# Asymptotically optimal policies for weakly coupled Markov decision processes

Diego Goldsztajn

joint work with

Konstantin Avrachenkov

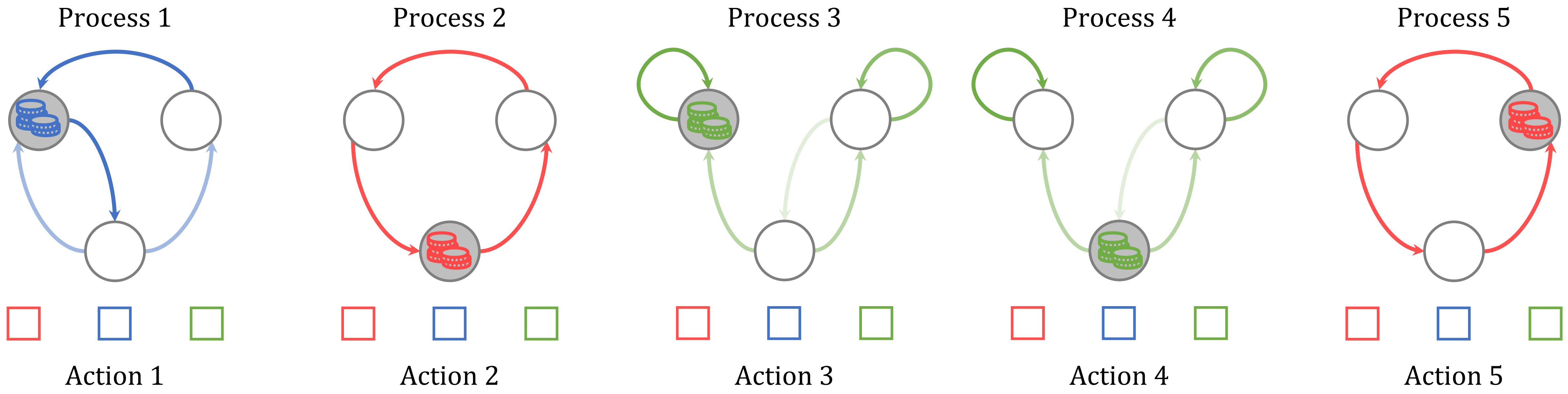
RÉPUBLIQUE FRANÇAISE
*Liberté Égalité Fraternité*

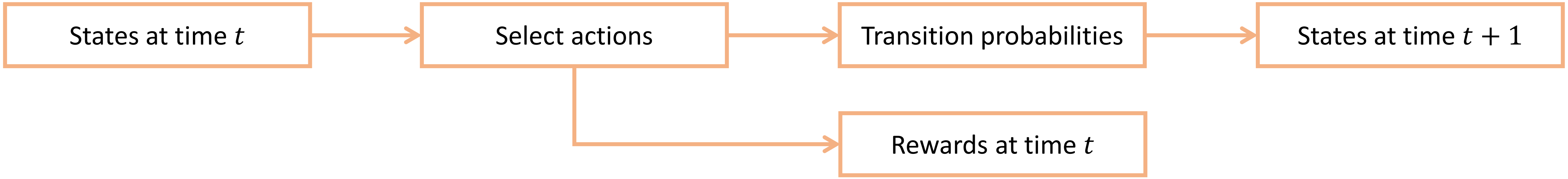
Inria

November 2024

# Problem formulation

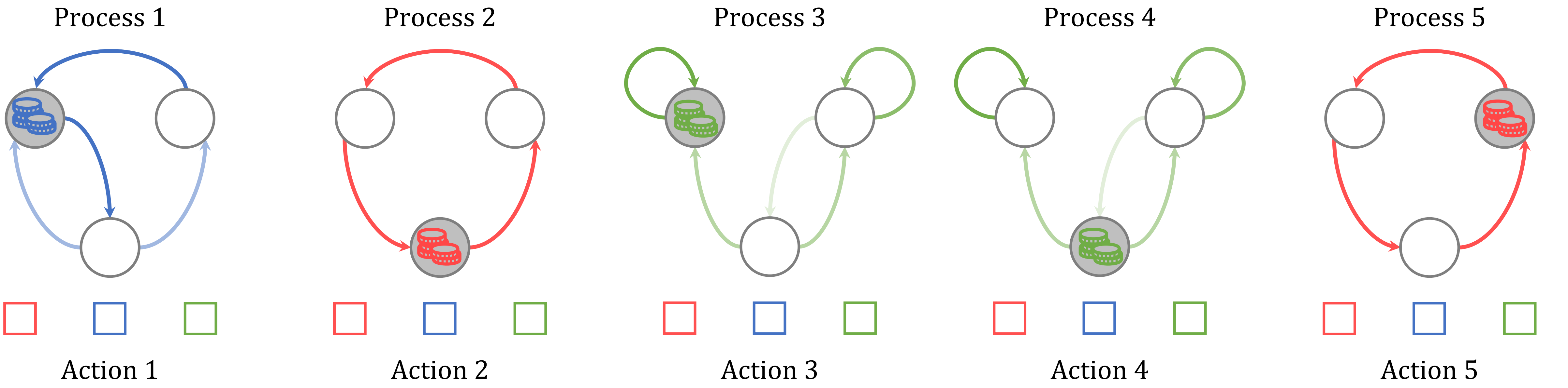
Identical Markov decision processes evolving in discrete time



# Problem formulation

Identical Markov decision processes evolving in discrete time




Transitions are independent (given states and actions) but processes coupled through action-selection constraints

Objective is to maximize expected average reward over infinite time horizon

Many applications: logistics, healthcare, communication networks, recommendation systems, etc.

Transition probabilities and rewards are known in advance

# Notation

Consider $n$ identical processes with finite action space $A$ and state space $S$

$$S_n(t,m) = \text{state of process } m \text{ at time } t \qquad A_n(t,m) = \text{action of process } m \text{ at time } t$$

$$x_n(t,i) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{I}_{\{S_n(t,m)=i\}}$$

State frequencies

$$y_n(t,i,a) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{I}_{\{S_n(t,m)=i, A_n(t,m)=a\}}$$

State-action frequencies

Action selection must respect multiple linear constraints at each time

$$\sum_{a\in A} y_n(t,a)C_n(a) = d_n \quad \text{and} \quad \sum_{a\in A} y_n(t,a)E_n(a) \preccurlyeq f_n \quad \text{for all} \quad t \geq 0 \quad \text{(matrix notation)}$$

Example

$$\sum_{a\in A} [y_n(t,1,a) \quad y_n(t,2,a) \quad y_n(t,3,a)] \begin{bmatrix} C_n(1,1,a) & C_n(2,1,a) \\ C_n(1,2,a) & C_n(2,2,a) \\ C_n(1,3,a) & C_n(2,3,a) \end{bmatrix} = [d_n(1) \quad d_n(2)]$$

# Notation

Consider $n$ identical processes with finite action space $A$ and state space $S$

$S_n(t,m) = \text{state of process } m \text{ at time } t$ $\qquad$ $A_n(t,m) = \text{action of process } m \text{ at time } t$

$$x_n(t,i) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{I}_{\{S_n(t,m)=i\}}$$

State frequencies

$$y_n(t,i,a) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{I}_{\{S_n(t,m)=i,A_n(t,m)=a\}}$$

State-action frequencies

Action selection must respect multiple linear constraints at each time

$$\sum_{a\in A} y_n(t,a)C_n(a) = d_n \quad \text{and} \quad \sum_{a\in A} y_n(t,a)E_n(a) \preccurlyeq f_n \quad \text{for all} \quad t \geq 0 \quad \text{(matrix notation)}$$

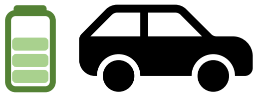
Ideally, we want policy that maximizes expected average reward over infinite time horizon

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{n}\sum_{m=1}^{n} E\big[r\big(S_n(t,m),A_n(t,m)\big)\big] = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{a\in A} E[y_n(t,a)]r(a) \quad \text{(informal)}$$

$$r(i,a) = \text{reward for process in state } i \text{ with action } a$$

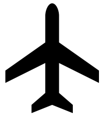
# Electric taxis example

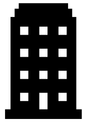
Each taxi is a process



$S_n(t,m) \in \{0,\dots,7\}$

Battery level

$A_n(t,m) = 0$ Send to airport  $S_n(t+1,m) \sim \min\{S_n(t,m) - X_0, 0\}$

$A_n(t,m) = 1$ Send downtown  $S_n(t+1,m) \sim \max\{S_n(t,m) - X_1, 0\}$

$A_n(t,m) = 2$ Charge taxi  $S_n(t+1,m) = \min\{S_n(t,m) + 2, 7\}$

**Action-selection constraints**

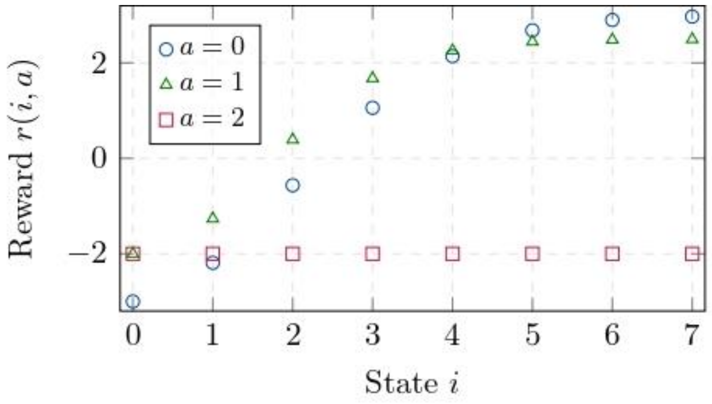
$$\sum_{i \in S} y_n(t,i,0) \geq 0.1 \quad \text{(at least 10\% at airport)}$$

$$\sum_{i \in S} y_n(t,i,2) \leq 0.7 \quad \text{(at most 70\% charging)}$$

**Implicit constraints**

$$\sum_{a \in A} y_n(t,i,a) = x_n(t,i)$$

$$y_n(t,i,a) \geq 0$$



# Restless bandits

Important particular case with two actions $A = \{0,1\}$ and a single constraint:

$$\sum_{i \in S} y_n(t,i,1) \leq \alpha \quad \text{(inequality constraint)} \quad \text{or} \quad \sum_{i \in S} y_n(t,i,1) = \frac{\lfloor \alpha n \rfloor}{n} \quad \text{(equality constraint)}$$

**Whittle index policy**

- [1988 - Whittle] Conjectures asymptotic optimality if indexability holds
- [1990 - Weber and Weiss] Counterexample but indexability and global attractivity sufficient
- [2023 - Gast, Gaujal and Yan] Bounds on optimality gap (exponentially small)

**LP-priority policies**

- [2016 - Verloop] LP-priority policies subsume Whittle index policy and are asymptotically optimal if global attractivity holds
- [2023 - Gast, Gaujal and Yan] Bounds on optimality gap (exponentially small)

**Non-priority policies**

- [2023 - Hong, Xie, Chen and Wang] Follow the Virtual Advice
- [2024 - Hong, Xie, Chen and Wang] Set-expansion and ID policies
- [2024 - Yan] Align and Steer policy

**Multiple actions and constraints**

- [2023 – Brown and Zhang] Finite horizon and discounted infinite horizon
- [2024 – Gast, Gaujal and Yan] Finite horizon

# Objective

All processes together form single MDP with state space $S^n$ and action spaces contained in $A^n$

History-dependent policies: map history to probability distribution on action vectors

Stationary policies: map state vectors to probability distributions on action vectors

Expected average reward (or gain) exists for stationary policies but may not exist for history-dependent policies

$$g_n^{\pi}(s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} \sum_{m=1}^{n} E_s^{\pi}\left[r\left(S_n(t,m), A_n(t,m)\right)\right]$$

Standard MDP theory implies that there exists stationary deterministic $\pi^*$ such that

$$g_n^{\pi^*}(s) = g_n^*(s) = \sup_{\pi} \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} \sum_{m=1}^{n} E_s^{\pi}\left[r\left(S_n(t,m), A_n(t,m)\right)\right] \quad \text{for all} \quad s \in S^n$$

## Curse of dimensionality

Optimal policy can be computed with dynamic programming but computation time grows exponentially with $n$

Objective is to find simple policy with asymptotically optimal gain

# Linear program relaxation

Assume that $C_n(a) \to C(a)$, $E_n(a) \to E(a)$, $d_n \to d$ and $f_n \to f$ as $n \to \infty$

**Linear program relaxation**

$$g_r = \underset{y \in \Delta_{S\times A}}{\text{maximize}} \sum_{a\in A} y(a) r(a)$$

$$\text{subject to} \sum_{a\in A} y(a)P(a) = \sum_{a \in A} y(a)$$

$$\sum_{a\in A} y(a)C(a) = d$$

$$\sum_{a\in A} y(a)E(a) \preccurlyeq f$$

$P(a) = \text{transition matrix given action } a$

**Interpretation**

$$y(i,a) = \lim_{n\to\infty} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} E_s^{\pi}[y_n(t,i,a)]$$

Inner limit exists for stationary policies

Constraints need only hold on average

**Upper bound**

$$\limsup_{n\to\infty} \sup_{s\in S^n} g_n^*(s) \le g_r$$

If $C_n(a) = C(a)$, $E_n(a) = E(a)$, $d_n = d$ and $e_n = e$, then $g_n^*(s) \le g_r$ for all $s \in S^n$

# Fluid problem

Define $X_n = \{x \in \Delta_S : nx \in \mathbb{Z}^S\} \subset \Delta_S = X$ and $Y_n = \{y \in \Delta_{S\times A} : ny \in \mathbb{Z}^{S\times A}\} \subset \Delta_{S\times A} = Y$

**Discrete control**

$\phi_n : X_n \to Y_n$ such that for all $x \in X_n$:

$$\sum_{a\in A} \phi_n(x)(a) = x,$$

$$\sum_{a\in A} \phi_n(x)(a) C_n(a) = d_n \quad \text{and} \quad \sum_{a\in A} \phi_n(x)(a) E_n(a) \preccurlyeq f_n$$

$\phi_n$ determines evolution of $(x_n, y_n)$ and we have:

$$y_n(t) = \phi_n\big(x_n(t)\big) \quad \text{and} \quad E[x_n(t+1)] = \sum_{a\in A} E[y_n(t,a)] P(a)$$

**Fluid control**

$\phi : X \to Y$ such that for all $x \in X$:

$$\sum_{a\in A} \phi(x)(a) = x,$$

$$\sum_{a\in A} \phi(x)(a) C(a) = d \quad \text{and} \quad \sum_{a\in A} \phi(x)(a) E(a) \preccurlyeq f$$

Fluid trajectory given by $\phi$ and $x(0) = x^0$ is:

$$y(t) = \phi\big(x(t)\big) \quad \text{and} \quad x(t+1) = \sum_{a\in A} y(t,a) P(a)$$

Recall that $g_r$ is upper bound for gain in the limit as $n \to \infty$

**Fluid problem** Find $\phi$ such that $\sum_{a\in A} y(t,a) r(a) \to g_r$ as $t \to \infty$ for all fluid trajectories (i.e., regardless of initial condition)

# Overview of main results

Theorem. Consider discrete controls $\phi_n$ and fluid control $\phi$ such that:

- $\phi$ solves the fluid problem and is continuous
- $\max_{x \in X_n} \|\phi(x) - \phi_n(x)\| \to 0$ as $n \to \infty$

The gain of the discrete controls $\phi_n$ approaches $g_r$ as $n \to \infty$ for arbitrary (and possibly random) initial conditions $\{x_n(0) : n \geq 1\}$

We can obtain asymptotically optimal policies in two steps:

1. Find continuous solution of fluid problem
2. Construct discrete controls that approximate solution (rounding)

Particular case   If $y^*$ solves LP and $y(t) \to y^*$ as $t \to \infty$ for all fluid trajectories, then $\phi$ solves the fluid problem

We provide conditions for:

1. Existence of solutions to fluid problem (sufficient and necessary for particular case)
2. Explicit constructions of solutions and asymptotically optimal discrete controls

# Sufficient conditions for asymptotic optimality

$$y^* \text{ solves LP}, \qquad x^* = \sum_{a \in A} y^*(a), \qquad S_+^* = \{i \in S : x^*(i) > 0\}$$

## 1. Structure of transition matrices

Single process (no constraints or rewards) admits $\pi$ such that:

- Markov chain associated with $\pi$ is unichain and aperiodic
- $S_+^*$ is contained in the unique irreducible class

Such policy $\pi$ exists if and only if the policy that selects actions uniformly at random has the above properties

## 2. Structure of constraints

Problem satisfies the following properties:

- Restless bandit problem or multiple inequality constraints
- Nonnegative coefficients (resource allocation)
- All coefficients are zero for one action

These conditions are for enforcing feasibility

If the above conditions hold, we provide explicit constructions for:

- A solution $\phi$ of the fluid problem such that $y(t) \to y^*$ as $t \to \infty$ for all fluid trajectories
- Asymptotically optimal discrete controls $\phi_n$ such that $\max_{x \in X_n} \|\phi(x) - \phi_n(x)\| \to 0$ as $n \to \infty$

# Conditions for asymptotic optimality of restless bandits

### Whittle index policy

- Indexability
- Global attractor property
- System is unichain and aperiodic

### Follow the Virtual Advice

Consider relaxed single-process problem

Optimal policy is unichain and satisfies synchronization assumption

### Relaxed problem

Same rewards but relaxed constraint:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{i\in S} y(t,i,1) = \alpha$$

### LP-priority policies

- Global attractor property
- System is unichain and aperiodic

### ID policy

Consider relaxed single-process problem

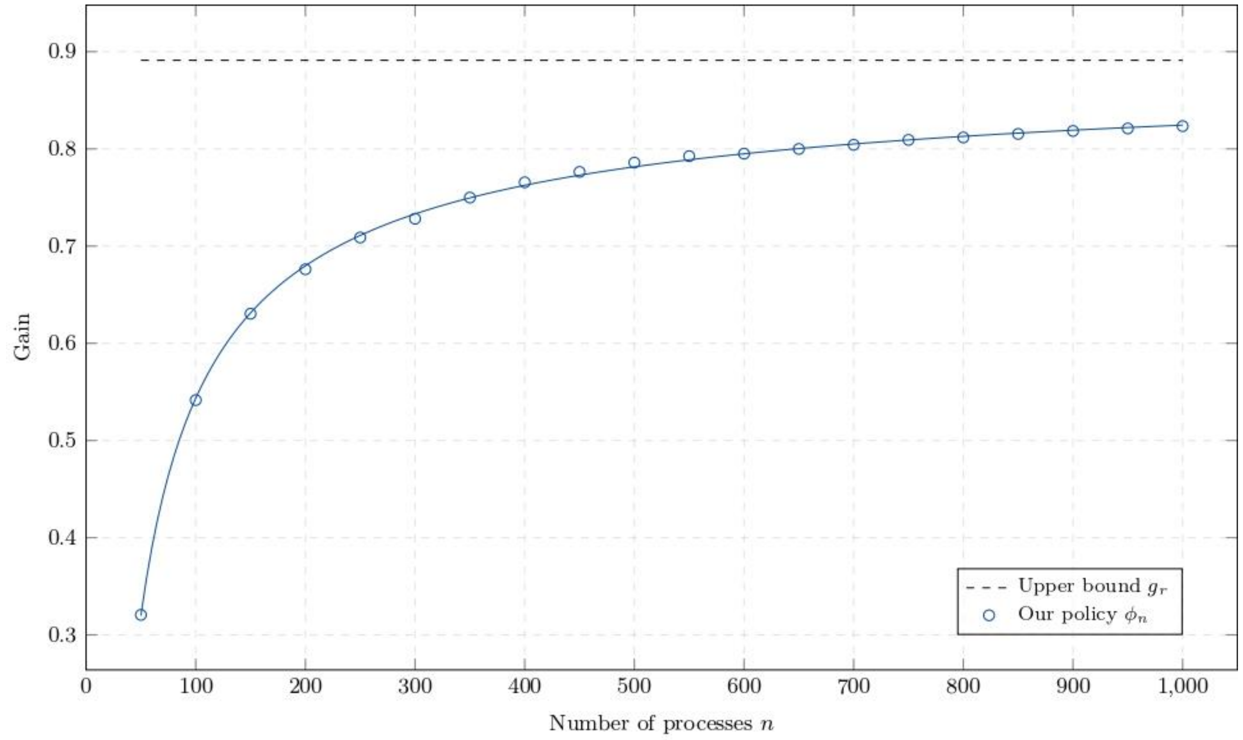
Optimal policy is unichain and aperiodic

### Our conditions

Consider policy that selects actions uniformly at random

This policy is unichain, aperiodic and $S_+^*$ is contained in irreducible class

# Electric taxis example




Resource allocation inequality constraints:

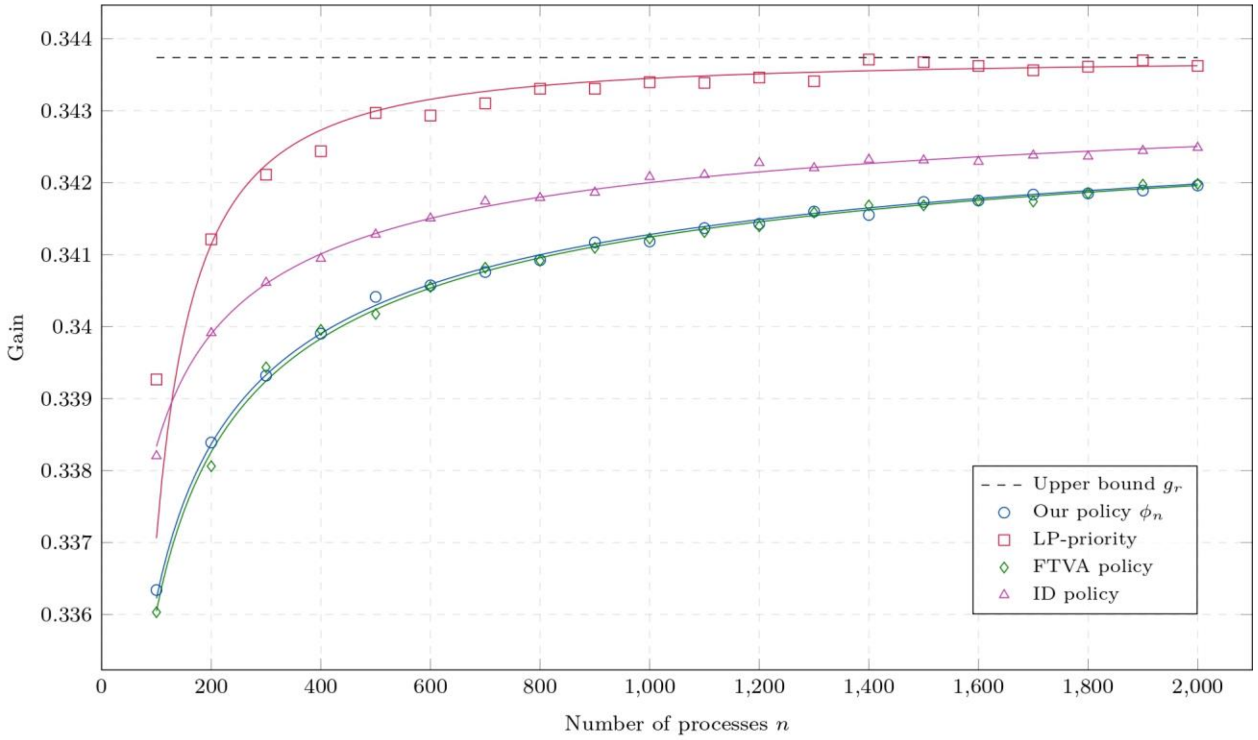
$$\sum_{i \in S} [y_n(t,i,1) + y_n(t,i,2)] \leq 0.9$$

$$\sum_{i \in S} y_n(t,i,2) \leq 0.7$$

All coefficients are zero for action “send to airport”

Multichain example

# Counterexample for Whittle index policy



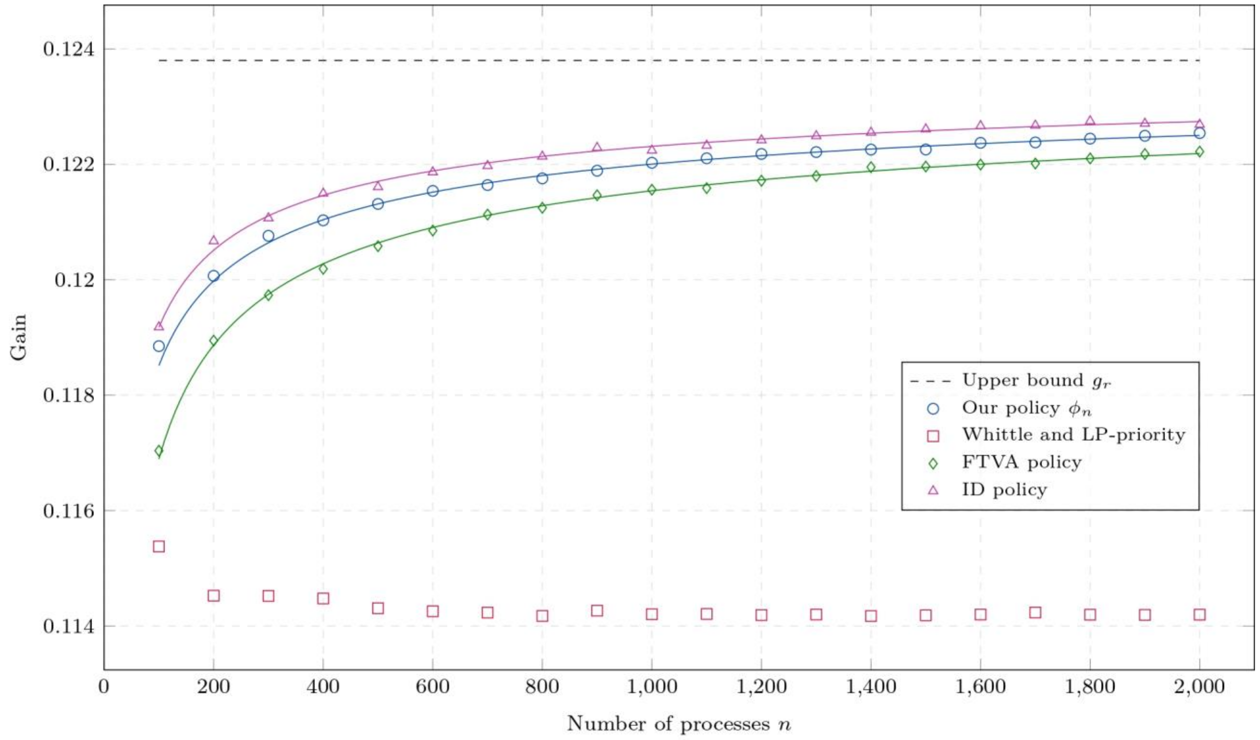

Equality-constrained restless bandits

Example from [2023 - Gast, Gaujal and Khun]

Indexability condition fails

Whittle indexes are not well-defined

# Counterexample for LP-priority policies

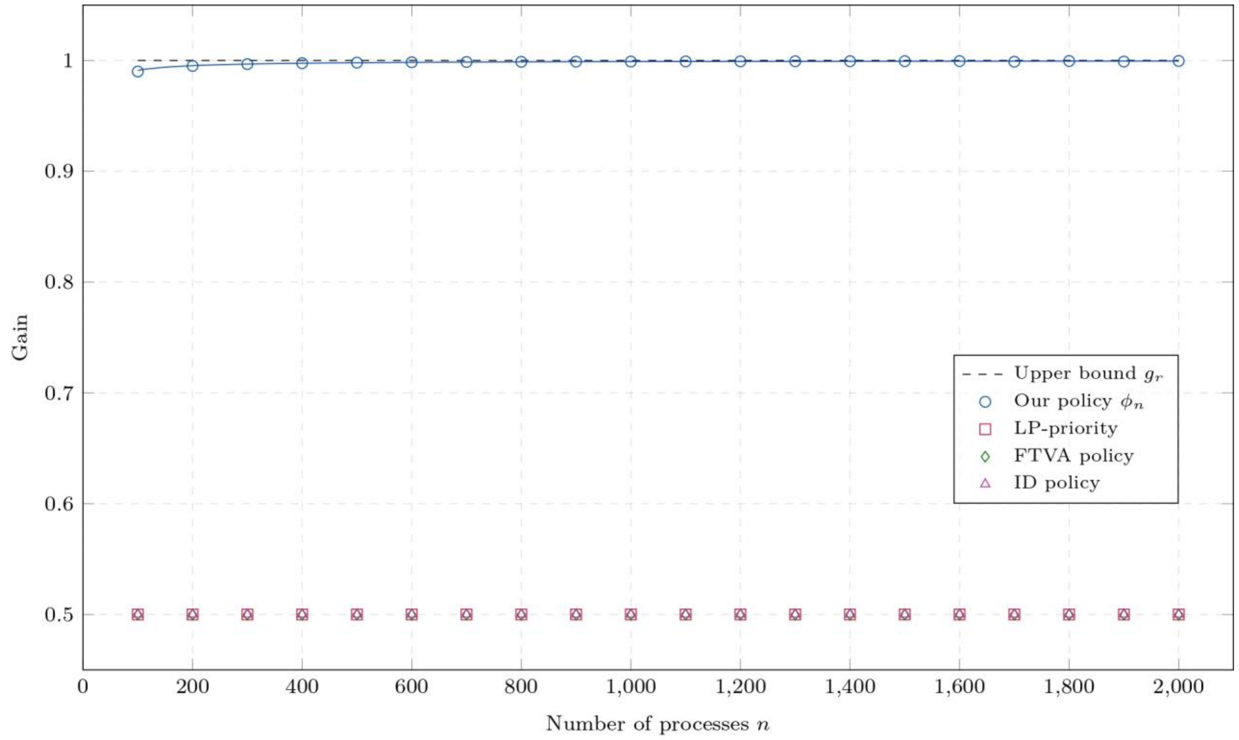


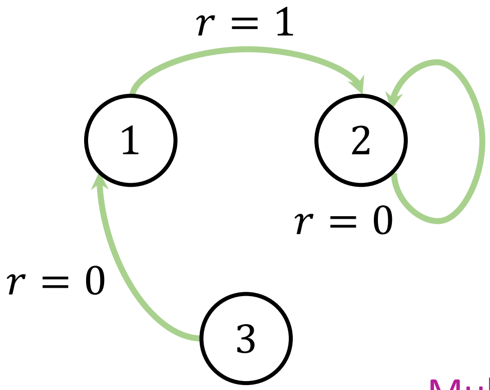

Equality-constrained restless bandits

Example from [2020 - Gast, Gaujal and Yan]

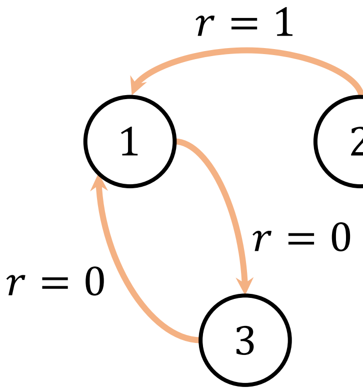
Global attractor property fails for LP-priority policy (and Whittle)

# Counterexample for FTVA and ID policy




Multichain example




Equality constraint with $\alpha = 1/2$

Initial condition: all processes in state 1

# Asymptotic optimality result (reminder)

Fluid problem    Find $\phi$ such that $\sum_{a \in A} y(t,a) r(a) \to g_r$ as $t \to \infty$ for all fluid trajectories (i.e., regardless of initial condition)

Fluid control is $\phi : X \to Y$ such that for all $x \in X$:

$$\sum_{a \in A} \phi(x)(a) = x,$$

$$\sum_{a \in A} \phi(x)(a) C(a) = d, \qquad \sum_{a \in A} \phi(x)(a) E(a) \preccurlyeq f$$

Fluid trajectory given by $\phi$ with initial condition $x^0$ is:

$$x(0) = x^0, \qquad y(t) = \phi\big(x(t)\big)$$

$$x(t+1) = \sum_{a \in A} y(t,a) P(a)$$

Theorem. Consider discrete controls $\phi_n$ and fluid control $\phi$ such that:

- $\phi$ solves the fluid problem and is continuous
- $\max_{x \in X_n} \|\phi(x) - \phi_n(x)\| \to 0$ as $n \to \infty$

The gain of the discrete controls $\phi_n$ approaches $g_r$ as $n \to \infty$ for arbitrary (and possibly random) initial conditions $\{x_n(0) : n \geq 1\}$

# Main ideas of proof

Suppose fluid control $\phi$ and discrete controls $\phi_n$ are as in theorem

$$x_n(t+1) = z_n(t+1) + \sum_{a \in A} y_n(t,a)P(a) \quad \text{with} \quad E[z_n(t)] = 0 \quad \text{and} \quad E[\|z_n(t)\|_2^2] \le \frac{1}{n} E[\|x_n(t)\|_1]$$

Lemma 1. Let $x^0$ be random variable on $\Omega$ such that $x_n(0) \Rightarrow x^0$ as $n \to \infty$

$$x_n(t) \Rightarrow x(t) \quad \text{and} \quad y_n(t) \Rightarrow y(t) \quad \text{as} \quad n \to \infty$$

where $\{x(\omega,t), y(\omega,t) : t \ge 0\}$ is fluid trajectory with $x(\omega,0) = x^0(\omega)$

Lemma 2. Let $x_n(0)$ be stationary distribution of $x_n$

$$\lim_{n\to\infty} \sum_{a \in A} E[y_n(0,a)]r(a) = g_r$$

Consider a MDP with finite state and action spaces

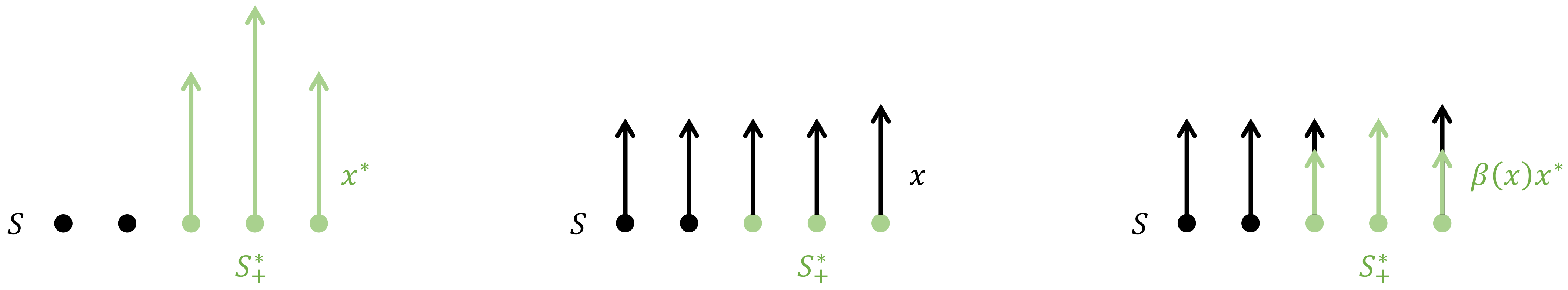
Let $\pi$ be a stationary (deterministic) policy with transition matrix $P_\pi$

$$g^\pi(\nu) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\nu^\pi\big[r\big(S(t),A(t)\big)\big] = \nu P_\pi^* r_\pi = E_{S\sim\nu P_\pi^*}\big[r\big(S,\pi(S)\big)\big] \quad \text{with} \quad P_\pi^* = \frac{1}{K}\sum_{k=0}^{K-1} P_\pi^k$$

Proof of Theorem: apply Lemma 2 with the stationary distribution $x_n(0)$ that gives gain of $\phi_n$ for the given initial distribution

# Solutions of the fluid problem

$y^*$ solves LP, $\quad x^* = \sum_{a \in A} y^*(a)\,, \quad S_+^* = \{i \in S : x^*(i) > 0\}, \quad \beta(x) = \max\{\lambda \geq 0 : \lambda x^* \preccurlyeq x\}$



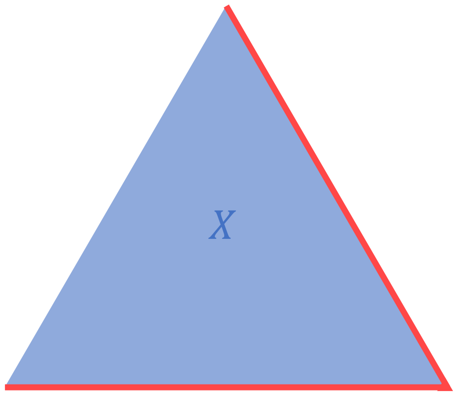
# Solutions of the fluid problem

$$y^* \text{ solves LP}, \qquad x^* = \sum_{a \in A} y^*(a), \qquad S_+^* = \{i \in S : x^*(i) > 0\}, \qquad \beta(x) = \max\{\lambda \geq 0 : \lambda x^* \preccurlyeq x\}$$

Theorem. The following statements are equivalent:

- There exists a continuous fluid control $\psi$ such that, for all fluid trajectories, $x(t)$ leaves $\{x \in X : \beta(x) = 0\}$ in finite time
- There exists a continuous solution of the fluid problem $\phi$ such that $y(t) \to y^*$ as $t \to \infty$ for all fluid trajectories

Furthermore, we can take $\phi(x) = \beta(x) y^* + [1 - \beta(x)] \psi([1 - \beta(x)]^{-1} [x - \beta(x) x^*])$




$\{x \in X : \beta(x) = 0\} = \{x \in X : x(i) = 0 \text{ for some } i \in S_+^*\}$

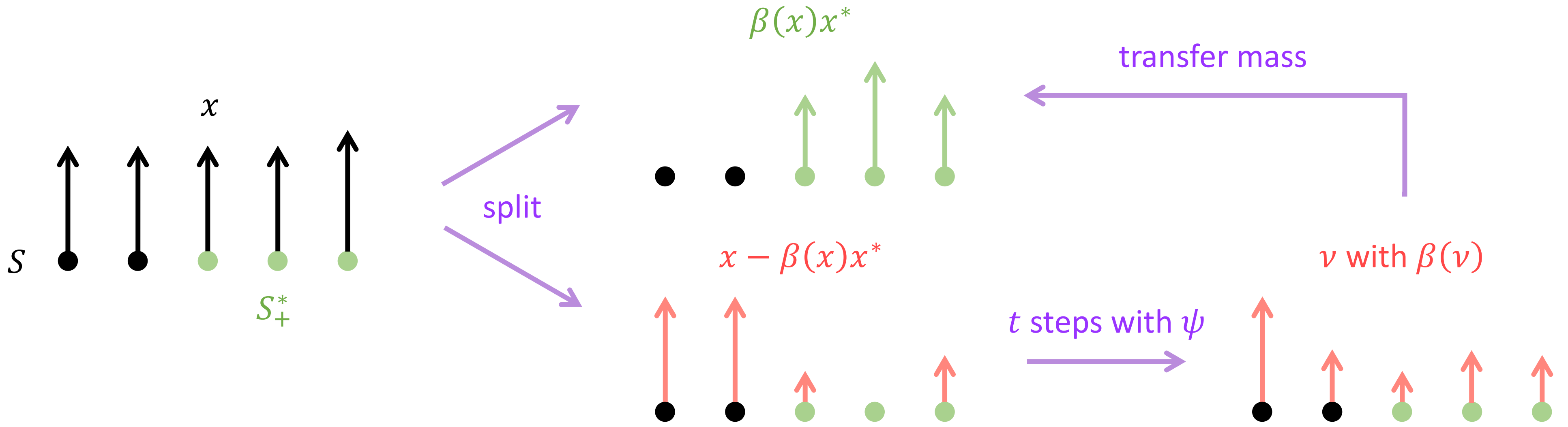
# Solutions of the fluid problem

$$y^* \text{ solves LP}, \qquad x^* = \sum_{a \in A} y^*(a), \qquad S_+^* = \{i \in S : x^*(i) > 0\}, \qquad \beta(x) = \max\{\lambda \geq 0 : \lambda x^* \preccurlyeq x\}$$

Theorem. The following statements are equivalent:

- There exists a continuous fluid control $\psi$ such that, for all fluid trajectories, $x(t)$ leaves $\{x \in X : \beta(x) = 0\}$ in finite time
- There exists a continuous solution of the fluid problem $\phi$ such that $y(t) \to y^*$ as $t \to \infty$ for all fluid trajectories

Furthermore, we can take $\phi(x) = \beta(x)y^* + [1 - \beta(x)]\psi([1 - \beta(x)]^{-1}[x - \beta(x)x^*])$



# Sufficient conditions for constructing solutions

Consider functions $\psi : X \to Y$ of the form

$$\psi(x) = \gamma\psi_1(x) + (1-\gamma)\psi_2(x) \quad \text{where} \quad \gamma \in (0,1] \quad \text{and} \quad \psi_1(x)(i,a) = x(i)\pi(a|i)$$

Here $\psi_1$ is based on a single-process policy $\pi$ and we assume that:

- $\pi$ is unichain and aperiodic and $S_+^*$ is contained in its unique irreducible class
- $\psi_2$ is such that the convex combination satisfies the constraints

We prove by induction that

$$(L \circ \psi)^t(x^0) = \gamma^t x^0 P_\pi^t + (1-\gamma)w_t(x^0) \quad \text{where} \quad L(y) = \sum_{a\in A} y(a)P(a) \quad \text{and} \quad w_t(x^0) \in \mathbb{R}_+^S$$

Conditions of theorem hold: if $x_\pi$ is the stationary distribution of $\pi$, then $x_\pi(i) > 0$ for all $i \in S_+^*$ and $x^0 P_\pi^t \to x_\pi$

We can define $\psi_2$ and $\gamma$ explicitly in the following cases:

- Restless bandit problem with equality or inequality constraints
- Problems with multiple resource allocation inequality constraints and one action that does not consume resources

In these cases we can also define discrete controls $\phi_n$ explicitly

# Conclusion

Asymptotically optimal policies can be obtained in two steps:

1. Find a continuous fluid control $\phi$ that solves the fluid problem
2. Define discrete controls $\phi_n$ that approach $\phi$ uniformly

We provided sufficient conditions and constructions for carrying out these steps

1. There exists a suitable unichain and aperiodic single-process policy (constraints and rewards not involved)
2. Restless bandit problem or multiple resource allocation inequality constraints

Second condition is for constructing feasible policies explicitly

We compared our policy with other policies for restless bandit problems

- Our results seem to hold under weaker assumptions
- Our policy is asymptotically optimal when other policies are not

Thanks for your attention