What is optimality is time-average MDPs?

Nicolas Gast, Bruno Gaujal, Kimang Khun

Inria

Workshop Mama 2023, June 19, 2023

Motivation: Whittle index

"A problem is indexable if and only if the optimal policy $\pi^*(\lambda)$ is non-decreasing".

Motivation: Whittle index

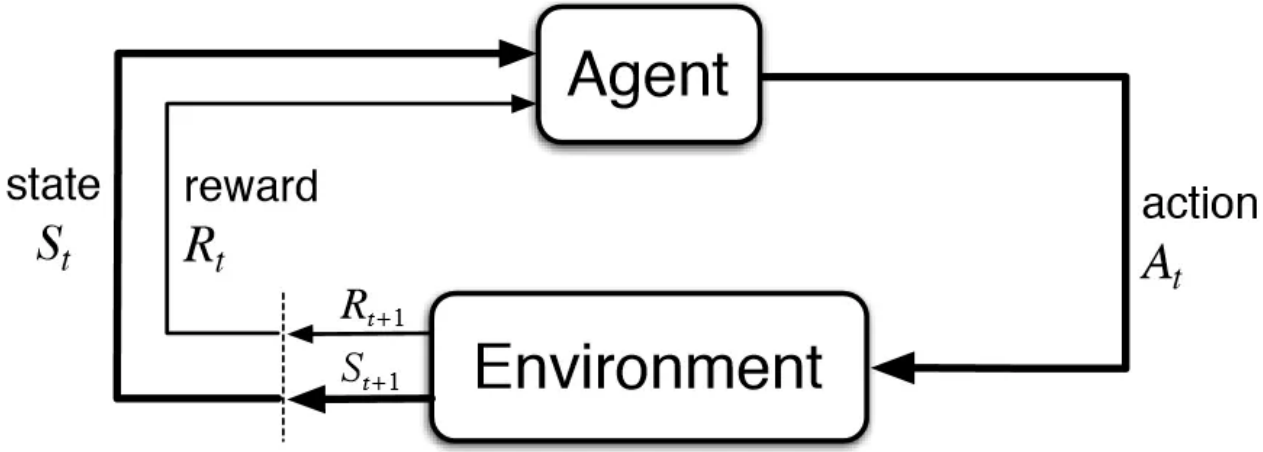
"A problem is indexable if and only if the optimal policy $\pi^*(\lambda)$ is non-decreasing".

Ambiguous definition:

1. What is "optimal"?
2. "The" optimal: is it unique?

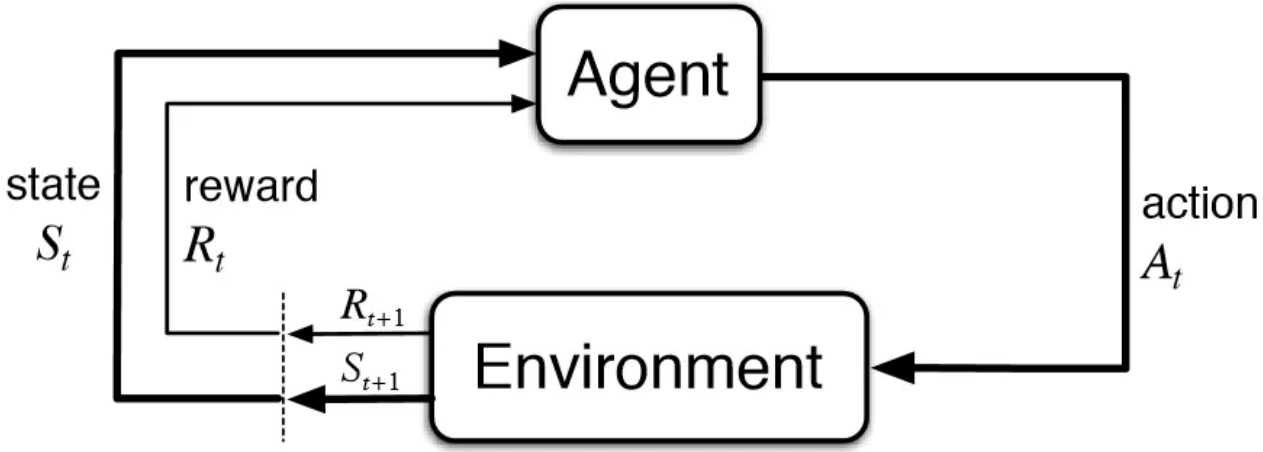
Markov deicison processes (MDPs)

Markov decision processes = Markov chains + actions and rewards.



Markov deicison processes (MDPs)

Markov decision processes = Markov chains + actions and rewards.



- Introduced in the 50s (Bellman)
- Very popular today because of reinforcement learning

Find a policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximize some *optimality criterion*.

What are optimality criteria?

1. Finite horizon: $\max \mathbb{E}\left[R_0 + R_1 + \cdots + R_T\right]$.
2. Discounted: $\max \mathbb{E}\left[R_0 + \gamma R_1 + \gamma^2 R_2 + \ldots\right]$ for some $\gamma < 1$.
3. Time-average (a.k.a. gain): $\max \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[R_0 + R_1 + \cdots + R_T\right]$.

What are optimality criteria?

1. Finite horizon: $\max \mathbb{E}\left[R_0 + R_1 + \cdots + R_T\right]$.
2. Discounted: $\max \mathbb{E}\left[R_0 + \gamma R_1 + \gamma^2 R_2 + \ldots\right]$ for some $\gamma < 1$.
3. Time-average (a.k.a. gain): $\max \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[R_0 + R_1 + \cdots + R_T\right]$.

What is the "good" notion of optimality for time-average MDPs?

- Time-average makes sense for queueing applications.
- Algorithms are designed for finite/discounted.
- Strong connections between the three notions

Outline

Definition of a MDP

At time t, you

- Observe $S_t \in \mathcal{S}$ and take an action $A_t \in \mathcal{A}$ (we assume $\mathcal{S}$, $\mathcal{A}$ finite).
- Receive $R_t = r(S_t, A_t)$.
- S_{t+1} jumps according to $P(\cdot|S_t, A_t)$.

A (deterministic) policy is a function $\pi : \mathcal{S} \to \mathcal{A}$.

Notion of gain-optimality

The gain of a policy is:

$$g^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r(S_t, A_t) \mid S_0 = s\right].$$

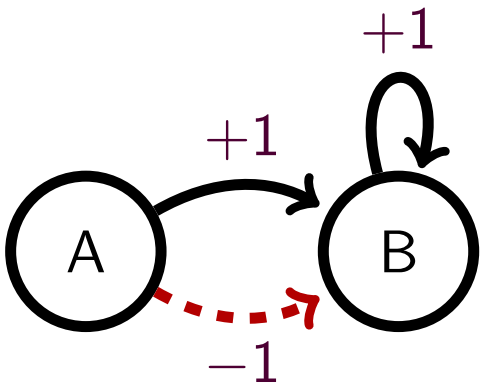
A gain-optimal policy[1] π^* maximizes the gain, *i.e.*, for all s:

$$g^{\pi^*}(s) = \arg\max_\pi g^\pi(s).$$

[1] Theorem (e.g., Puterman 2005). When $\mathcal{S}$ and $\mathcal{A}$ are finite, gain and optimal policies are well defined.

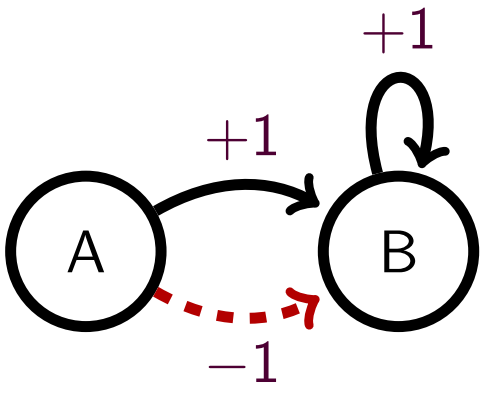
Is gain-optimal a good notion of optimality?

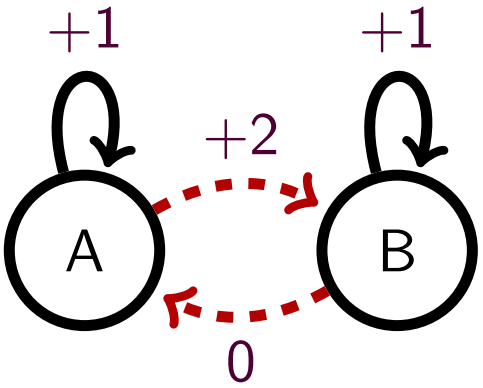
Example of deterministic MDPs



Is gain-optimal a good notion of optimality?

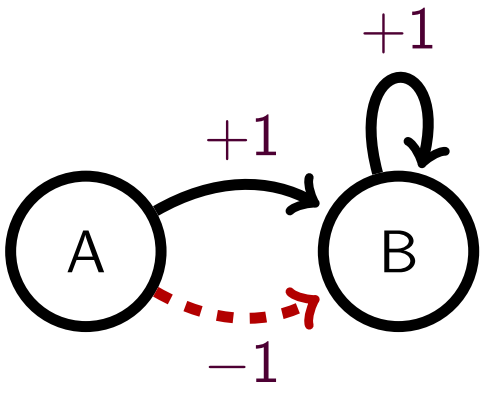
Example of deterministic MDPs

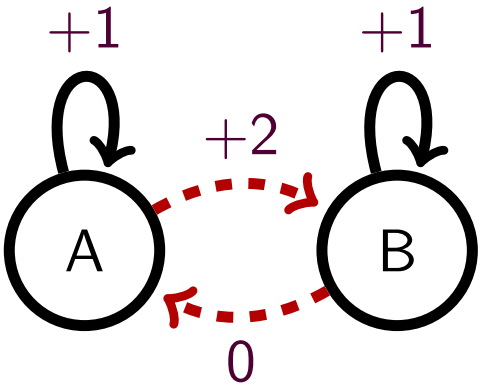




Is gain-optimal a good notion of optimality?

Example of deterministic MDPs





Here: all policies are gain-optimal but some look better than others.

- What is the right notion of optimality?
- What do algorithms compute?

Outline

Discounted optimality

For discounted problems, the quantity of importance is the value function:

$$V_\gamma^\pi(s) = \mathbb{E}\left[R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots \mid S_0 = s\right].$$

Discounted optimality

For discounted problems, the quantity of importance is the value function:

$$V_\gamma^\pi(s) = \mathbb{E}\left[R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots \mid S_0 = s\right].$$

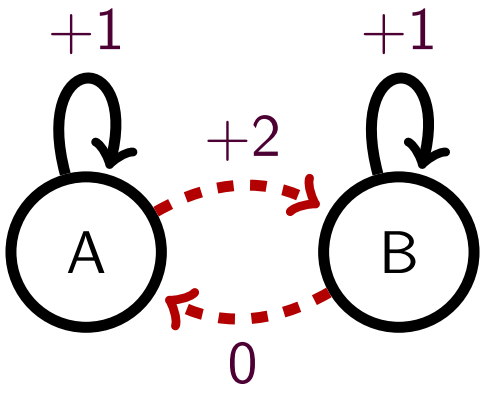
The optimal policy satisfies Bellman equation:

$$V_\gamma^*(s) = \max_a \underbrace{r(s,a) + \gamma \sum_{s'} V_\gamma^*(s') p(s'|s,a)}_{=:Q^*(s,a)}.$$

Bellman equation is very powerful:

- Unique solution for any $\gamma < 1$.
- Many algorithms to solve it (VI, PI, Q-learning).
- Can be used to define advantage $Q^*(s,a) - V^*(s)$ (e.g., actor-critic).

Discounted-optimality when $\gamma \to 1$



Policy π	discounted gain $V_\gamma^\pi(A)$	
Stay left	$1+\gamma+\gamma^2+\dots$	$= \dfrac{1}{1-\gamma}$
Go right:	$2+\gamma+\gamma^2+\gamma^3+\dots$	$= \dfrac{1}{1-\gamma}+1.$
Alternate:	$2+2\gamma^2+2\gamma^4\dots$	$= \dfrac{1}{1-\gamma}+\dfrac{1}{1+\gamma}.$

Best policy?

Bias-optimality

(Puterman 2015): For any policy π, there exists a bias vector $h^\pi(s)$ s.t.:

$$V_\gamma^\pi(s) = \frac{1}{1-\gamma}\underbrace{g^\pi(s)}_{\text{gain}} + \underbrace{h^\pi(s)}_{\text{bias}} + o(1-\gamma).$$

Bias-optimality

(Puterman 2015): For any policy π, there exists a bias vector $h^\pi(s)$ s.t.:

$$V_\gamma^\pi(s) = \frac{1}{1-\gamma}\underbrace{g^\pi(s)}_{\text{gain}} + \underbrace{h^\pi(s)}_{\text{bias}} + o(1-\gamma).$$

A policy is:

- Gain-optimal if it maximizes $g^\pi(s)$ (for all s).
- Bias-optimal if it maximizes $g^\pi(s)$ (for all s) and then $h^\pi(s)$.

Bias-optimality

(Puterman 2015): For any policy π, there exists a bias vector $h^\pi(s)$ s.t.:

$$V_\gamma^\pi(s) = \frac{1}{1-\gamma}\underbrace{g^\pi(s)}_{\text{gain}} + \underbrace{h^\pi(s)}_{\text{bias}} + o(1-\gamma).$$

A policy is:

- Gain-optimal if it maximizes $g^\pi(s)$ (for all s).
- Bias-optimal if it maximizes $g^\pi(s)$ (for all s) and then $h^\pi(s)$.

What do algorithms compute?

Gain-optimal policies? Bias-optimal policies?

Bias-optimality

(Puterman 2015): For any policy π, there exists a bias vector $h^\pi(s)$ s.t.:

$$V_\gamma^\pi(s) = \frac{1}{1-\gamma}\underbrace{g^\pi(s)}_{\text{gain}} + \underbrace{h^\pi{}_0(s)}_{\text{0-bias}} + (1-\gamma)\underbrace{h_1^\pi(s)}_{\text{1-bias}} + (1-\gamma)^2\underbrace{h_2^\pi(s)}_{\text{2-bias}} + \dots$$

A policy is:

- Gain-optimal if it maximizes $g^\pi(s)$ (for all s).
- Bias-optimal if it maximizes $g^\pi(s)$ (for all s) and then $h^\pi(s)$.
- (Blackwell-optimal if it maximizes $g^\pi(s)$ then $h_0^\pi(s)$ then $h_1^\pi(s)$,...)

What do algorithms compute?

Gain-optimal policies? Bias-optimal policies?

Outline

The (modified) Bellman equation

Let g^* be the optimal gain. There exists a bias vector h such that:

$$\begin{aligned} g^*(s) &= \max_a g^*(s')p(s'|s,a) \\ h(s) + g^*(s) &= \max_a r(s,a) + \sum_{s'} h(s')p(s'|s,a). \end{aligned} \tag{1}$$

Note: the solution is not unique (not just up to a constant).

The (modified) Bellman equation

Let g^* be the optimal gain. There exists a bias vector h such that:

$$\begin{aligned} g^*(s) &= \max_a g^*(s')p(s'|s,a) \\ h(s) + g^*(s) &= \max_a r(s,a) + \sum_{s'} h(s')p(s'|s,a). \end{aligned} \tag{1}$$

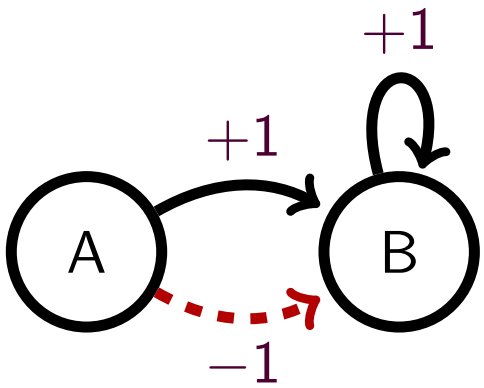
Note: the solution is not unique (not just up to a constant).

We call a best response to (1) a Bellman-optimal policy.

What is a Bellman optimal policy?

Bellman-optimal policies are not gain/bias optimal policies

gain-optimal $\supsetneq$ Bellman-optimal $\supsetneq$ bias-optimal

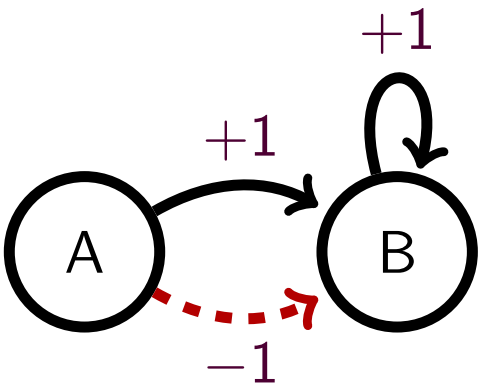


$h^*(A) = h^*(B)$

Black is Bellman-opt
All are Gain-optimal

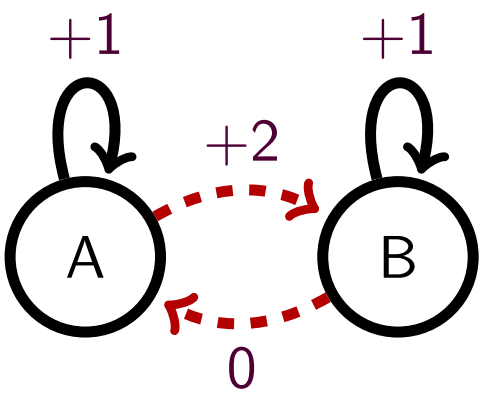
Bellman-optimal policies are not gain/bias optimal policies

gain-optimal $\supsetneq$ Bellman-optimal $\supsetneq$ bias-optimal



$h^*(A) = h^*(B)$

Black is Bellman-opt
All are Gain-optimal

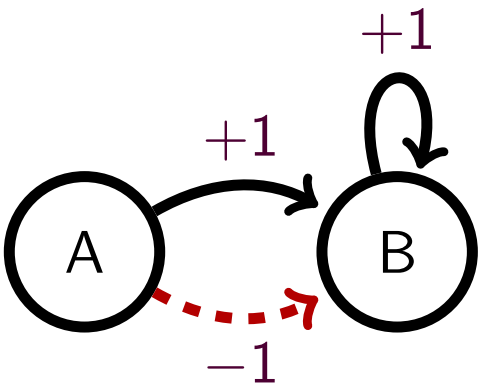


$h^*(A) = 1 + h^*(B)$

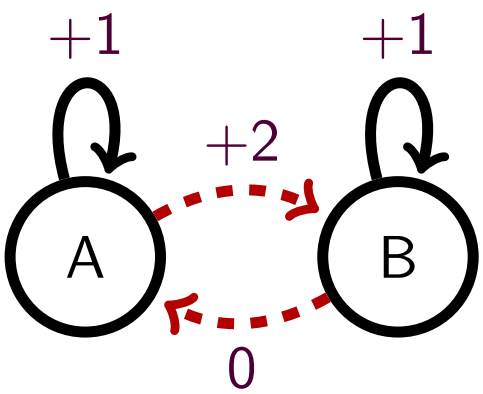
All policies are Bellman-optimal
There exists a unique bias-optimal policy

Bellman-optimal policies are not gain/bias optimal policies

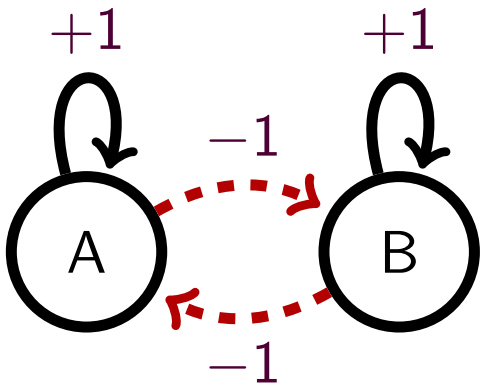
gain-optimal $\supsetneq$ Bellman-optimal $\supsetneq$ bias-optimal



$h^*(A) = h^*(B)$



$h^*(A) = 1 + h^*(B)$



$h^*(A) \in h^*(B) + [-1, 1]$

Black is Bellman-opt
All are Gain-optimal

All policies are Bellman-optimal
There exists a unique bias-optimal policy

Bellman optimal policy is a natural notion

Theorem

The set of optimal policies that can be output of policy iteration of value iteration is the set of Bellman-optimal policy.

Bellman optimal policy is a natural notion

Theorem

The set of optimal policies that can be output of policy iteration of value iteration is the set of Bellman-optimal policy.

A policy π is *canonical optimal* if there exists a final reward F such that π is optimal for for all finite horizon with final reward F, *i.e.* it maximmizes

$$\mathbb{E}\left[R_0 + R_1 + \cdots + R_T + F(S_T)\right].$$

Theorem (Yushkevich 1974)

- A policy is Bellman-optimal if and only if it is canonical optimal.

Does it matter: Definition of computation of Whittle index

Consider a two-action MDP, with a penalty:

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

Does it matter: Definition of computation of Whittle index

Consider a two-action MDP, with a penalty:

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

Classical definition of index

The Whittle index of s is a penalty λ_s such that that the optimal policy chooses $\pi(s) = 1$ when $\lambda < \lambda_s$ and $\pi(s) = 0$ when $\lambda > \lambda_s$.

Does it matter: Definition of computation of Whittle index

Consider a two-action MDP, with a penalty:

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

Non-ambiguous definition of index

The Whittle index of s is the unique penalty λ_s such that that any (Bellman-)optimal policy chooses $\pi(s) = 1$ when $\lambda < \lambda_s$ and $\pi(s) = 0$ when $\lambda > \lambda_s$.

How to compute Whittle indices?

A Bellman-optimal policies satisfies Bellman equations:

$$g_\lambda^*(s) + h_\lambda^*(s) = \max_a \underbrace{r(s,a) + a\lambda + \sum_{s'} P(s'|s,a) h_\lambda^*(s')}_{q_\lambda(s,a)}$$

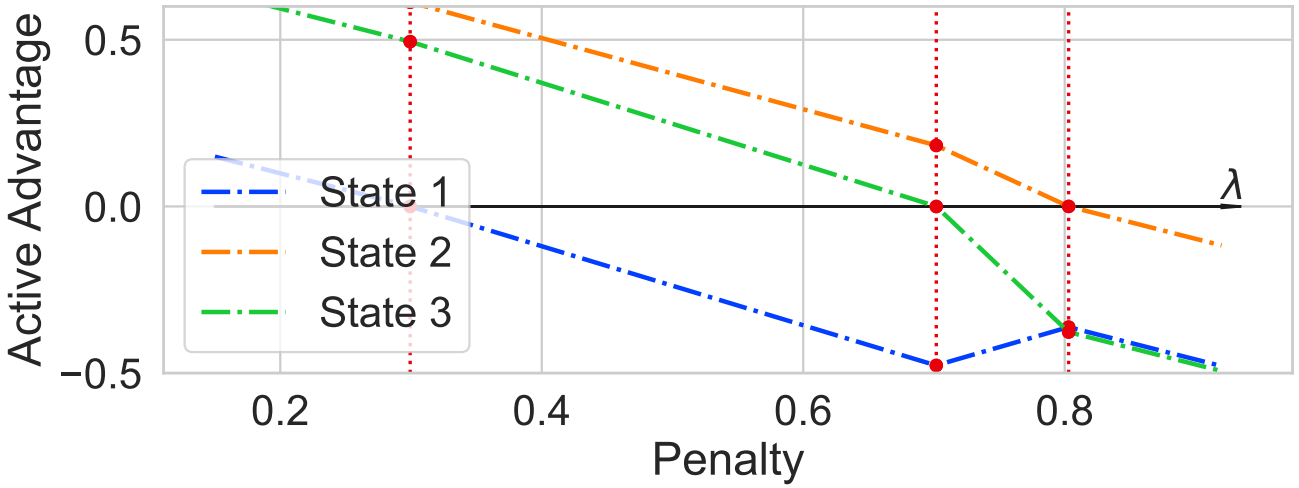
We define the *active advantage* $b_\lambda(s) := q_\lambda(s,1) - q_\lambda(s,0)$.

How to compute Whittle indices?

A Bellman-optimal policies satisfies Bellman equations:

$$g_\lambda^*(s) + h_\lambda^*(s) = \max_a \underbrace{r(s,a) + a\lambda + \sum_{s'} P(s'|s,a) h_\lambda^*(s')}_{q_\lambda(s,a)}$$

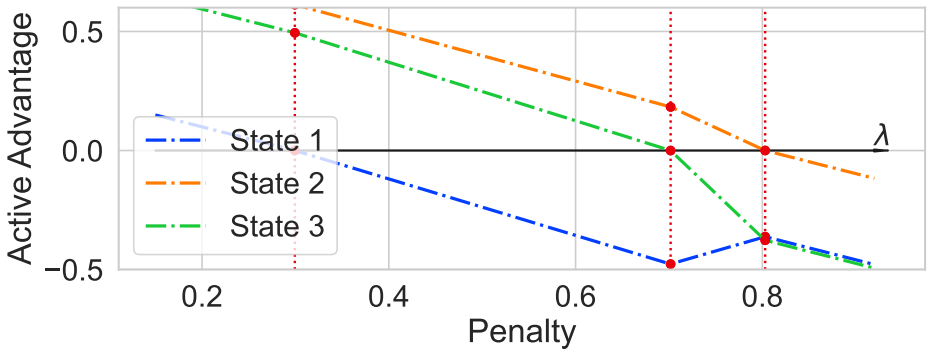
We define the *active advantage* $b_\lambda(s) := q_\lambda(s,1) - q_\lambda(s,0)$.



Theorem (G,Gaujal,Khun, 22)

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

We can use to build an subcubic algorithm

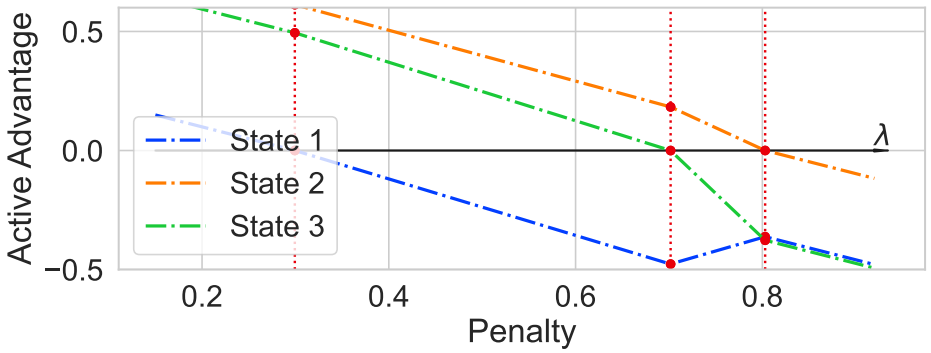


Three ingredients:

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

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

We can use to build an subcubic algorithm



Three ingredients:

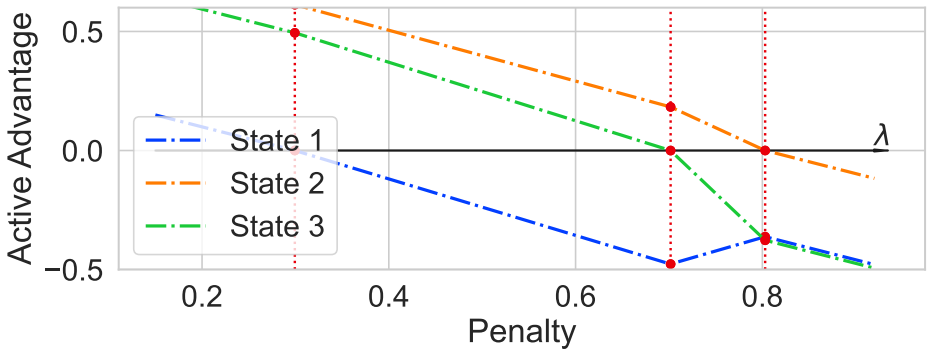
1. For MDP, the advantage function is piecewise linear:

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

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 to build an subcubic algorithm



Three ingredients:

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

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

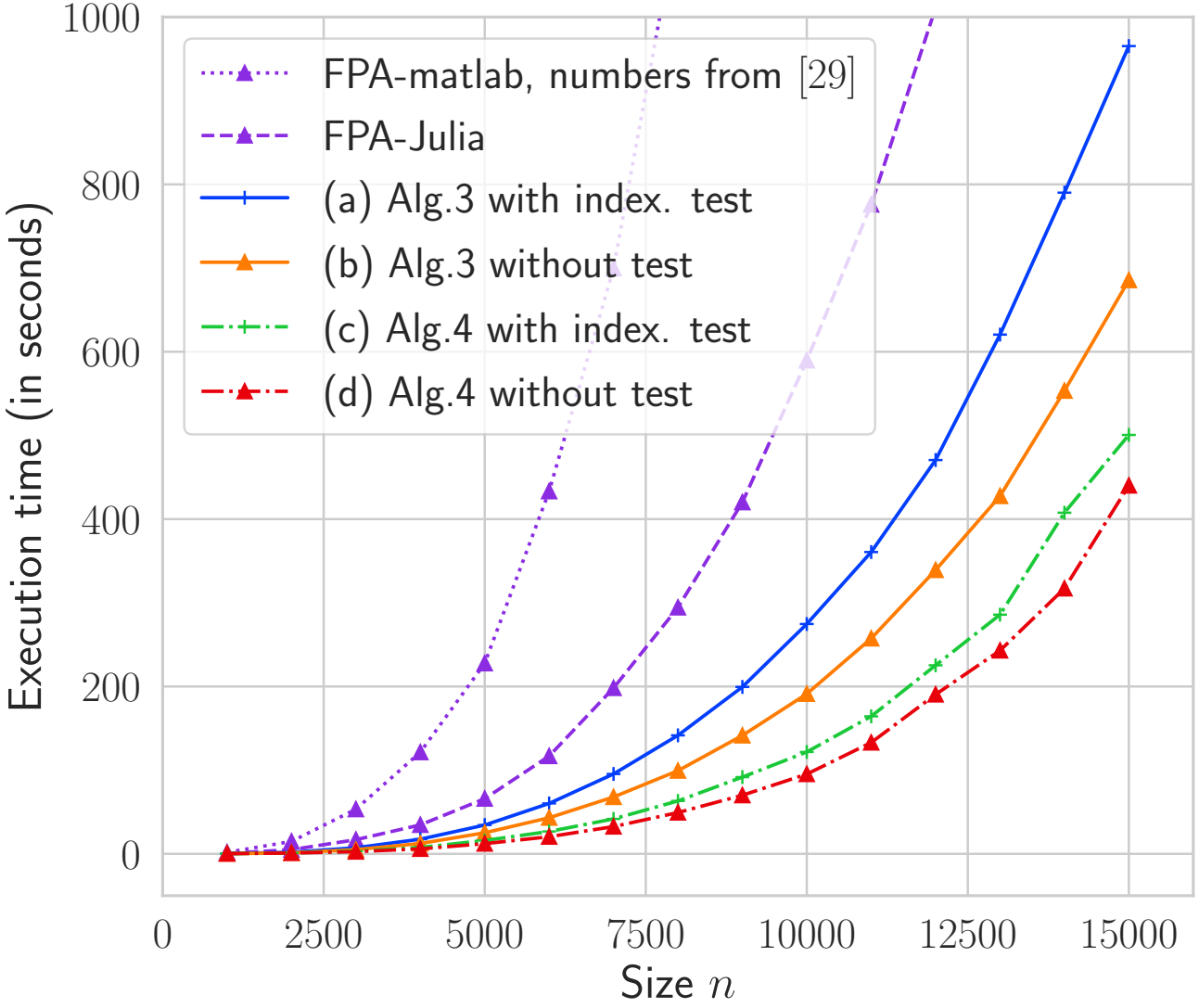
2. Sherman-Morisson formula: Let A be an invertible matrix, u and v vectors $1D$ such that $1 + v^T A^{-1} u \neq 0$. Then:

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

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

Conclusion

Time-average MDPs are complicated:

- Gain-optimality is defined but stronger notions of optimality are used.
- Be careful about structure.

$$\text{bias-optimal} \subsetneq \begin{array}{c}\text{Bellman-optimal}\\ \equiv \\ \text{canonical-optimal}\end{array} \subsetneq \text{gain-optimal}$$

- Bellman-optimality allows to define the advantage (as for discounted problems).

`http://polaris.imag.fr/nicolas.gast/`

- *Computing Whittle (and Gittins) Index in Subcubic Time*, G. Gaujal, Khun `https://arxiv.org/abs/2203.05207`