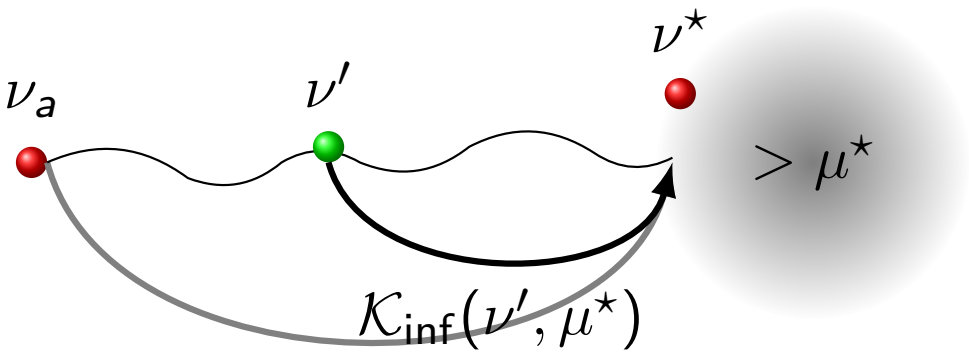
- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



We can move ν_a towards the region of high mean.

Intuition: Information complexity of sub-optimal arms

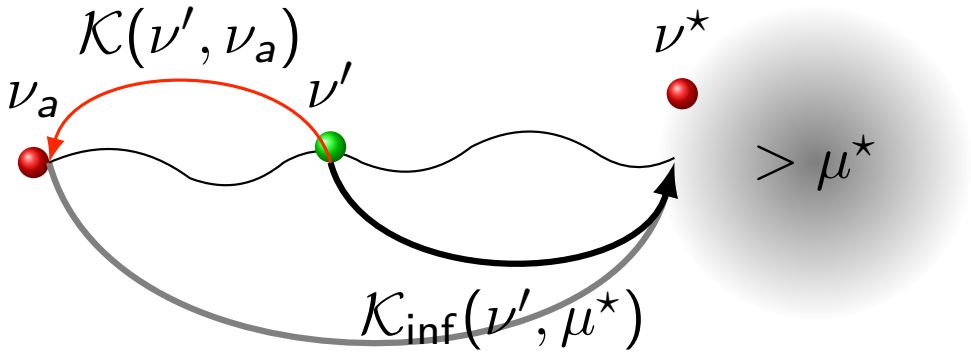
- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



We can move ν_a towards the region of high mean.

Intuition: Information complexity of sub-optimal arms

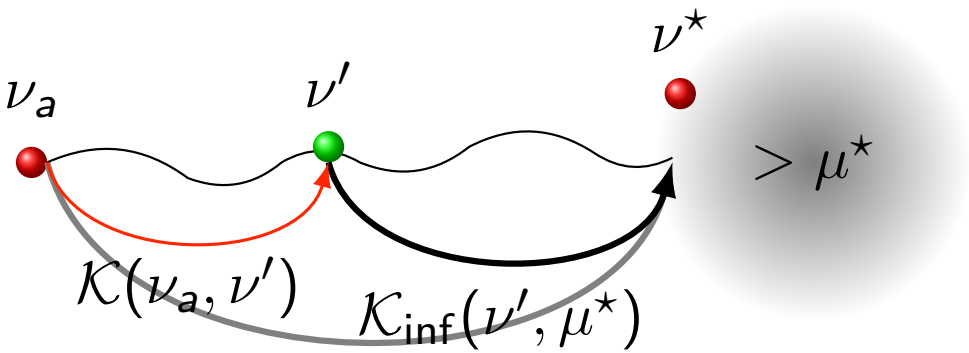
- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



This is the meaning of $\theta_a(\gamma)$ for $\gamma = \frac{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}{1+c_a}$.

Intuition: Information complexity of sub-optimal arms

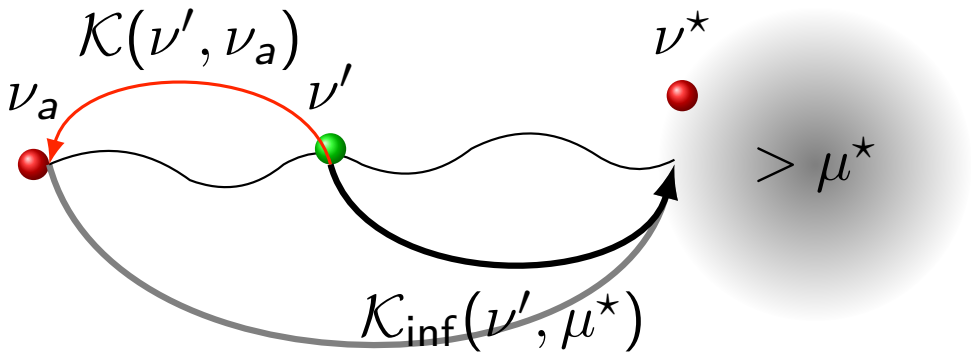
- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



This has nothing to do and gives no information on θ_a.

Intuition: Information complexity of sub-optimal arms

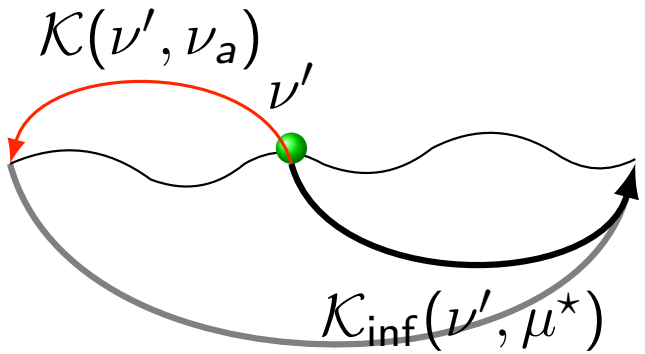
- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



Meaning: how $\mathcal{K}(\nu', \nu_a)$ evolves when we move ν' from ν_a to the gray region?

Intuition: Information complexity of sub-optimal arms

- $\theta_a(\gamma) = \inf\left\{\mathcal{K}(\nu', \nu_a) : \quad \nu' \quad \text{s.t.} \quad \mathcal{K}_{\inf}(\nu', \mu^\star) < \gamma\right\}$ vanishes for $\mathcal{K}_{\inf}(\nu_a, \mu^\star) < \gamma$. Now since the algorithm uses $\gamma = \frac{f(t)}{N_t(a)}$, we have to wait that $N_t(a) > \frac{f(t)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}$ before having an exponential decay of bad choices.
- The rate of decay then depends on the struture of the distributions.



- It can be made explicit for the Bernoulli case (see the paper). However in general, this intrinsically depends on the considered class of distributions.

Conclusion and Future work

- We provided a finite-time analysis of the (asymptotically optimal) $\mathcal{K}_{\inf}$-strategy in the case of finitely supported distributions $\mathcal{P}_F([0,1])$.
- The extension to the case of general distributions needs new tools:
  (1) Ensuring that $\exists\, \gamma < \mathcal{K}_{\inf}(\nu_a, \mu^\star)$ such that $0 < \theta_a(\gamma) < \infty$ seems not that easy for general distributions.
  (2) The method of types only applies to $\mathcal{P}_F([0,1])$, thus we need extensions of non-asymptotic Sanov's lemma to non-convex sets.
- Exploring other directions for such extensions (exponential families, histograms...) as well as finite-time lower bounds is left for future work.

Köszönöm!