# Multi-Armed Bandit Formulations for Identification and Control

Cristian R. Rojas

Joint work with Matías I. Müller and Alexandre Proutiere

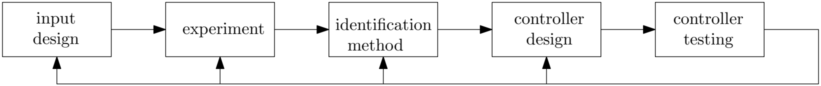
*KTH Royal Institute of Technology, Sweden*

ERNSI, September 24-27, 2017

# Outline

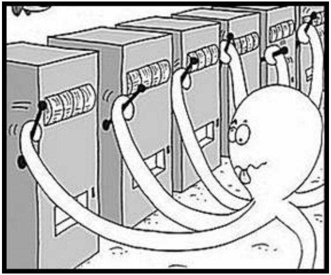
- Motivation: identification and control
- Introduction to multi-armed bandits
- Lower bounds and optimal algorithms
- Application to $\mathcal{H}_\infty$-norm estimation
- Summary

# Motivation: Identification and control




- Need to make choices in identification procedure focusing on end goal
- *E.g.*, input design can emphasize system properties of interest, while properties of little or no interest can be *hidden*
- However, to design a good input requires knowing what we don't know yet: the true system!
- This can be solved by *adaptively* tuning the input:
  - S.D. Silvey, *Optimal Design: An Introduction to the Theory for Parameter Estimation*. Chapman and Hall, 1980
  - L. Pronzato, "Optimal experimental design and some related control problems". *Automatica*, 44(2):303–325, 2008
  - L. Gerencsér, H. Hjalmarsson and L. Huang. "Adaptive input design for LTI systems". *IEEE Transactions on Automatic Control*, 62(5):2390–2405, 2017

# Introduction to Multi-Armed Bandits



- Popular machine learning framework for adaptive control (but where the “plant” is static)
- Name coined in 1952 by Herbert Robbins, in the context of Sequential Design of Experiments
- Exploration vs exploitation dilemma
- First asymptotically optimal solution proposed in 1985 by T.L. Lai and H. Robbins

**Basic setup:**

- There are $K$ *arms* (slot machines) to choose from
- One can play one arm in each round
- Each arm $j$ gives a reward $X_{j,t} \in \{0, 1\}$ ($t$: round)
  (**Obs** only the reward of the selected arm $j$ is revealed)
- Problem: which machine should one play in each round?

Performance of a strategy measured in terms of *expected cumulative regret*:

$$R(T) = \sum_{i=1}^{T} (\mu^* - \mathrm{E}\{\mu_{a(i)}\})$$

where: $T$: number of rounds
$\mu^* = \max_i \mu_i$: best reward
$a(i)$: arm chosen at round $i$

# Introduction to Multi-Armed Bandits (cont.)

**Different formulations:**

- **Stochastic:** rewards sampled from an unknown distribution (independent between rounds)
  Example: $X_{j,t}$: i.i.d. Bernoulli variables with unknown mean $\mu_j$
- **Adversarial:** rewards chosen by an *adversary*
  - *Oblivious adversary:*
    $X_{j,t}$ chosen a priori (at round 0)
  - *Adaptive adversary:*
    $X_{j,t}$ chosen based on history of selected arms and rewards so far
- **Markovian:** rewards are Markov processes (evolving only when the respective arm is chosen)
  - Large literature from the 70's based on Gittins indices

We will focus mostly on stochastic MABs

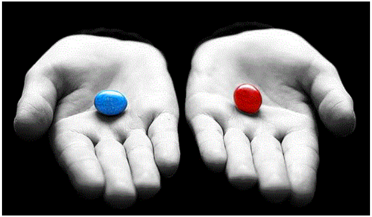
(for applications of adversarial MABs to identification, check

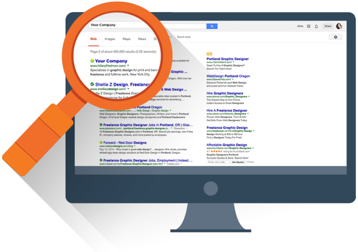
G. Rallo *et. al.* "Data-driven $\mathcal{H}_\infty$-norm estimation via expert advice". *CDC*'17.)

**Applications:**

- Clinical trials
- Ad placement on webpages
- Recommender systems
- Computer game-playing
- . . .


| | | | | | |
|---|---|---|---|---|---|
| 2 | | | 4 | 5 | 2.94* |
| 5 | | 4 | | | 1 |
| | | 5 | | 2 | 2.48* |
| | 1 | | 5 | | 4 |
| | | 4 | | | 2 |
| 4 | 5 | | 1 | | 1.12* |





# Lower bounds and optimal algorithms

- The performance of a strategy depends on the true reward distribution of the arms
- To obtain reasonable problem-dependent lower bounds on the achievable performance, one needs to restrict the class of strategies

Consider a Bernoulli MAB:

**Definition (Uniformly good strategy)**

A strategy is called *uniformly good / efficient* if, for any mean reward distribution $(\mu_1, \ldots, \mu_K)$, the number of times $T_j(t)$ that any suboptimal arm $j$ ($\mu_j \neq \mu^*$) is chosen up to round $t$ satisfies

$$\mathrm{E}\{T_j(t)\} = o(t^\alpha), \qquad \text{for all } \alpha > 0$$

**Note** This definition should be suitably changed for other types of MABs

# Lower bounds and optimal algorithms (cont.)

**Theorem (Lower bound (Lai&Robbins, 1985))**

*For any uniformly good strategy and suboptimal arm $j$,*

$$\liminf_{t\to\infty} \frac{T_j(t)}{\log t} \geqslant \frac{1}{I(\mu_j, \mu^*)} \quad \text{w. p. } 1,$$

*where*

$$I(x,y) = x \log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$$

*is the KL divergence between two Bernoulli distributions with means $x$ and $y$. Therefore,*

$$\liminf_{t\to\infty} \frac{R(t)}{\log t} \geqslant \sum_{j=1}^{K} \frac{\mu^* - \mu_j}{I(\mu_j, \mu^*)}$$

**Idea of proof** Transform problem into hypothesis testing: a unif. good strategy should detect quickly the best arm, but for that it needs to collect enough samples of every suboptimal arm. Stein-Chernoff's Lemma provides a lower bound for $T_j(t)$ to achieve consistent detection

# Lower bounds and optimal algorithms (cont.)

Two large families of asymptotically optimal algorithms:

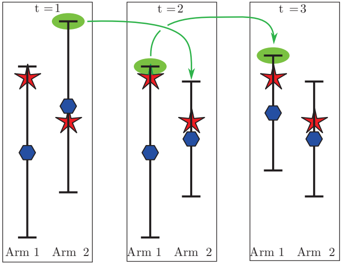
- Upper confidence bound (UCB)
- Thompson sampling

# Lower bounds and optimal algorithms (cont.)

**UCB algorithm:**

At each round $t$:

- construct a confidence interval around $\mu_j$ for each arm $j$, of significance level $\alpha_t$
- choose arm whose upper confidence bound is the largest (*Optimism in the face of uncertainty*)




Significance level $\alpha_t$ should be carefully tuned so that $\alpha_t \to 1$, to obtain an asymptotically optimal strategy. The resulting upper bounds are

$$b_j(t) = \hat{\mu}_j(t) + \sqrt{\frac{2\log(t)}{T_j(t)}}, \quad \hat{\mu}_j(t): \text{ average reward of arm } j$$

$$T_j(t): \text{ \# times arm } j \text{ has been played up to round } t$$

Similar to the *Bet on the Best* (BoB) principle of S. Bittanti and M.C. Campi (*Comm. Inf. & Syst.*, 6(4):299–320, 2006)

# Lower bounds and optimal algorithms (cont.)

For Bernoulli rewards, UCB algorithm gives logarithmic regret, but its regret does not exactly match the lower bound

A variant, called KL-UCB, does match the lower bound; the upper bound used is

$$b_j(t) = \max\{q \leqslant 1 : \; T_j(t) I(\hat{\mu}_j(t), q) \leqslant f(t)\}$$

where $f(t) = \log(t) + 3\log(\log(t))$ is the confidence level

**Interpretation** For optimal performance, an algorithm has to sample each suboptimal arm as many times as given by the lower bound

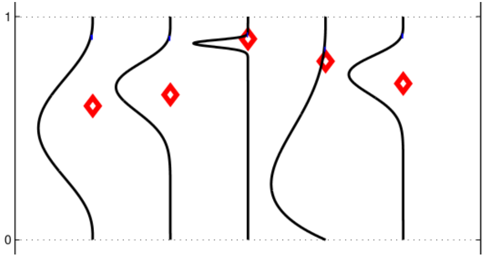
$b_j(t)$ keeps track of how far from this *quota* arm $j$ has been sampled

Term $3\log(\log(t))$ accounts for uncertainty on $\hat{\mu}_j$ and optimal arm

# Lower bounds and optimal algorithms (cont.)

**Thompson sampling:** (Thompson, 1933)

- Much older than UCB, conceived for adaptive clinical trials
- Bayesian origin: Assume a uniform prior on $\mu_j$ for every $j$, and update the posterior $p_{\mu_j}$ based on samples up to round $t$
- At round $t$, sample $\hat{\mu}_j$ from posterior $p_{\mu_j}$, and pick arm for which $\hat{\mu}_j$ is largest




- Empirically shown that TS has better finite sample mean performance than UCB algorithms, but its variance can be higher
- Kaufmann, Korda & Munos (ALT, 2012) showed that Thompson Sampling is asymptotically optimal

# Application to $\mathcal{H}_\infty$-norm estimation

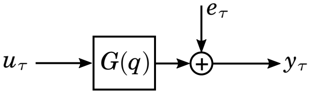
Applications of MAB to control problems are very sparse. Some examples:

- P.R. Kumar. "An adaptive controller inspired by recent results on learning from experts". In K.J. Åström, G.C. Goodwin & P.R. Kumar, *Adaptive Control, Filtering, and Signal Processing*, Springer, 1995
- M. Raginsky, A. Rakhlin, and S. Yüksel. "Online convex programming and regularization in adaptive control". *CDC*, 2010

**Our goal:** apply MAB theory to problems of iterative identification

# Application to $\mathcal{H}_\infty$-norm estimation (cont.)

**Setup**

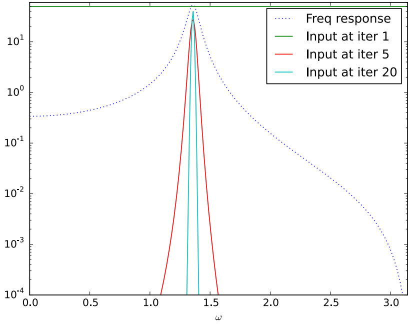



- Work with data batches, of length $N$, sufficiently spaced in time
- At each iteration $\tau$, an input batch $\boldsymbol{u}_\tau = (u_1, \ldots, u_N)$ is designed and applied to the system
- The output of the system, $\boldsymbol{y}_\tau = (y_1, \ldots, y_N)$, is collected
- **Goal** Determine the $\mathcal{H}_\infty$ norm of the system, as accurately as possible
- **Why** $\mathcal{H}_\infty$-norm is important for bounding model error (needed for robust control, *etc.*)

## Application to $\mathcal{H}_\infty$-norm estimation (cont.)

**Main Idea:**

Design $\boldsymbol{u}_\tau$ in frequency domain, considering each freq. $2\pi k/N$ as an arm!

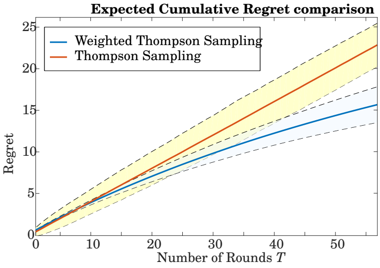


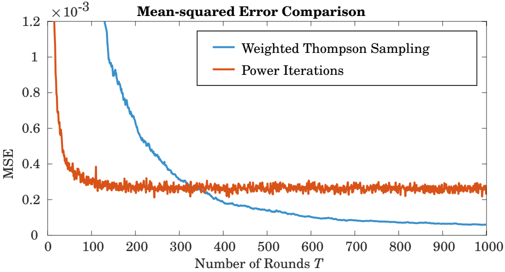

This is a standard MAB problem, except that:

- More than one arm can be pulled at once (in fact, we can choose a *distribution* over the arms!)
- The outcomes are complex-valued Gaussian distributed (variance inversely proportional to applied power)

# Application to $\mathcal{H}_\infty$-norm estimation (cont.)

- Derived a lower bound for the problem, which shows that choosing only one freq. is not more restrictive (asymptotically in $\tau$) than a continuous spectrum for $\boldsymbol{u}_\tau$
- Proposed a *weighted Thompson sampling* algorithm with better regret than standard TS
- Still... power iterations has better initial transient than MAB algorithms!




More information on Matias' poster!

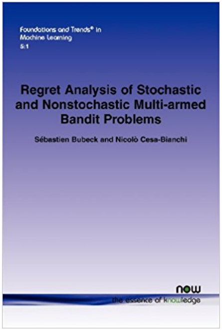
# Summary

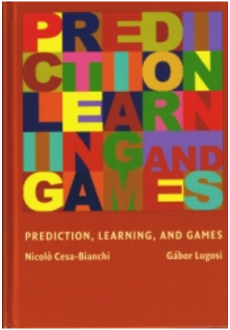
- MABs are a useful approach to adaptive control
- Standard theory applicable to some problems of iterative identification and control
- A relevant example: $\mathcal{H}_\infty$-norm estimation
- Control applications require non-trivial extensions to basic MAB framework:

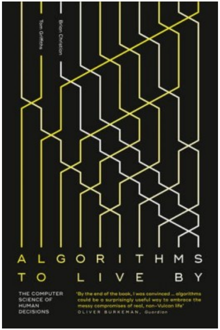
Interesting research directions!

# Some references

- S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. NOW, 2012
- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006
- B. Christian and T. Griffiths. *Algorithms to Live By: The Computer Science of Human Decisions*. William Collins, 2016



Thank you for your attention!