Mean-Field Control for Restless Bandits and Weakly Coupled MDPs

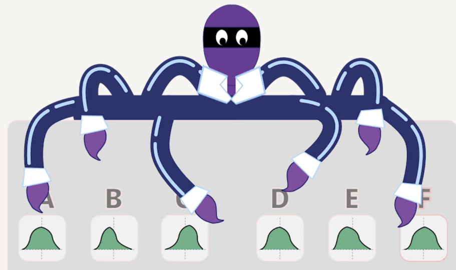
Nicolas Gast

joint work with Bruno Gaujal, Kimang Khun, Chen Yan

Inria

CNI Seminar series, May 2nd, 2023

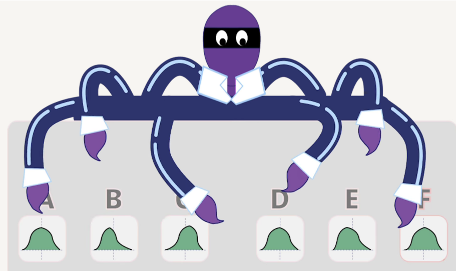
The Markovian bandit problem



Classical bandit problem:

- N arms
- I.i.d. unknown reward
- Goal: identify the best

The Markovian bandit problem



Classical bandit problem:

- N arms
- I.i.d. unknown reward
- Goal: identify the best

Markovian bandit:

- N *statistically identical* arms.
- Each arm has a state: you know $P(\cdot|s_n, a_n)$ and $r(s_n, a_n)$.
- Goal: compute a policy $\pi : \mathcal{S}^N \to \mathcal{A}^N$.

Example 1: Applicant screening problem

N applicants, T rounds of interview.
Each round: you can interview up to αN candidates.
Goal: maximize the expected quality of selected candidates.

Example 1: Applicant screening problem

N applicants, T rounds of interview.
Each round: you can interview up to αN candidates.
Goal: maximize the expected quality of selected candidates.

Each candidate has an (unknown) quality q_n.

- Result of an interview: Bernoulli(q_n)

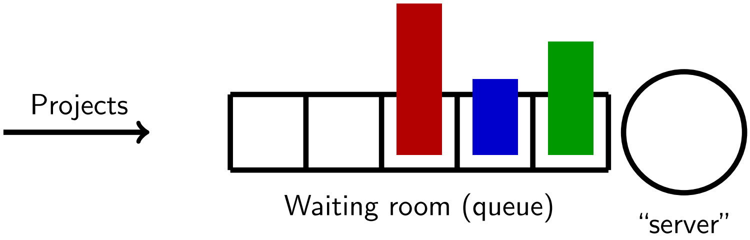
Goal: find the βN highest q_n.

Example 2: What to work on?

Job Scheduling

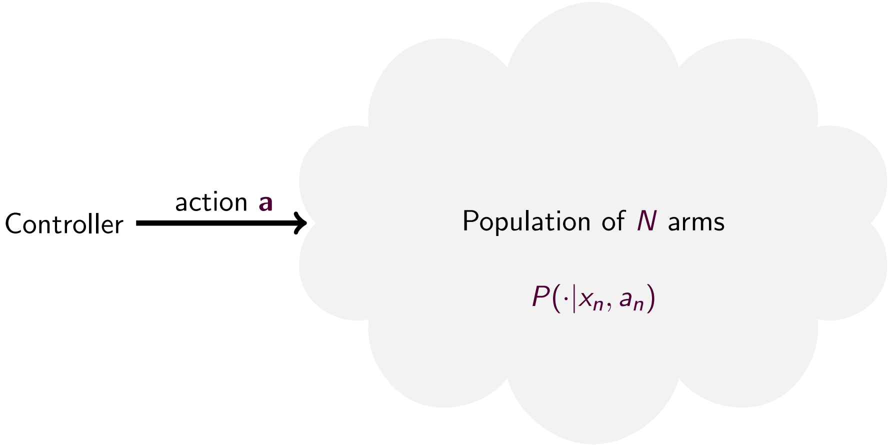
Example 2: What to work on?

Job Scheduling

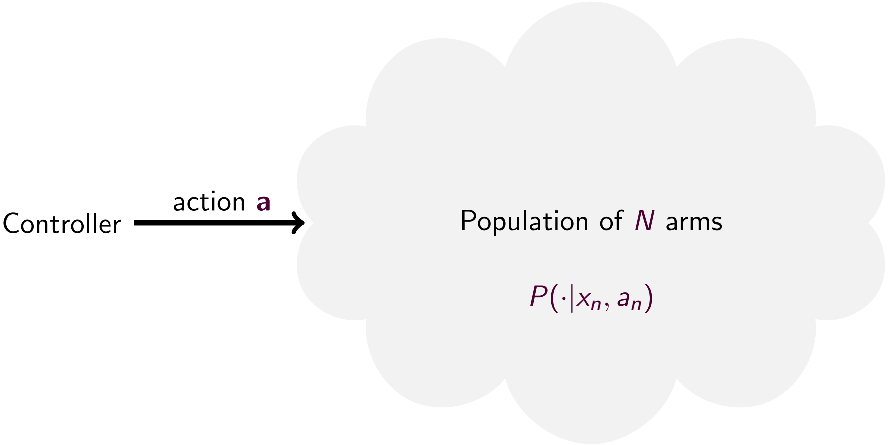


- Examples: research projects, tasks allocations, electric vehicle charging, wireless scheduling,...
- Heuristics: SRPT, EDF,...

We use tools from mean field control



We use tools from 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?)

Outline

1. The mean-field control problem
2. Infinite-horizon and index policies
3. Asymptotic optimality and index computation
4. Finite-horizon restless bandits
5. Conclusion

Original model for finite N

N statistically identical arms

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

Maximize expected reward

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

Original model for finite N

N statistically identical arms

- 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: $\forall t : \sum_{n=1}^{N} a_n(t) \leq C.$

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

The mean-field control problem

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

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

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

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

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

$$\sum_{s} x_s = 1.$$

$$\sum_{s} y_{s,1} = \alpha \qquad \text{relaxed budget contraint}$$

The 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.

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

$$
\begin{aligned}
\max_{x\geq 0, y\geq 0} \quad & \sum_{t=1}^{T} \sum_{s,a} r_{s,a} y_{s,a}(t) \\
\text{s.t.} \quad & x_{s'}(t+1) = \sum_{s} y_{s,a}(t) P(s'|s,a) \\
& x_s(t) = \sum_{a} y_{s,a}(t) \\
& \sum_{s} x_s = x_s(0). \\
& \sum_{s} y_{s,1}(t) = \alpha(t) && \text{relaxed budget contraint}
\end{aligned}
$$

Can I apply this to $N < \infty$?

$$\sum_s a_n(t) \leq \alpha$$

Original problem
(Hard)

$$\longrightarrow$$

$$\sum_s \mathbb{E}\left[a_n(t)\right] \leq \alpha$$

LP relaxation
(Easy)

$$V_N^* \qquad \leq \qquad V_{rel}^*$$

Can I apply this to $N < \infty$?

$$\sum_s a_n(t) \leq \alpha$$

Original problem
(Hard)

$$\sum_s \mathbb{E}\left[a_n(t)\right] \leq \alpha$$

LP relaxation
(Easy)

$$V_N^* \quad \leq \quad V_{rel}^*$$

$$V_{rel}^* \longrightarrow x^*, \; y^*$$

Can I apply this to $N < \infty$?

$$\sum_s a_n(t) \le \alpha \qquad\qquad \sum_s \mathbb{E}\left[a_n(t)\right] \le \alpha$$

Original problem (Hard) $\longrightarrow$ LP relaxation (Easy)

$$V_N^* \quad \le \quad V_{rel}^*$$

$$\wr\wr \qquad\qquad \big\downarrow$$

V_N^π — Can we build π that is close to optimal? — x^*, y^*

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).

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).

- 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

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

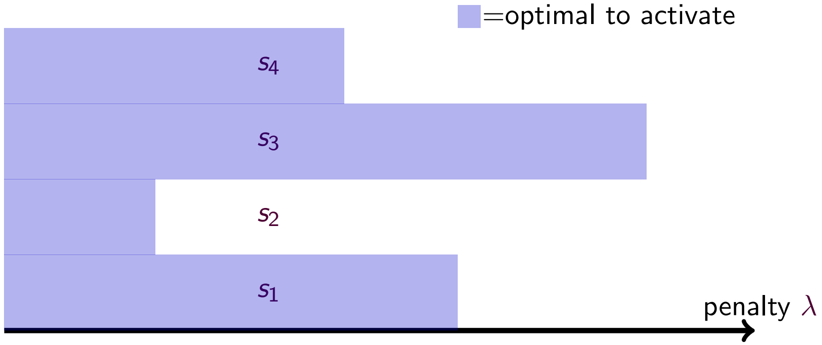
The $N = \infty$ is a constraint MDP:

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

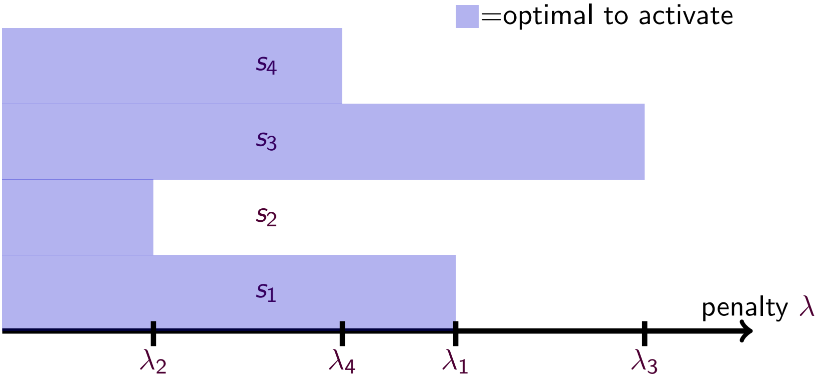
Idea: use a Lagrangian relaxation:

- $P(\cdot|s_n, a_n)$ and $r(s_n, a_n) - \underbrace{\lambda}_{\text{Penalty for activation}} a_n$.

The penaly can be used to define a priority policy



The penaly 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 λ_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 λ_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[1].

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

[1] 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)

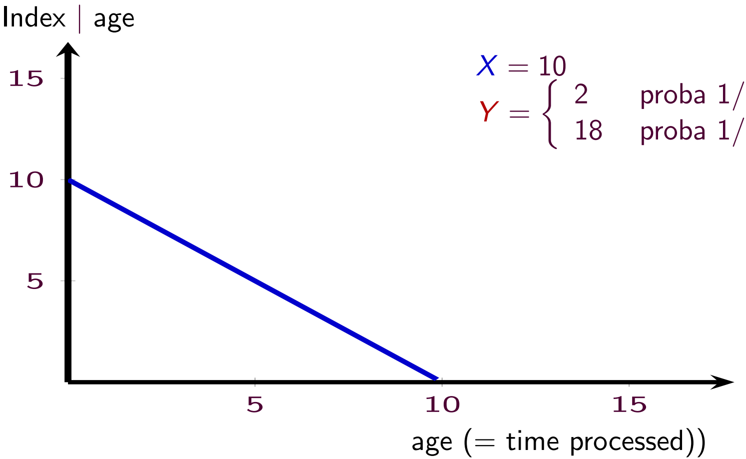
Jobs of sizes X and Y with:

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

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)



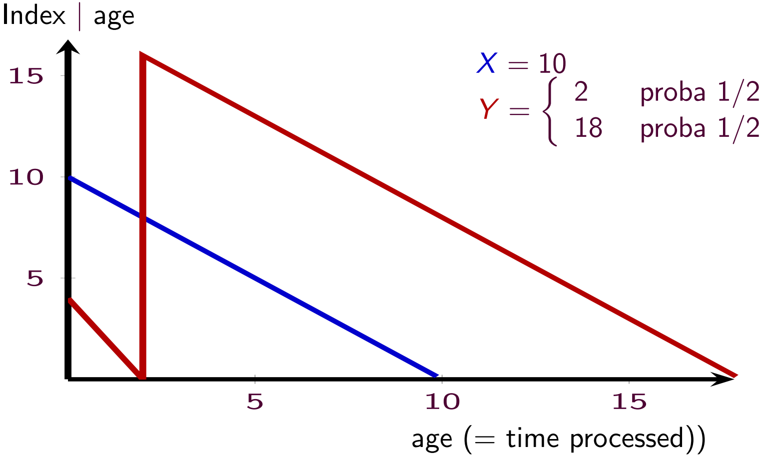
$X = 10$

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

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

Illustration of what is Whittle policy

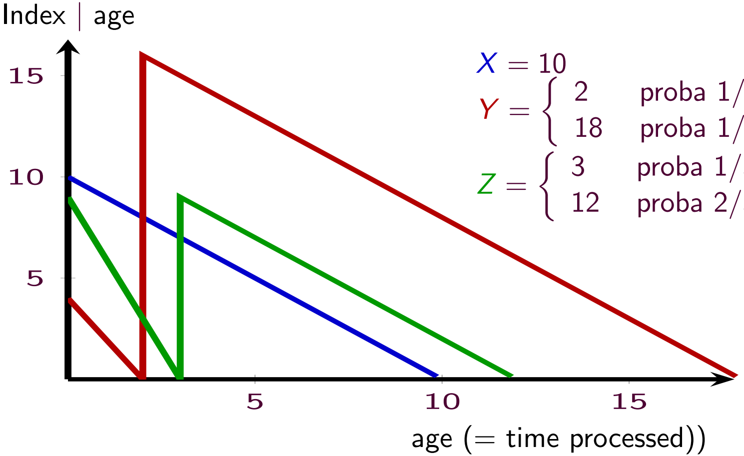
(stochastic scheduling)



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

Illustration of what is Whittle policy

(stochastic scheduling)



$X = 10$

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

$Z = \begin{cases} 3 & \text{proba } 1/3 \\ 12 & \text{proba } 2/3 \end{cases}$

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

Outline

Are Whittle index asymptotic optimal?

Assume indexability. For the infinite model, π^{WIP} defines a (piecewise linear) dynamical system:

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

Theorem

1. *If π^{WIP} has a unique attractor, then WIP 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]*

Sketch of proof

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

We have:

$$\mathbf{X}^{(N)}(t+1) = \pi^{WIP}(\mathbf{X}^{(N)}(t)) + \underbrace{O(1/\sqrt{N})}_{\text{stochastic noise. CLT}} .$$

Sketch of proof

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

We have:

$$\mathbf{X}^{(N)}(t+1) = \pi^{WIP}(\mathbf{X}^{(N)}(t)) + \underbrace{O(1/\sqrt{N})}_{\text{stochastic noise. CLT}}.$$

Hence:

1. If π^{WIP} has a unique attractor x^*, then $\mathbf{X}^N(\infty)$ concentrates on x^* (Hoeffding bound / large deviation).
2. Non-degenerate $= \pi^{WIP}$ is locally linear around x^*. We use the linearity of expectation.

How to compute Whittle indices?

Classical definition:

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

How to compute Whittle indices?

Refined definition:

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

How to compute Whittle indices?

A Bellman-optimal policies satisfies Bellman equations:

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

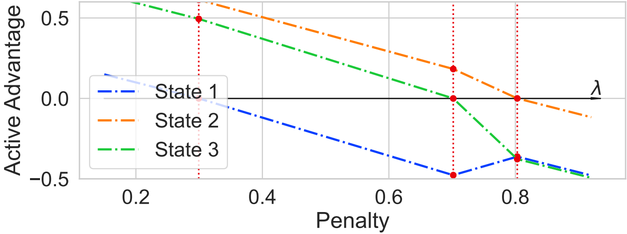
We define the *active advantage* $b_s(\lambda) := q_{s,1}(\lambda) - q_{s,0}(\lambda)$.

How to compute Whittle indices?

A Bellman-optimal policies satisfies Bellman equations:

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

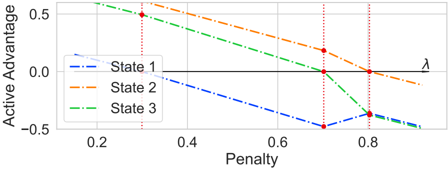
We define the *active advantage* $b_s(\lambda) := q_{s,1}(\lambda) - q_{s,0}(\lambda)$.



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

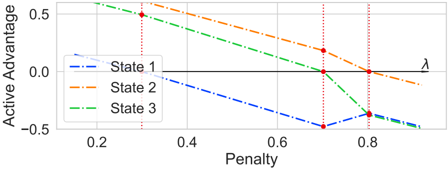


Three ingredients:

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:

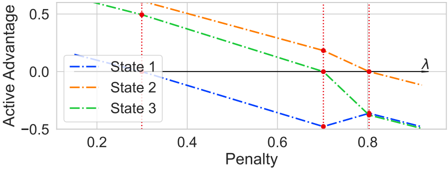
1. For MDP, the advantage function is piecewise linear:

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

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

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

We can use this characterization to build an efficient algorithm



Three ingredients:

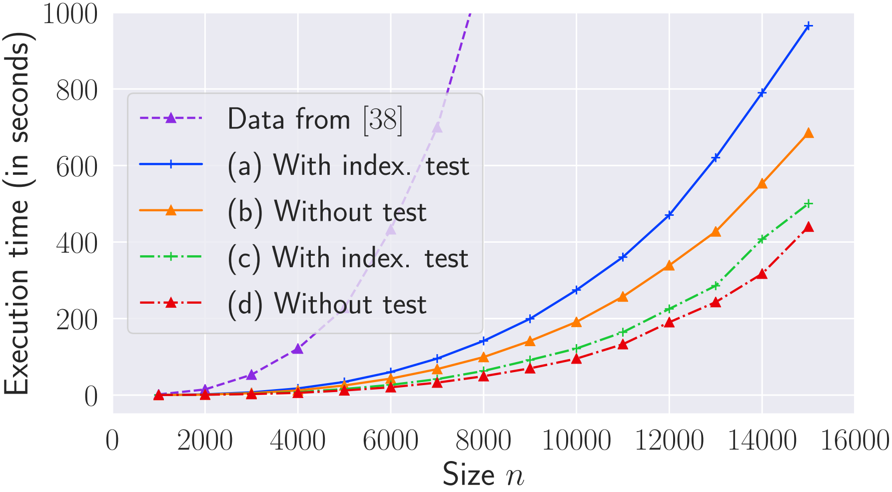
1. For MDP, the advantage function is piecewise linear:
$$b^{\pi}(\lambda) = (A^{\pi})^{-1}(r + \lambda\pi).$$

2. Sherman-Morisson formula: Let A be an invertible matrix, u and v vectors $1D$ such that $1 + v^T A^{-1} u \neq 0$. Then:
$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}.$$

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

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

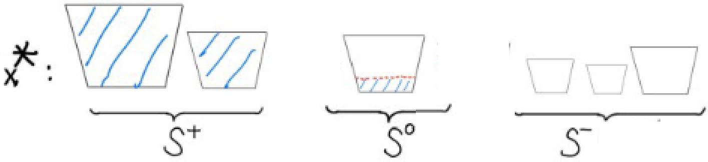
https://pypi.org/project/markovianbandit-pkg/



Outline

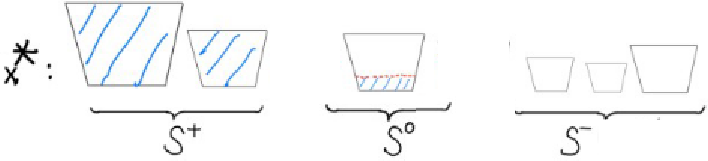
How to construct a policy for the original problem?

Relaxed problem: Optimal sequence $x^*_s(t)$, $y^*_{s,a}(t)$.

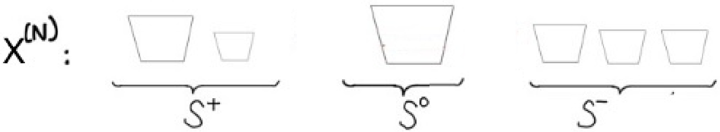


How to construct a policy for the original problem?

Relaxed problem: Optimal sequence $x^*_s(t)$, $y^*_{s,a}(t)$.



Original problem: Sequence $\pi_t : \mathcal{X} \to \mathcal{Y}$ such that $\pi(x^*) = y^*$.



You can implement π_t as a priority rule iff $|calS^0(t)| = 1$,

- It is locally linear.

Asymptotic optimality

Theorem

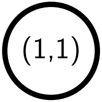
- *There exists an priority rule that is asymptotically optimal if and only if for all* t, $|S^0(t)| \leq 1$.
- *It becomes optimal exponentially fast if for all* t, $|S^0(t)| = 1$.

Proof ingredients.

1. Concentration argument: π continuous implies $\lim_{N\to\infty} X_\pi^{(N)}(t) = x_\pi(t)$.
2. Linearity of expectation.

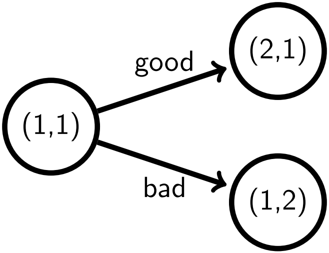
Many finite-horizon problems do not admit asymptotically optimal priority rules Example: Applicant screening problem (Brown Smith 2020)

Candidates with prior quality Beta(1,1), Interview budget $\alpha=0.25$



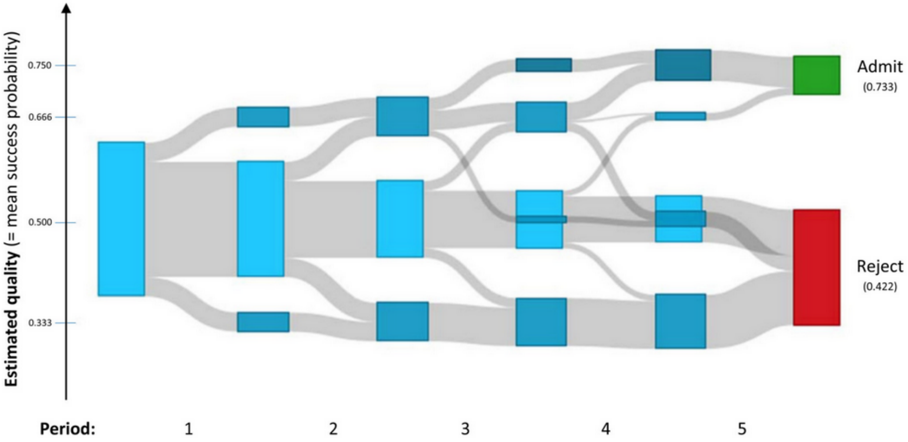
Many finite-horizon problems do not admit asymptotically optimal priority rules Example: Applicant screening problem (Brown Smith 2020)

Candidates with prior quality Beta(1,1), Interview budget $\alpha = 0.25$



Many finite-horizon problems do not admit asymptotically optimal priority rules Example: Applicant screening problem (Brown Smith 2020)

Candidates with prior quality Beta(1,1), Interview budget α=0.25



No asymptotically optimal priority policy: after two interviews:

- $x^*_{(1,1)} = 2x^*_{(2,1)} = 2x^*_{(1,2)} = 0.5$.
- $y^*_{(2,1),interview} = y^*_{(1,1),interview} = 0.125$.

Outline

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 actication 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 actication 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. Gaujal, Khun https://arxiv.org/abs/2203.05207
- *LP-based policies for restless bandits: necessary and sufficient conditions for (exponentially fast) asymptotic optimality*. G. Gaujal Yan. https://arxiv.org/abs/2106.10067