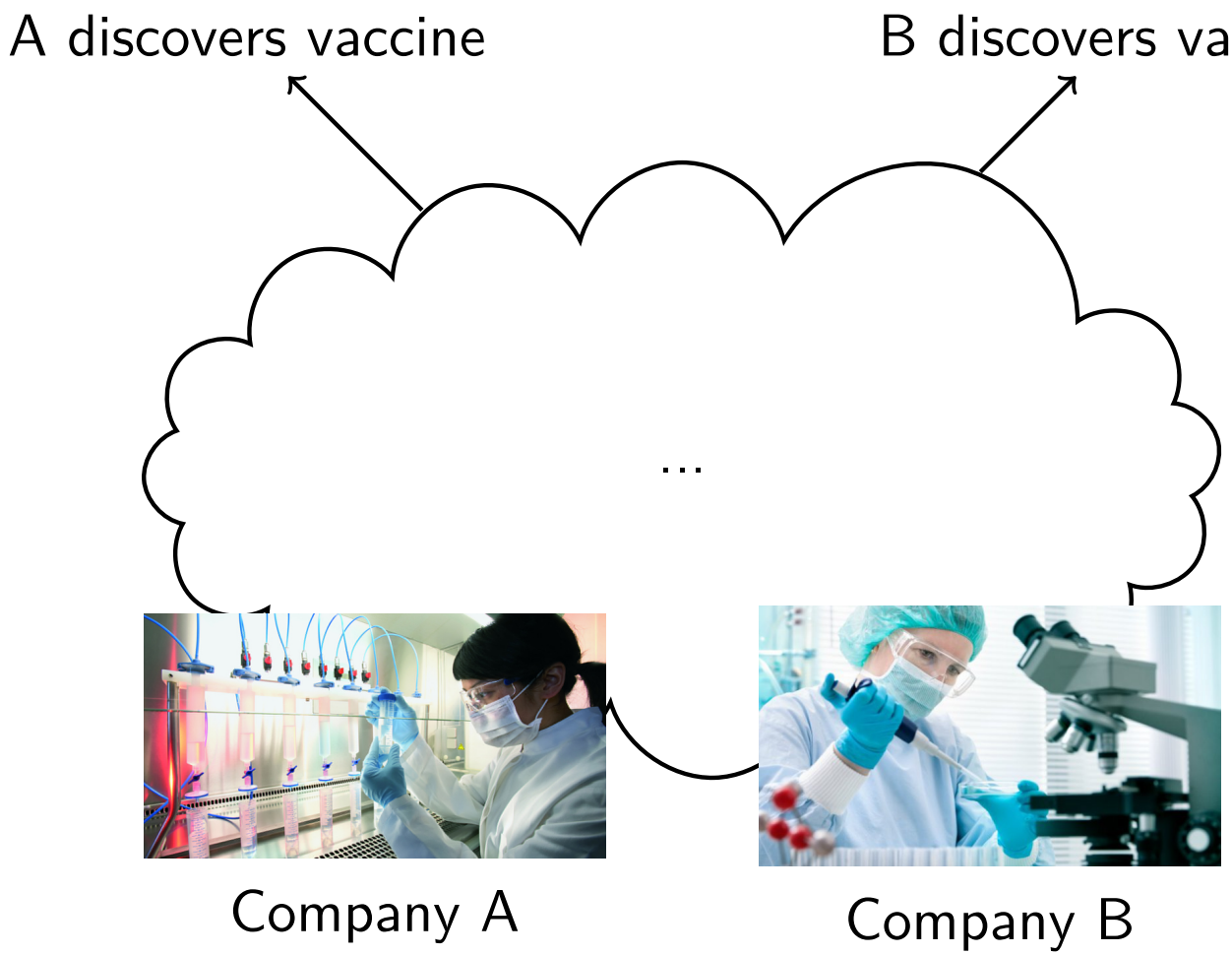
Hedging bets in Markov decision processes

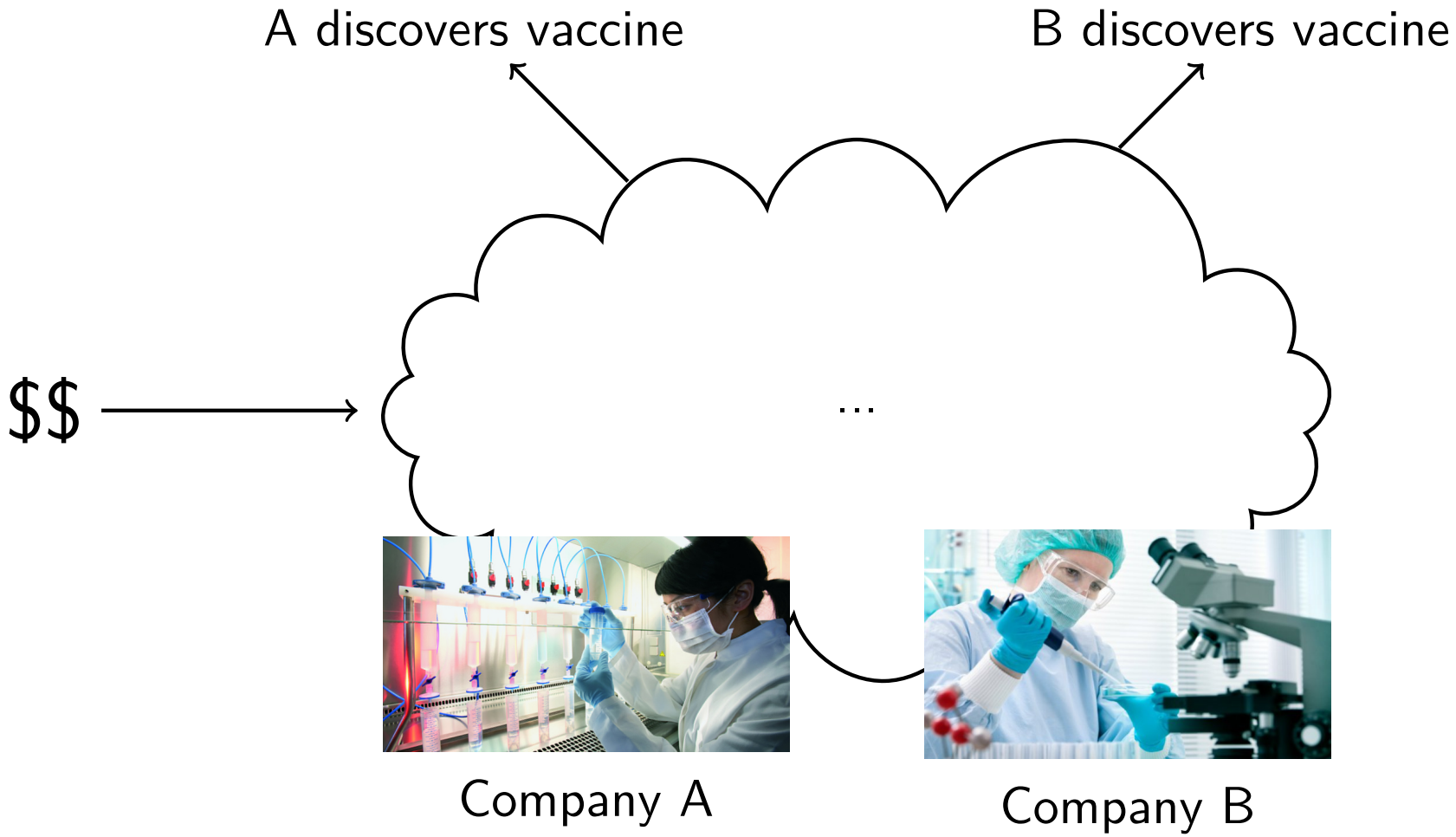
Rajeev Alur[2] **Marco Faella**[1] Sampath Kannan[2] Nimit Singhania[2]

[1]University of Naples "Federico II", Italy

[2]University of Pennsylvania, USA

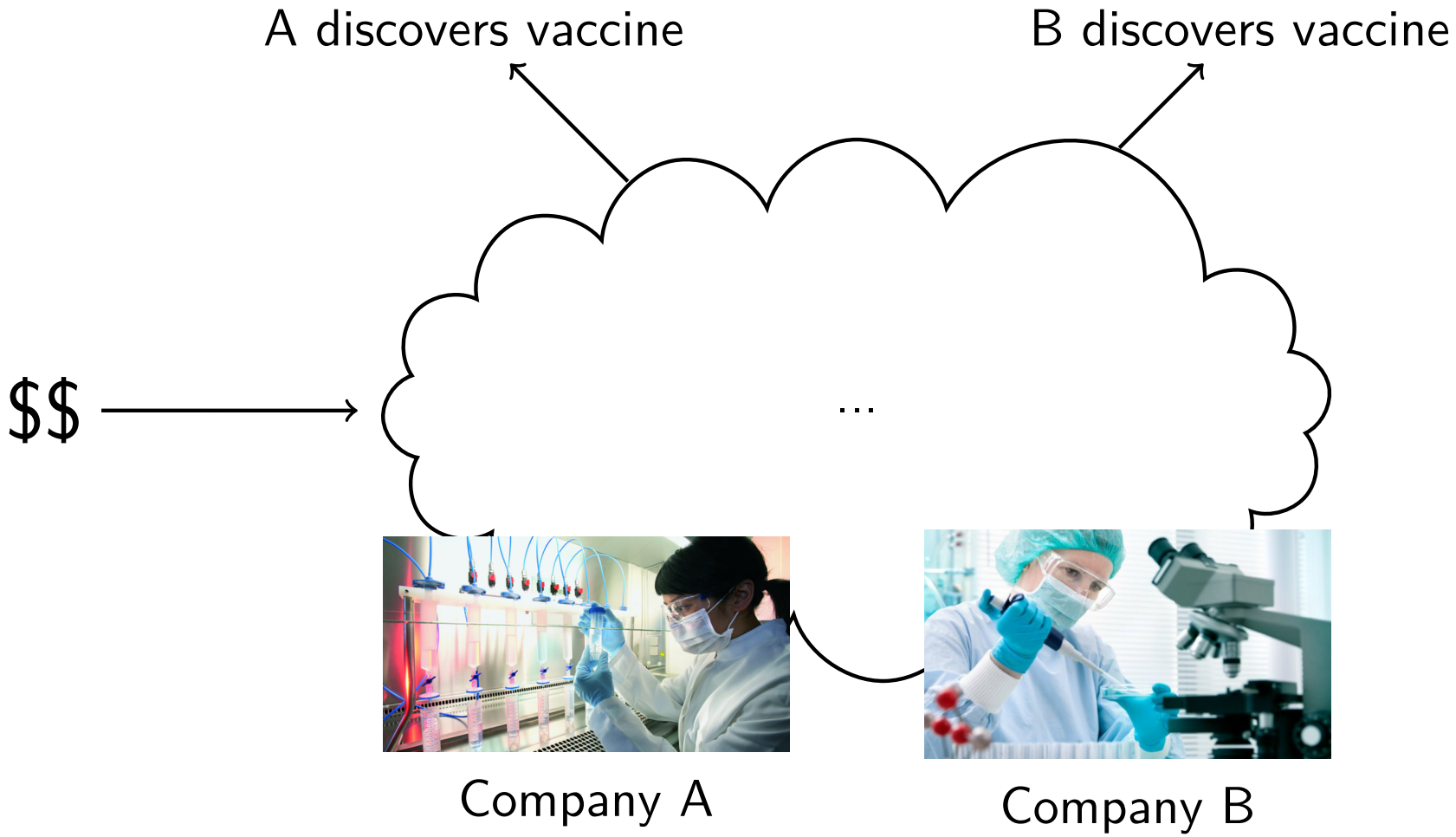
CSL 2016


A discovers vaccine
B discovers vaccine
...
Company A
Company B

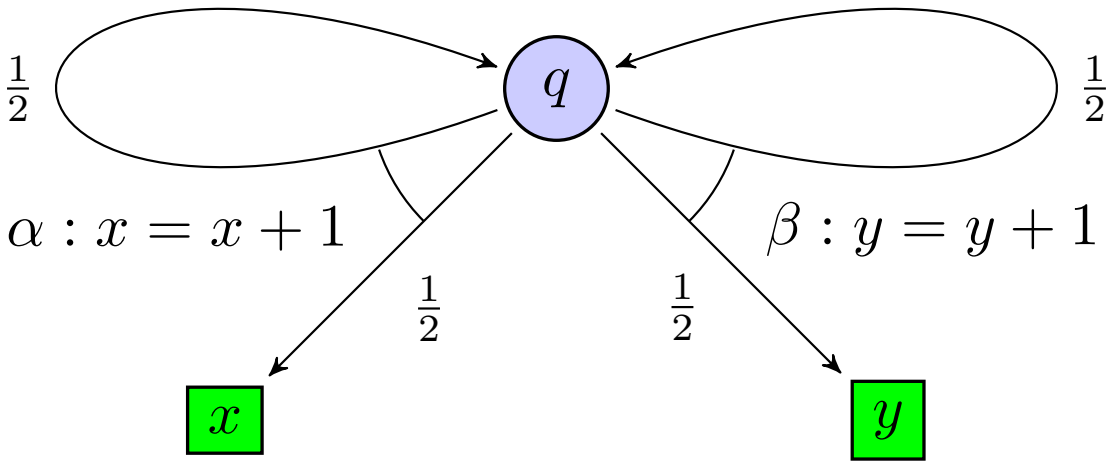

A discovers vaccine
B discovers vaccine
$$
...
Company A
Company B

The model: MDP with Alternative Objectives (MDPAO)

- Finite-state, discrete-time MDP
- **Multiple integer registers**
- An action specifies:
  - an integer update δ_x for each register x
  - prob. p_q of transitioning into each state q
  - prob. p_x of immediate termination in each register x
- The process eventually terminates in a register (almost surely)
- Cost of the run is the **final value of that register**


A discovers vaccine
B discovers vaccine
$$
...
Company A
Company B

Example



Two runs:

$$(q,0,0) \xrightarrow{\beta} (q,0,1) \xrightarrow{\beta} (q,0,2) \xrightarrow{\alpha} x \qquad \text{cost} = 0$$

$$(q,0,0) \xrightarrow{\beta} (q,0,1) \xrightarrow{\beta} y \qquad \text{cost} = 1$$

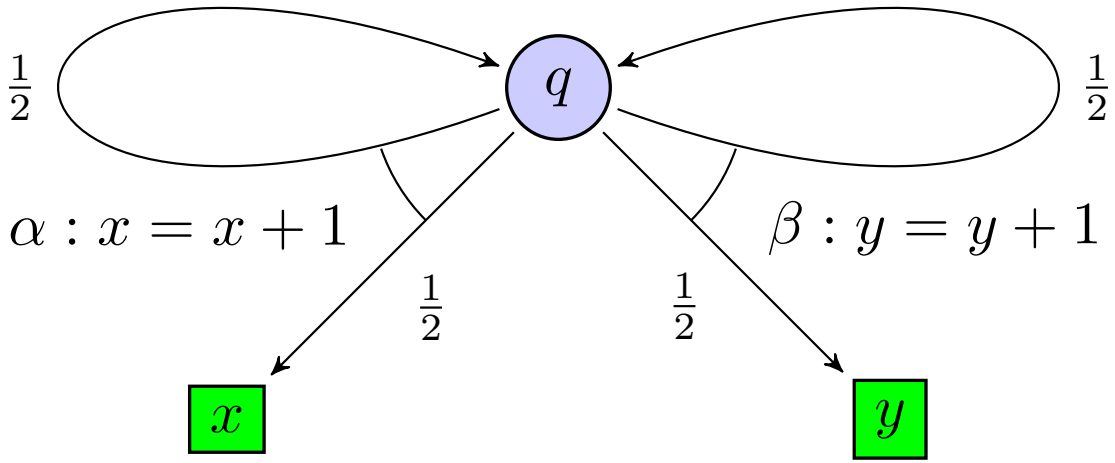
Decision Problem

A *strategy* chooses the next action, given past states and register values

Problem: Find the **minimum expected cost** of any strategy

(given a rational number, compare it with the min. exp. cost)

Example



Exp. cost of "always α": 1

Exp. cost of "always β": 1

Exp. cost of alternating: $\frac{1}{3}$ (optimal, requires memory)

Results

- 1 action (MC with Alt. Objs.): expected cost is computable in PTIME
- 2 actions and 1 state: decidable (EXPTIME)
- 2 registers: decidable under *tie-less* assumption (2EXPTIME)
- general: approximation algorithm

Section 1

2-action 1-state MDPAOs

Approach

Two actions α, β

We define a *preference function* $P : \mathbb{R}^n \to \mathbb{R}$ from register values to real numbers

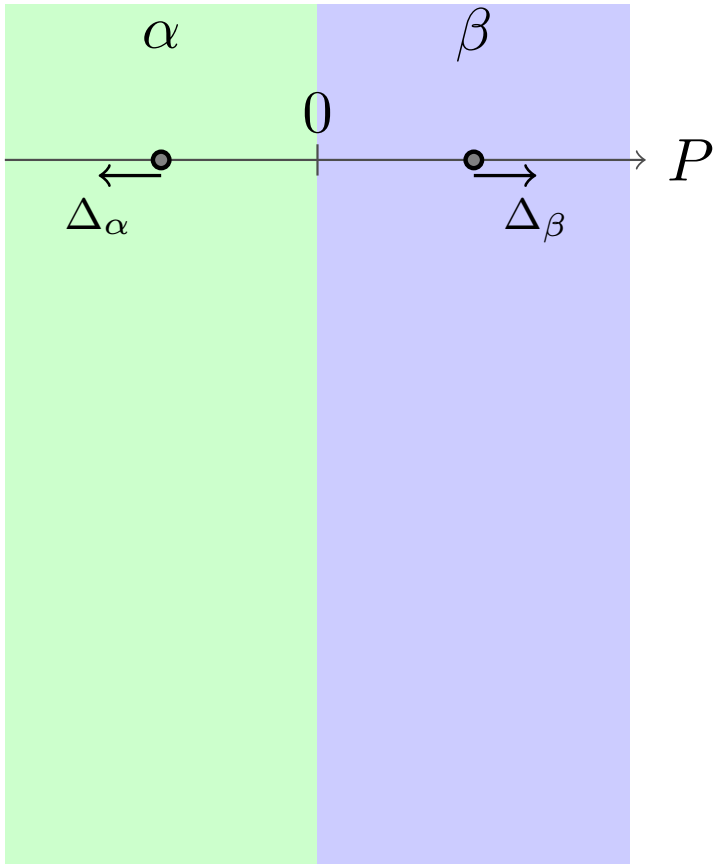
When $P(\nu) \leq 0$, α is optimal; otherwise, β is

Preference trends

Performing action $\gamma \in \{\alpha, \beta\}$ adds a constant Δ_γ to the preference P

Four cases:

1. $\Delta_\alpha < 0$, $\Delta_\beta \geq 0$



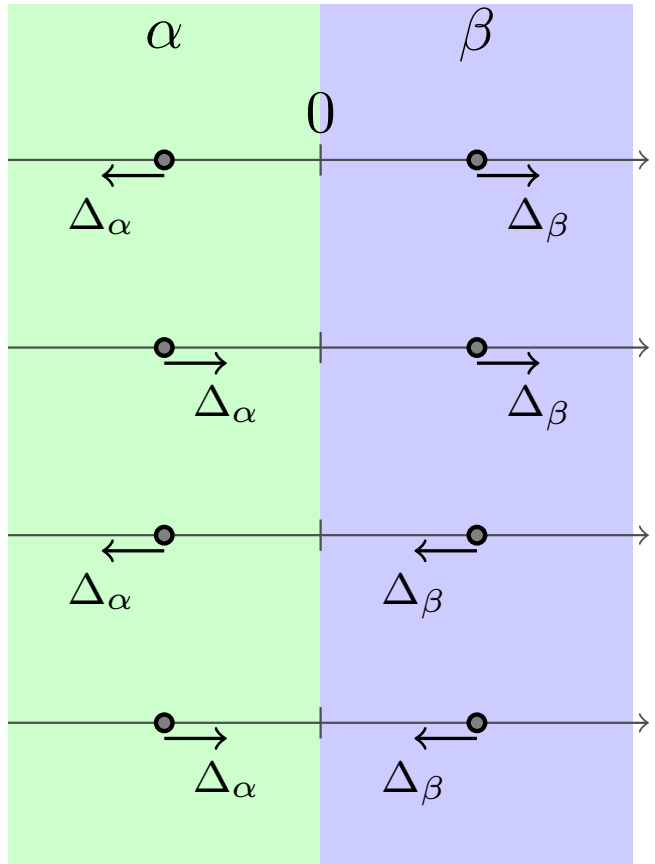
optimal strategies

$\alpha^\omega + \beta^\omega$

Preference trends

Performing action $\gamma \in \{\alpha, \beta\}$ adds a constant Δ_γ to the preference P

Four cases:



	optimal strategies
1. $\Delta_\alpha < 0,\ \Delta_\beta \geq 0$	$\alpha^\omega + \beta^\omega$
2. $\Delta_\alpha \geq 0,\ \Delta_\beta \geq 0$	$\alpha^k \beta^\omega$
3. $\Delta_\alpha < 0,\ \Delta_\beta < 0$	$\beta^k \alpha^\omega$
4. $\Delta_\alpha \geq 0,\ \Delta_\beta < 0$	$(\alpha^m + \beta^n)(\alpha^k \beta^l)^\omega$

The algorithm

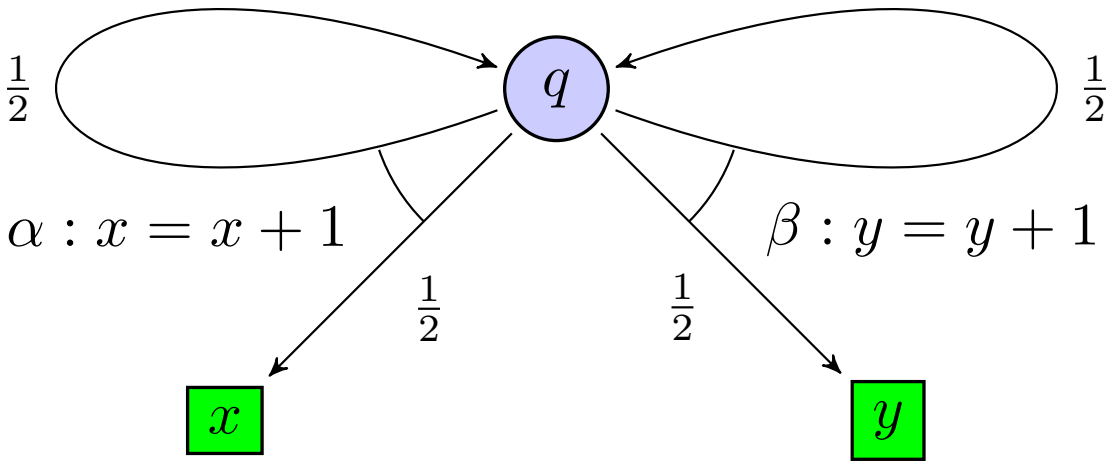
Given an MDPAO with 1 state and 2 actions:

1. Find optimal strategy by computing Δ_α, Δ_β, $P(\mathbf{0})$
   - $P(\mathbf{0})$ is the value of the preference when all registers are zero
2. Build the corresponding MC with Alt. Objs.
3. Return its expected cost

Theorem

The min. exp. cost of a 2-action 1-state MDPAO can be computed in EXPTIME.

Example

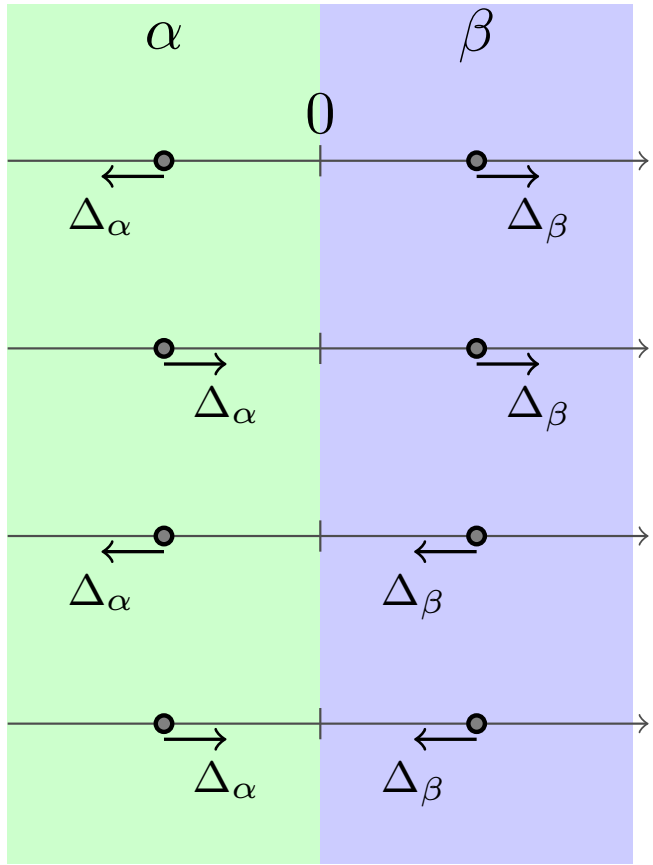


$$P(\mathbf{0}) = 0,\ \Delta_\alpha = 1,\ \Delta_\beta = -1$$

Preference trends

Performing action $\gamma \in \{\alpha, \beta\}$ adds a constant Δ_γ to the preference P

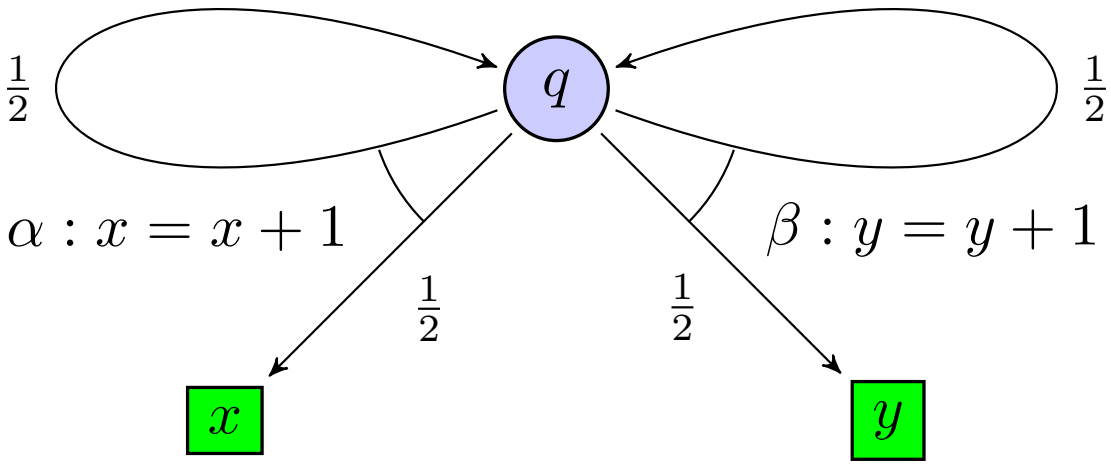
Four cases:



optimal strategies

1. $\Delta_\alpha < 0,\ \Delta_\beta \geq 0$ — $\alpha^\omega + \beta^\omega$
2. $\Delta_\alpha \geq 0,\ \Delta_\beta \geq 0$ — $\alpha^k \beta^\omega$
3. $\Delta_\alpha < 0,\ \Delta_\beta < 0$ — $\beta^k \alpha^\omega$
4. $\boldsymbol{\Delta_\alpha \geq 0,\ \Delta_\beta < 0}$ — $\boldsymbol{(\alpha^m + \beta^n)(\alpha^k \beta^l)^\omega}$

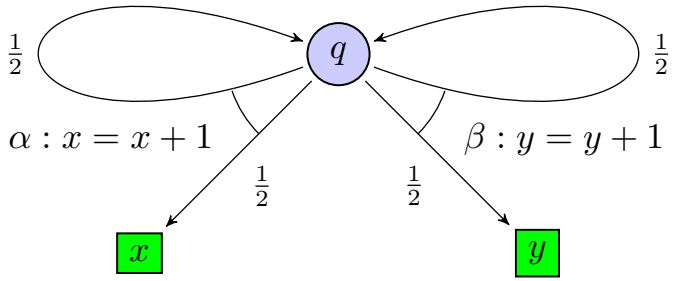
Example



$$P(\mathbf{0}) = 0,\ \Delta_\alpha = 1,\ \Delta_\beta = -1$$

optimal strategy from $\mathbf{0}$: $(\alpha\beta)^\omega$

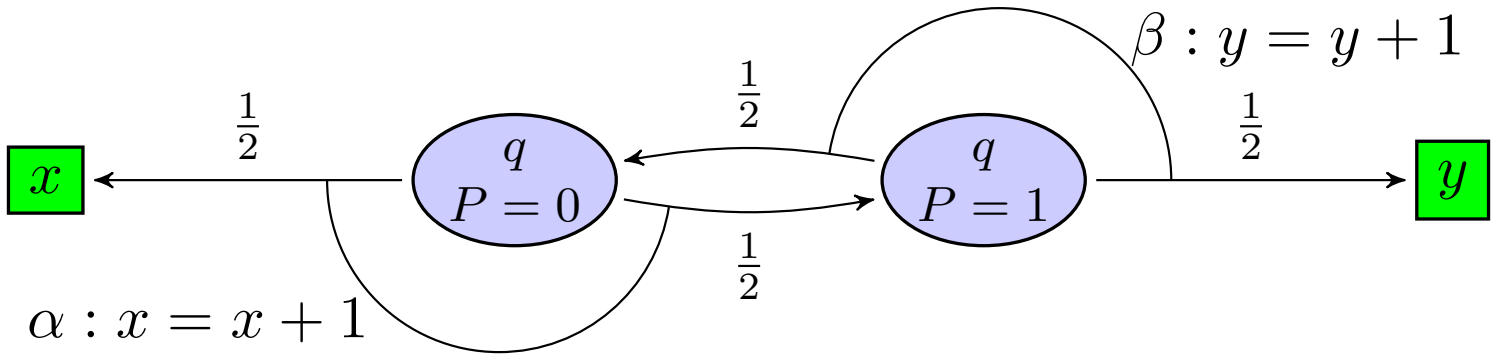
Example



$P(\mathbf{0}) = 0,\ \Delta_\alpha = 1,\ \Delta_\beta = -1$

optimal strategy from $\mathbf{0}$: $(\alpha\beta)^\omega$

The corresponding MC with Alt. Objs.:



Section 2

2-register MDPAOs

Approach

Reduction to p1CA via OC-MDP

Two-Register MDPAOs

- this paper
- minimum expected cost?

One-Counter MDPs

- OC-MDPs *with boundary*
- minimum expected accumulated reward problem is open
- [Brȁdzil et al., '10]

Probabilistic One-Counter Automata

- p1CAs
- no control (i.e., 1 action)
- expected accumulated reward problem is decidable
- [Esparza et al., '05] [Brȁdzil et al., '11]

From 2 registers to 1 counter

Difference Lemma

*Only the **difference** between the registers is relevant to optimality.*

- Such difference can be encoded with a (non-negative) counter and a sign bit: an OC-MDP
- However, min. exp. acc. reward problem for OC-MDPs is open!

From infinitely many strategies to finitely many

Classes of strategies for OC-MDPs:

- **counter-oblivious beyond M:** action depends on current state and counter value *up to* M ($M \in \mathbb{N}$)
  - i.e., all counter values beyond M are considered equivalent
  - given M, there is a **finite number** of strategies counter-oblivious beyond M
  - an OC-MDP + a strategy counter-oblivious beyond M = a p1CA
- **counter-oblivious:** action depends on current state

From infinitely many strategies to finitely many

Classes of strategies for OC-MDPs:

- **counter-oblivious beyond M:** action depends on current state and counter value *up to* M ($M \in \mathbb{N}$)
  - i.e., all counter values beyond M are considered equivalent
  - given M, there is a **finite number** of strategies counter-oblivious beyond M
  - an OC-MDP + a strategy counter-oblivious beyond M = a p1CA
- **counter-oblivious:** action depends on current state

Fact

*For all $M > 0$, strategies **counter-oblivious beyond M** do not suffice to achieve min. exp. acc. reward.*

From infinitely many strategies to finitely many

Bound Theorem

*There is $M > 0$ s.t. strategies **counter-oblivious beyond M** suffice to achieve min. exp. acc. reward in OC-MDPs induced by tie-less MDPAOs.*

Moreover, we can compute M

Great: We only need to consider a finite number of strategies/p1CAs!

The algorithm

Given a tie-less MDPAO with 2 registers:

1. Build the corresponding OC-MDP (difference lemma)
2. Compute the bound M (bound theorem)
3. For all strategies in the OC-MDP that are counter-oblivious beyond M:
   - Build the corresponding p1CA
   - Compute its exp. acc. reward
4. Return the minimum of the above rewards

Theorem

The min. exp. cost of a tie-less 2-register MDPAO can be computed in 2EXPTIME.

Tie-less MDPAOs

An MDPAO is x-*tie-less* if there is a **unique** counter-oblivious strategy (σ_x) having **minimal probability of terminating in** x

A 2-register MDPAO is *tie-less* if it is x-tie-less and y-tie-less

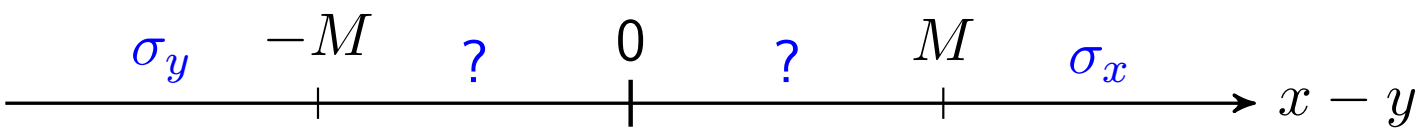
Tie-less MDPAOs

An MDPAO is *x-tie-less* if there is a **unique** counter-oblivious strategy (σ_x) having **minimal probability of terminating in** x

A 2-register MDPAO is *tie-less* if it is x-tie-less and y-tie-less

Lemma

When $x \gg y$, the optimal strategy plays like σ_x.



Corollary

There exists $M > 0$ s.t. the optimal strategy in the induced OC-MDP is counter-oblivious beyond M.

On the likelihood of being tie-less

If an MDPAO is tie-less, all MDPAOs *close enough* are also tie-less

If an MDPAO is *not* tie-less, in all neighborhoods there exists a tie-less MDPAO

Hence, being tie-less is a *robust* property, its opposite is not

(tie-less is *dense*, its opposite is not)

Section 3

General MDPAOs

Approximations

The minimum expected cost with *bounded horizon* is close to the real one

Approximations

The minimum expected cost with *bounded horizon* is close to the real one

Theorem

The min. exp. cost of an MDPAO can be computed up to an additive error ε *in* $|\Gamma||Q|(|\frac{\log \varepsilon}{\log p_M}|\delta_M)^{O(|X|)}$ *time.*

Γ actions

Q states

X registers

p_M maximum probability of *continuation*

δ_M maximum register update

Open problems

- Is the general case decidable? (multiple states, registers, and actions)
- Can tie-less property be applied to OC-MDPs?
  - characterize OC-MDPs admitting optimal strategies that are counter-oblivious beyond some M