

# Mean-Field Control for Restless Bandits and Weakly Coupled MDPs

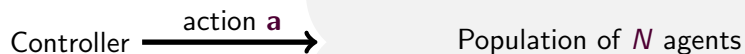
Nicolas Gast

joint work with Bruno Gaujal, Kimang Khun, Chen Yan

Inria

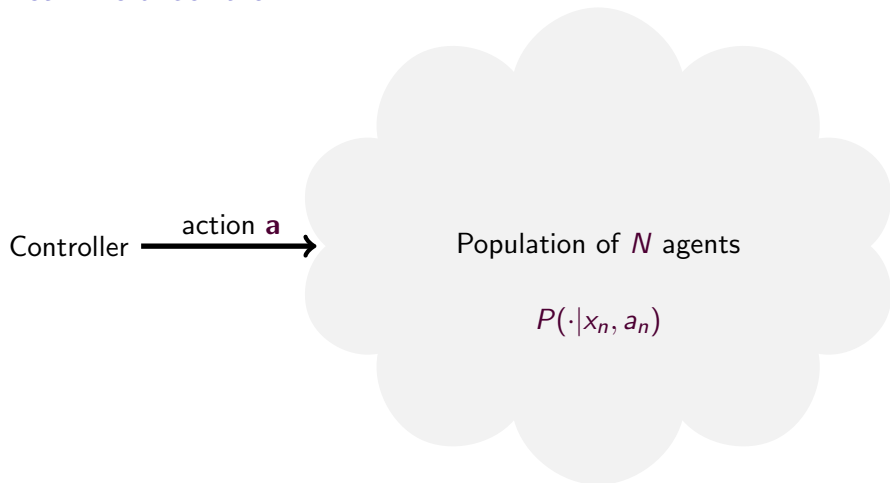
EDF R&D. September 27th, 2023

# Mean field control



$$P(\cdot | x_n, a_n)$$

# Mean field control



The computational difficulty increases with  $N$  but “ $N = \infty$ ” is easy.

- How to use the  $N = +\infty$  solution for finite  $N$ ?
- How efficient is this? (i.e., how fast does it become optimal?)

# This talk will focus on *Markovian bandits*

$N$  statistically identical **arms** (=agents)

- Discrete time, finite state space.
- $P(\cdot|s_n, a_n)$  and  $r(s_n, a_n)$ .

Maximize expected reward

$$\frac{1}{T} \sum_{t=1}^T \sum_{n=1}^N r(s_n(t), a_n(t)).$$

# This talk will focus on *Markovian bandits*

$N$  statistically identical **arms** (=agents)

- Discrete time, finite state space.
- $P(\cdot|s_n, a_n)$  and  $r(s_n, a_n)$ .

Maximize expected reward

$$\frac{1}{T} \sum_{t=1}^T \sum_{n=1}^N r(s_n(t), a_n(t)).$$

**Hard constraint:**  $\forall t : \sum_{n=1}^N a_n(t) \leq \alpha N.$

- If  $a_n(t) \in \{0, 1\}$ : Markovian bandit (**this talk**)
- If  $a_n(t) \in \{0, 1\}^d$ : Weakly coupled MDP.

## Example 1: Applicant screening problem



$N$  applicants.

Bayesian model of quality.

- $T$  rounds of interview.
- Each round: you interview up to  $\alpha N$  candidates.

Goal: Maximize the expected quality of selected candidates.

## Example 2: What to work on?

### Job Scheduling



# Outline

- 1 The (relaxed) mean-field control problem
- 2 Asymptotic optimality of priority policies
  - Finite-horizon problem
  - Infinite-horizon problems
- 3 Index policies and computation of Whittle indices
- 4 Conclusion



## The (relaxed) mean-field control problem

Original model: For all  $t$ ,  $\sum_{n=1}^N a_n(t) \leq \alpha N$ .  $\Rightarrow$  PSPACE-hard

## The (relaxed) mean-field control problem

**Relaxed** model: For all  $t$ ,  $\mathbb{E} \left[ \sum_{n=1}^N a_n(t) \right] \leq \alpha N. \Rightarrow$  Independence relaxation.

This can be solve by an LP.

# The (relaxed) mean-field control problem

**Relaxed** model: For all  $t$ ,  $\mathbb{E} \left[ \sum_{n=1}^N a_n(t) \right] \leq \alpha N$ .  $\Rightarrow$  Independence relaxation.

This can be solve by an LP.

- $x_s = \mathbf{P} [s_n = s]$  and  $y_{s,a} = \mathbf{P} [s_n = s, a_n = a]$ .

$$\max_{x \geq 0, y \geq 0} \sum_{s,a} r_{s,a} y_{s,a}$$

$$\text{s.t. } x_{s'} = \sum_s y_{s,a} P(s'|s, a)$$

$$x_s = \sum_a y_{s,a}$$

$$\sum_s x_s = 1.$$

$$\sum_s y_{s,1} = \alpha$$

relaxed budget constraint

# The (relaxed) mean-field control problem

**Relaxed** model: For all  $t$ ,  $\mathbb{E} \left[ \sum_{n=1}^N a_n(t) \right] \leq \alpha N$ .  $\Rightarrow$  Independence relaxation.

This can be solve by an LP.

- $x_s(t) = \mathbf{P}[s_n(t) = s]$  and  $y_{s,a}(t) = \mathbf{P}[s_n(t) = s, a_n(t) = a]$ .

$$\max_{x \geq 0, y \geq 0} \sum_{t=1}^T \sum_{s,a} r_{s,a} y_{s,a}(t)$$

$$\text{s.t. } x_{s'}(t+1) = \sum_s y_{s,a}(t) P(s'|s, a)$$

$$x_s(t) = \sum_a y_{s,a}(t)$$

$$\sum_s x_s = x_s(0).$$

$$\sum_s y_{s,1}(t) = \alpha(t)$$

relaxed budget constraint

## Going back to the original problem

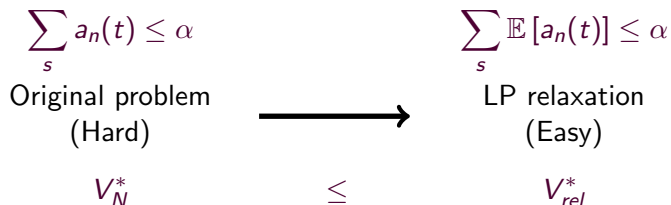
We can write the state of the system as:

$$X_s^N(t) = \frac{1}{N} \# \{ \text{Number of arms in state } i \text{ at time } t \}.$$

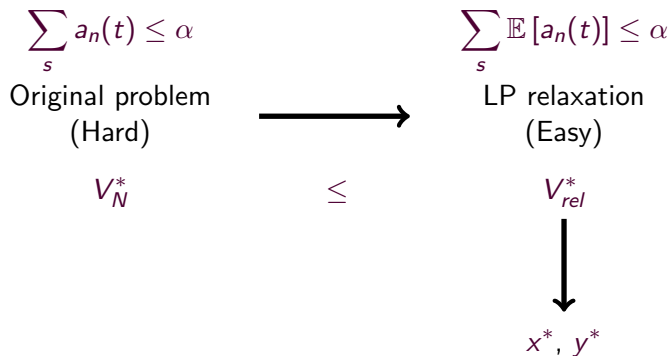
The control variable are:

$$Y_{s,a}^N(t) = \frac{1}{N} \# \{ \text{Number of arms in state } i \text{ for which action } a \text{ is taken} \}.$$

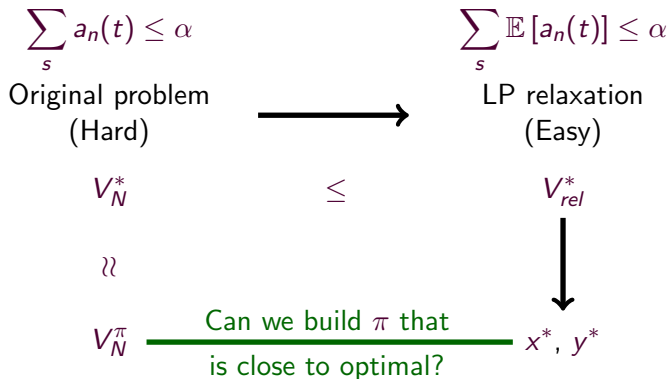
Can I apply this to the original (non-relaxed) problem?



Can I apply this to the original (non-relaxed) problem?



# Can I apply this to the original (non-relaxed) problem?



Main difficulty: in general  $\mathbf{X}^N(t) \neq \mathbf{x}^*(t)$ .

- We cannot choose  $\mathbf{Y}^N(t) = \mathbf{y}^*(t)$ .



## Some historical perspective

- Infinite horizon: Index policies (Gittins 60s, Whittle index (89), Nino-Mora, 90s-2000s)
  - ▶ Often asymptotically optimal. (Weber and Weiss 91).
  - 1. When they are: exponentially fast. (G, Gaujal, Yan 2021).
  - 2. We can compute index efficiently. (G, Gaujal, Khun 2022).
  - ▶ More recently: optimality without UGAP (but  $O(1/\sqrt{N})$ -suboptimality gap.)

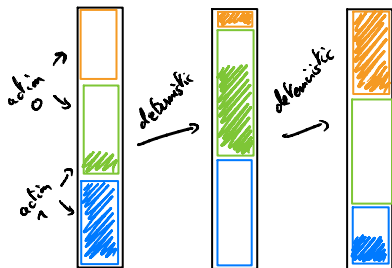
# Some historical perspective

- Infinite horizon: Index policies (Gittins 60s, Whittle index (89), Nino-Mora, 90s-2000s)
  - ▶ Often asymptotically optimal. (Weber and Weiss 91).
  - 1. When they are: exponentially fast. (G, Gaujal, Yan 2021).
  - 2. We can compute index efficiently. (G, Gaujal, Khun 2022).
  - ▶ More recently: optimality without UGAP (but  $O(1/\sqrt{N})$ -suboptimality gap.)
- Finite horizon: LP-index
  - ▶ Priority rule not always asymptotically optimal (Brown and Smith 2019), (Frazier et al 2020).
  - 3. When they are: exponentially fast (G, Gaujal, Yan 2022)

# Outline

- 1 The (relaxed) mean-field control problem
- 2 Asymptotic optimality of priority policies
  - Finite-horizon problem
  - Infinite-horizon problems
- 3 Index policies and computation of Whittle indices
- 4 Conclusion

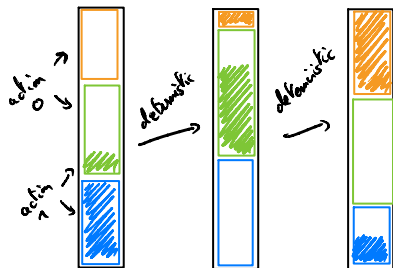
# From the LP solution to the finite- $N$ policies



Relaxed problem

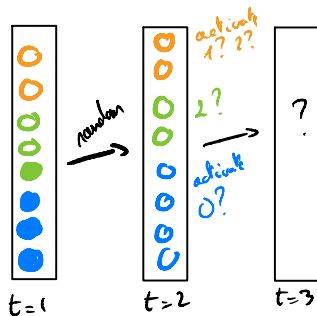
Optimal sequence  $x_s^*(t), y_{s,a}^*(t)$ .

# From the LP solution to the finite- $N$ policies



Relaxed problem

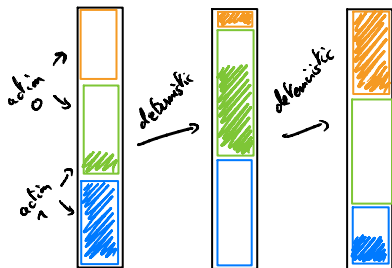
Optimal sequence  $x_s^*(t), y_{s,a}^*(t)$ .



Problem with  $N$  arms

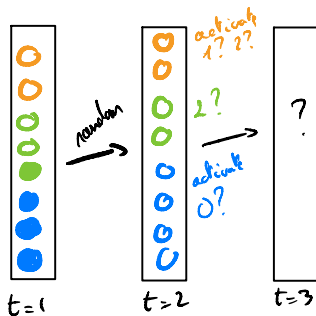
Need a policy.

# From the LP solution to the finite- $N$ policies



Relaxed problem

Optimal sequence  $x_s^*(t), y_{s,a}^*(t)$ .

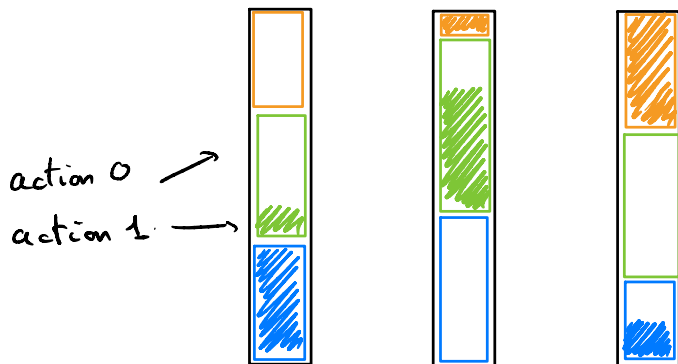


Problem with  $N$  arms

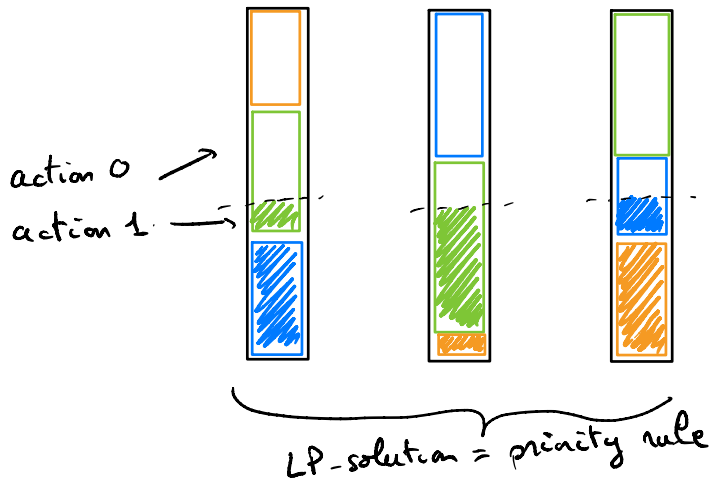
Need a policy.

- $\pi_t : X^N(t) \rightarrow Y^N(t)$  (this talk).
- Local policy (FtVA).

We want a policy such that  $\pi_t(x^*(t)) = y^*(t)$ .

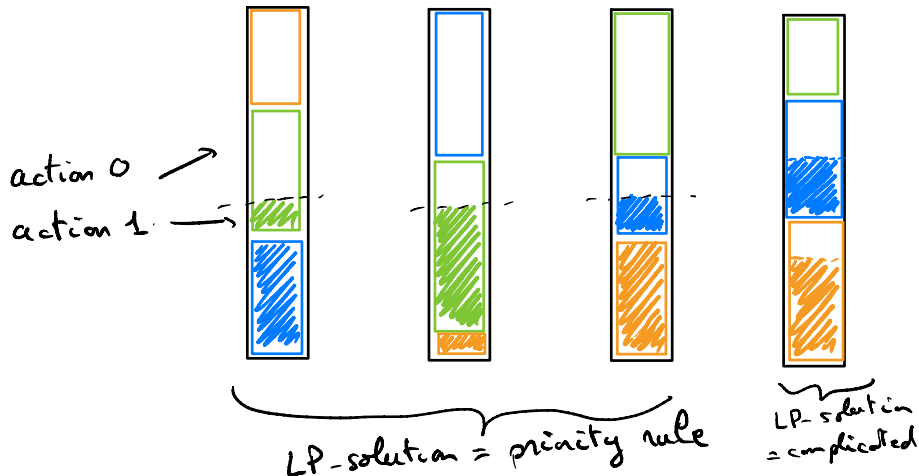


We want a policy such that  $\pi_t(x^*(t)) = y^*(t)$ .





We want a policy such that  $\pi_t(x^*(t)) = y^*(t)$ .



# Asymptotic optimality

Let  $S^0(t)$  be the number of states that are half-activated for the LP-solution at time  $t$ .

## Theorem

- There exists an *priority rule* that is *asymptotically optimal* if and only if for all  $t$ ,  $S^0(t) \leq 1$ .
- It becomes optimal *exponentially fast* if for all  $t$ ,  $S^0(t) = 1$ .

# Sketch of proof

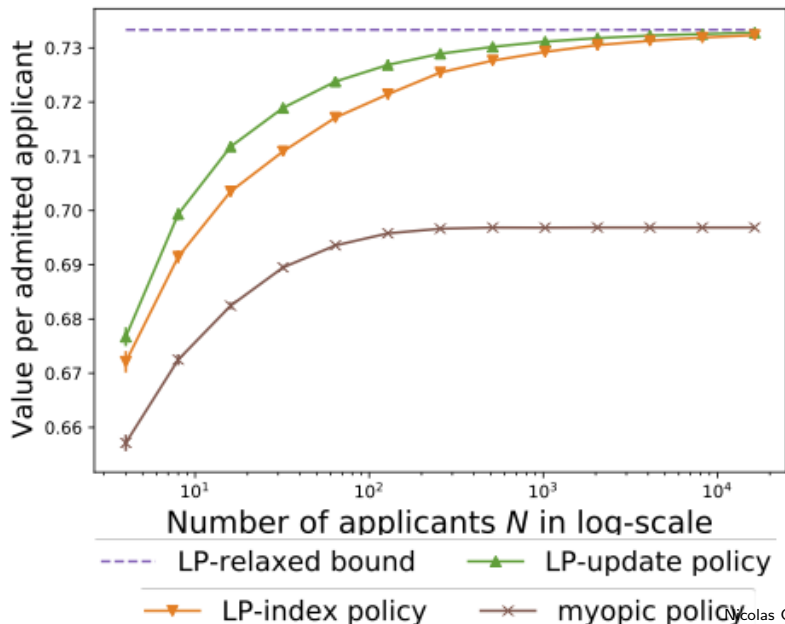
Recall that  $X_s^{(N)}(t) = \frac{1}{N} \# \{\text{arms in state } s \text{ at time } t\}$ .

$$X_{s'}^{(N)}(t+1) = \sum_{s,a} Y_{s,a}^{(N)}(t) \mathbf{P}[s'|s, a] + \underbrace{O(1/\sqrt{N})}_{\text{stochastic noise. CLT}}.$$

- ① Concentration argument:  $\pi$  continuous implies  $\lim_{N \rightarrow \infty} X_\pi^{(N)}(t) = x_\pi(t)$ .
  - ▶ This gives  $O(1/\sqrt{N})$  sub-optimality gap.
- ② A priority rule is *locally linear*. We use the linearity of expectation.
  - ▶ This gives  $e^{-\Omega(N)}$  sub-optimality gap.

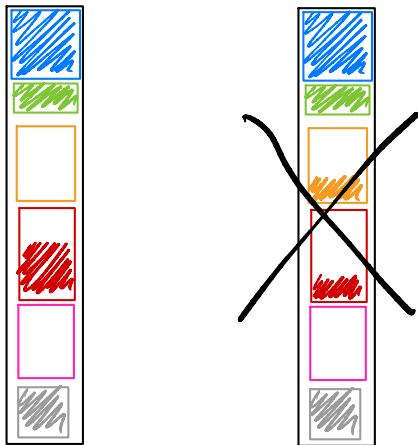
# Illustration of the asymptotic optimality

LP-index = policy from (Brown Smith 2020), LP-update = update LP solution.



For infinite-horizon problems, the LP-solution is a fixed point.

Theorem (LP folk). There exists a solution of the LP such that  $S^0 \leq 1$ .



- This defines priority policies (many because of tie-breaking).

# Are these priority policies asymptotic optimal?

For the infinite model, a priority policy defines a (piecewise linear) dynamical system:

$$\mathbf{x}(t+1) = \pi^{prio}(\mathbf{x}(t)).$$

## Theorem

- 1 If  $\pi^{prio}$  has a *unique attractor* (UGAP<sup>a</sup>), then it is *asymptotically optimal*. [Weber Weiss 90s, Verloop 2016]
- 2 For these problems, the suboptimality gap is exponentially small for *non-degenerate* problems. [G. Gaujal Yan 2021]

---

<sup>a</sup>See recent work of Hong, Xie, Chen, Wang for policies without UGAP

# Are these priority policies asymptotic optimal?

For the infinite model, a priority policy defines a (piecewise linear) dynamical system:

$$\mathbf{x}(t+1) = \pi^{\text{prio}}(\mathbf{x}(t)).$$

## Theorem

- 1 If  $\pi^{\text{prio}}$  has a *unique attractor* (UGAP<sup>a</sup>), then it is *asymptotically optimal*. [Weber Weiss 90s, Verloop 2016]
- 2 For these problems, the suboptimality gap is exponentially small for *non-degenerate* problems. [G. Gaujál Yan 2021]

---

<sup>a</sup>See recent work of Hong, Xie, Chen, Wang for policies without UGAP

## Proof.

- Concentration of measure.
- Local linearity + Stein's method.

Are all tie-breaking rules equivalent?

No.



# Outline

- 1 The (relaxed) mean-field control problem
- 2 Asymptotic optimality of priority policies
  - Finite-horizon problem
  - Infinite-horizon problems
- 3 Index policies and computation of Whittle indices
- 4 Conclusion

# Penalty and indexability

The  $N = \infty$  is a constraint MDP:

- $P(\cdot|s_n, a_n)$  and  $r(s_n, a_n)$  s.t. in steady-state,  $\mathbf{P}[a_n] = \alpha$ .

# Penalty and indexability

The  $N = \infty$  is a constraint MDP:

- $P(\cdot|s_n, a_n)$  and  $r(s_n, a_n)$  s.t. in steady-state,  $\mathbf{P}[a_n] = \alpha$ .

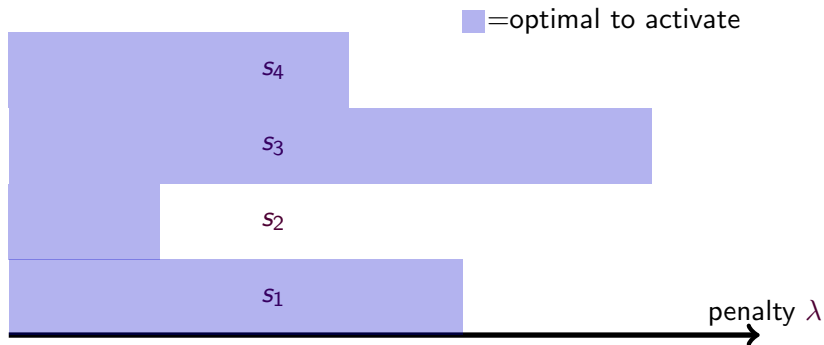
Idea: use a Lagrangian relaxation:

- $P(\cdot|s_n, a_n)$  and  $r(s_n, a_n) - \lambda a_n$ .

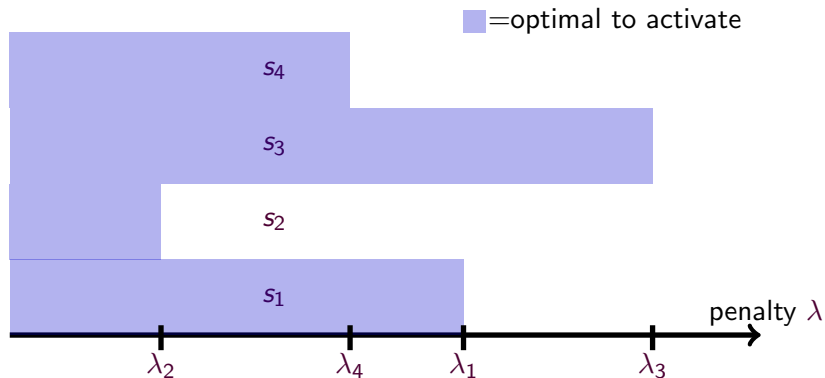


Penalty for activation

The penalty can be used to define a priority policy



The penalty can be used to define a priority policy



This is **Whittle index policy**.

For this example:  $s_3 \succ s_1 \succ s_4 \succ s_2$ .

# Definition of Whittle index

Intuitively, for each state, there exists a  $\lambda_s$  such that any optimal policy is such that:

- The optimal action in  $s$  is 0 (rest) if  $\lambda < \lambda_s$ ;
- The optimal action in  $s$  is 1 (activate) if  $\lambda > \lambda_s$ .

## Definition of Whittle index

Intuitively, for each state, there exists a  $\lambda_s$  such that any optimal policy is such that:

- The optimal action in  $s$  is 0 (rest) if  $\lambda < \lambda_s$ ;
- The optimal action in  $s$  is 1 (activate) if  $\lambda > \lambda_s$ .

This is **not always true**<sup>1</sup>.

If the model satisfies this assumption, we say that the model is indexable. Whittle index policy is the corresponding priority policy.

---

<sup>1</sup>True with high probability? Yes: (Nino-Mora 01), No (G, Gaujal, Khun 21).

# Illustration of what is Whittle policy

(stochastic scheduling)

Jobs of sizes  $X$  and  $Y$  with:

- $X = 10$
- $Y = \begin{cases} 2 & \text{proba } 1/2 \\ 18 & \text{proba } 1/2 \end{cases}$

Who should you run first to minimize expected completion time?



# Illustration of what is Whittle policy

(stochastic scheduling)

Jobs of sizes  $X$  and  $Y$  with:

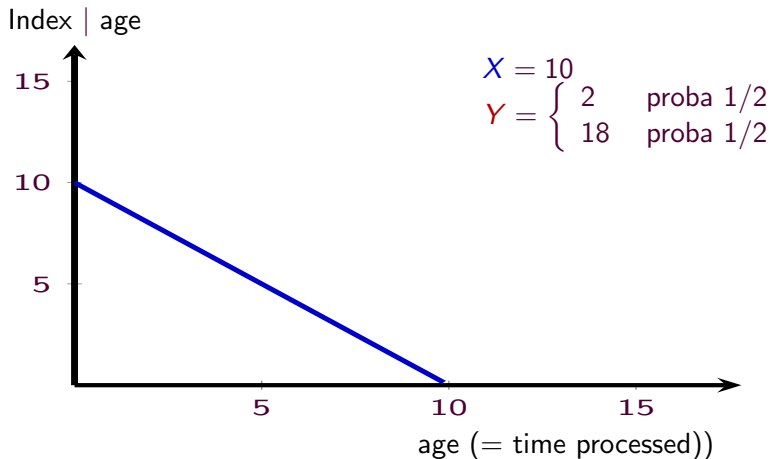
- $X = 10$
- $Y = \begin{cases} 2 & \text{proba } 1/2 \\ 18 & \text{proba } 1/2 \end{cases}$

Who should you run first to minimize expected completion time?

Running a job costs 1€/sec and you can stop anytime. If you finish the job, you earn  $x$ . **Whittle (=Gittins) index** is the smallest  $x$  so that you start running the job.

# Illustration of what is Whittle policy

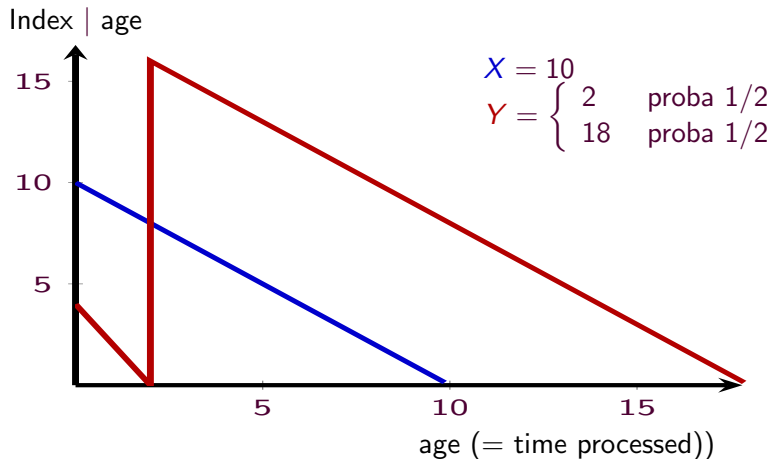
(stochastic scheduling)



Index can be computed independently for each job (=arm).

# Illustration of what is Whittle policy

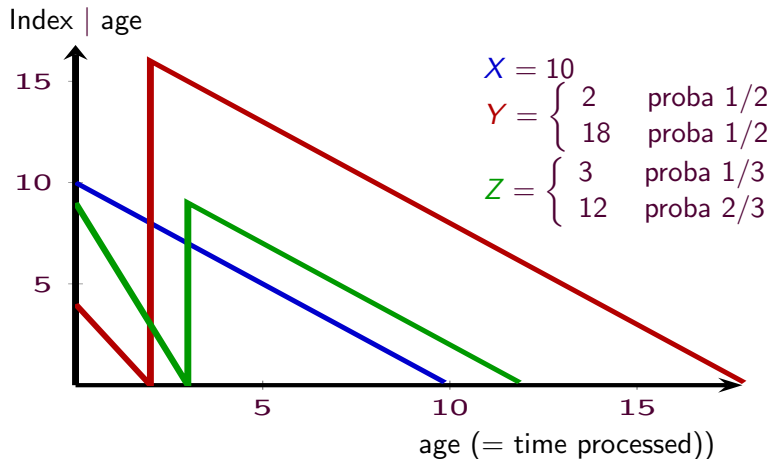
(stochastic scheduling)



Index can be computed independently for each job (=arm).

# Illustration of what is Whittle policy

(stochastic scheduling)



Index can be computed independently for each job (=arm).

# How to compute Whittle indices?

Classical definition:

- The index is the penalty  $\lambda_s$  such that that an optimal policy can choose to activate or not the state  $s$  when the penalty is  $\lambda_s$ .

# How to compute Whittle indices?

Refined definition:

- The index is the (**unique**) penalty  $\lambda_s$  such that that an (**Bellman-**)optimal policy can choose to activate or not the state  $s$  when the penalty is  $\lambda_s$ .

## How to compute Whittle indices?

A Bellman-optimal policies satisfies Bellman equations:

$$g^*(\lambda) + h_s^*(\lambda) = \max_a \underbrace{r(s, a) + a\lambda + \sum_j P(j|s, a)h_j^*(\lambda)}_{q_{s,a}(\lambda)}$$

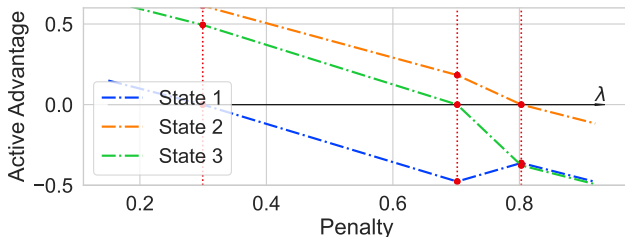
We define the *active advantage*  $b_s(\lambda) := q_{s,1}(\lambda) - q_{s,0}(\lambda)$ .

# How to compute Whittle indices?

A Bellman-optimal policy satisfies Bellman equations:

$$g^*(\lambda) + h_s^*(\lambda) = \max_a \underbrace{r(s, a) + a\lambda + \sum_j P(j|s, a)h_j^*(\lambda)}_{q_{s,a}(\lambda)}$$

We define the *active advantage*  $b_s(\lambda) := q_{s,1}(\lambda) - q_{s,0}(\lambda)$ .

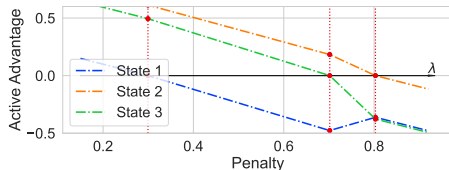


## Theorem (G,Gaujal,Khun, 22)

An arm is indexable if and only if for all  $s$ :  $b_{s,1}(\lambda) = 0$  has a unique solution.



We can use this characterization to build an efficient algorithm

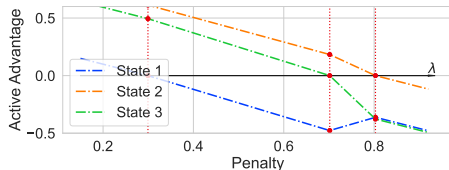


Three ingredients:

- 1 For MDP, the advantage function is piecewise linear:

$$b^{\pi}(\lambda) = (A^{\pi})^{-1}(r + \lambda\pi).$$

We can use this characterization to build an efficient algorithm



Three ingredients:

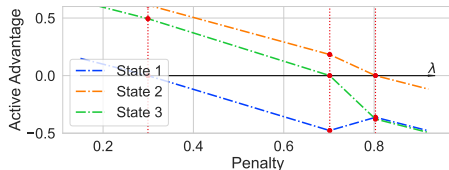
- 1 For MDP, the advantage function is piecewise linear:

$$b^\pi(\lambda) = (A^\pi)^{-1}(r + \lambda\pi).$$

- 2 Sherman-Morrison formula: Let  $A$  be an invertible matrix,  $u$  and  $v$  vectors  $1D$  such that  $1 + v^T A^{-1}u \neq 0$ . Then:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}.$$

We can use this characterization to build an efficient algorithm



Three ingredients:

- 1 For MDP, the advantage function is piecewise linear:

$$b^\pi(\lambda) = (A^\pi)^{-1}(r + \lambda\pi).$$

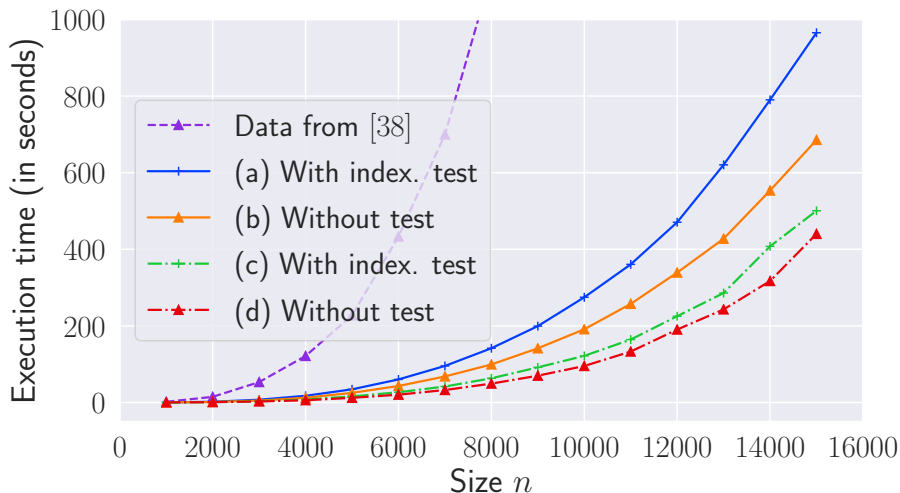
- 2 Sherman-Morrisson formula: Let  $A$  be an invertible matrix,  $u$  and  $v$  vectors  $1D$  such that  $1 + v^T A^{-1}u \neq 0$ . Then:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}.$$

- 3 We can reorder operations to use Strassen's like operations.

We obtain a theoretical complexity of  $O(S^{2.53})$  and an efficient implementation

<https://pypi.org/project/markovianbandit-pkg/>



# Outline

- 1 The (relaxed) mean-field control problem
- 2 Asymptotic optimality of priority policies
  - Finite-horizon problem
  - Infinite-horizon problems
- 3 Index policies and computation of Whittle indices
- 4 Conclusion

# Conclusion

For Markovian bandits, mean-field control can be solved by an LP.

- Can be generalized to weakly coupled MDPs.

Simple policies (priority rule) are not always optimal.

- When they are, they become optimal exponentially fast.
- Index policy (= “right activation price”) are very efficient.

# Conclusion

For Markovian bandits, mean-field control can be solved by an LP.

- Can be generalized to weakly coupled MDPs.

Simple policies (priority rule) are not always optimal.

- When they are, they become optimal exponentially fast.
  - Index policy (= “right activation price”) are very efficient.
- This talk: finite-state space, computation of policies.
  - Open questions: learning, continuous state-spaces.

<http://polaris.imag.fr/nicolas.gast/>

- *Computing Whittle (and Gittins) Index in Subcubic Time*, G. Gajal, Khun <https://arxiv.org/abs/2203.05207>
- *LP-based policies for restless bandits: necessary and sufficient conditions for (exponentially fast) asymptotic optimality*. G. Gajal Yan. <https://arxiv.org/abs/2106.10067>