CDC'12

# The complexity of Policy Iteration is exponential for discounted Markov Decision Processes

**Romain Hollanders**

Joint work with Raphaël Jungers and Jean-Charles-Delvenne

Prezi

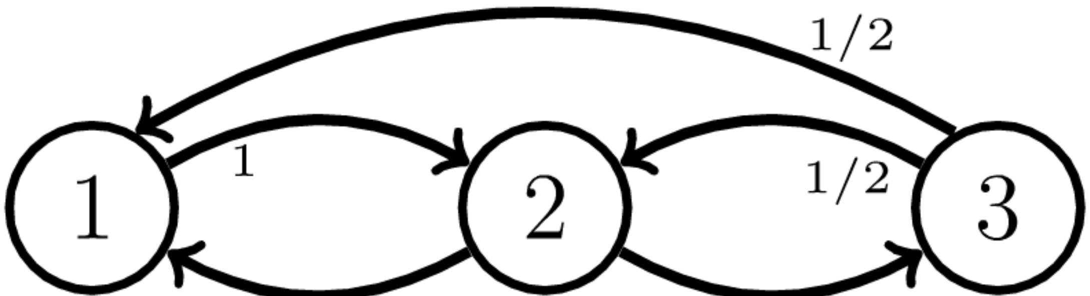
CDC'12

# The complexity of Policy Iteration is exponential for discounted Markov Decision Processes

**Romain Hollanders**

Joint work with Raphaël Jungers and Jean-Charles-Delvenne

# Markov Chains



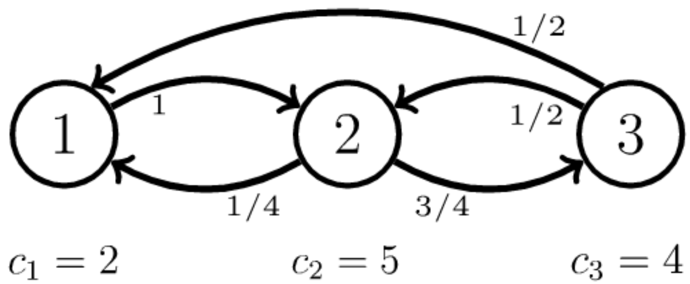

$c_1 = 2$ $c_2 = 5$ $c_3 = 4$

$$P = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

$$p_k^T = p_0^T P^k$$

How much will I pay

# Markov Chains




$$P = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

$$p_k^T = p_0^T P^k$$

$$p_0^T = 1 \quad 0 \quad 0$$
$$p_1^T = 0 \quad 1 \quad 0$$
$$p_2^T = \frac{3}{4} \quad 0 \quad \frac{1}{4}$$
$$p_3^T = \frac{1}{8} \quad \frac{7}{8} \quad 0$$
$$p_4^T = \frac{21}{32} \quad \frac{4}{32} \quad \frac{7}{32}$$
$$\vdots$$

How much will I pay if I start from state 1?

- Total cost

$$x(1) = \sum_{k=0}^{H} p_k^T c$$

- Average cost

$$x(1) = \lim_{H\to\infty} \frac{1}{H} \cdot \sum_{k=0}^{H} p_k^T c$$

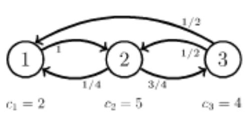
- Discounted cost

$$x(1) = \sum_{k=0}^{\infty} \gamma^k p_k^T c$$

# Markov Decision Processes

One action per state

## Markov Chains




$$P = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

$$p_k^T = p_0^T P^k$$

$$\begin{aligned} p_0^T &= 1 \quad 0 \quad 0 \\ p_1^T &= 0 \quad 1 \quad 0 \\ p_2^T &= \tfrac{3}{4} \quad 0 \quad \tfrac{1}{4} \\ p_3^T &= \tfrac{1}{8} \quad \tfrac{7}{8} \quad 0 \\ p_4^T &= \tfrac{21}{32} \quad \tfrac{4}{32} \quad \tfrac{7}{32} \\ &\vdots \end{aligned}$$

How much will I pay if I start from state 1?

- Total cost

$$x(1) = \sum_{k=0}^{H} p_k^T c$$

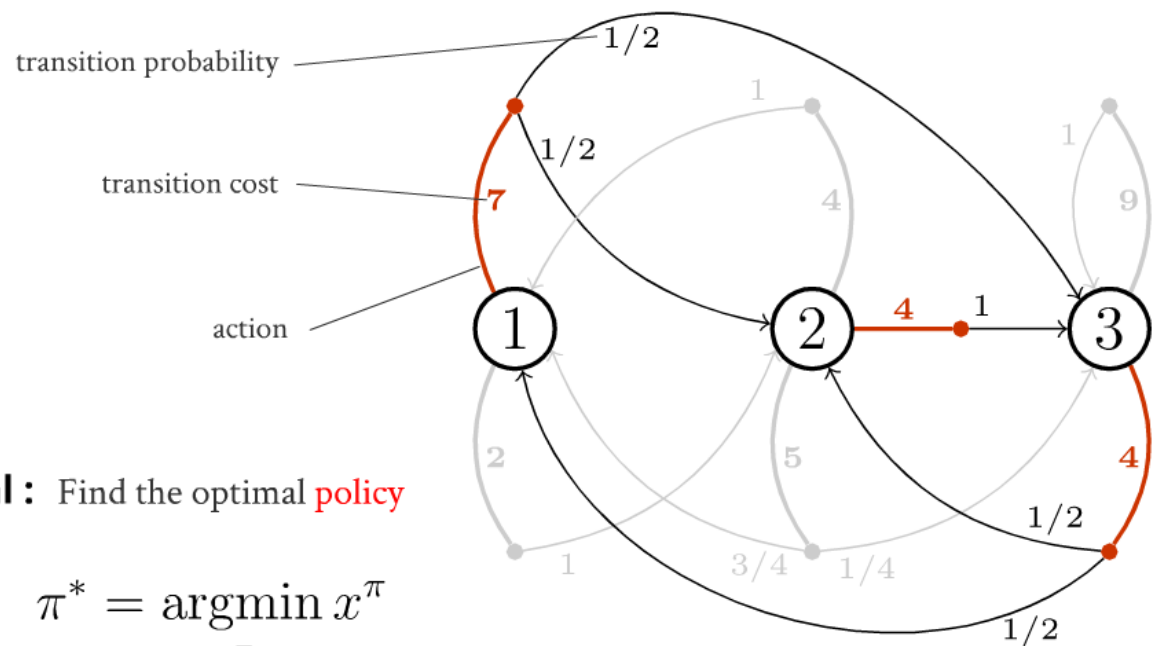
- Average cost

$$x(1) = \lim_{H \to \infty} \frac{1}{H} \cdot \sum_{k=0}^{H} p_k^T c$$

- Discounted cost

$$x(1) = \sum_{k=0}^{\infty} \gamma^k p_k^T c$$

In general

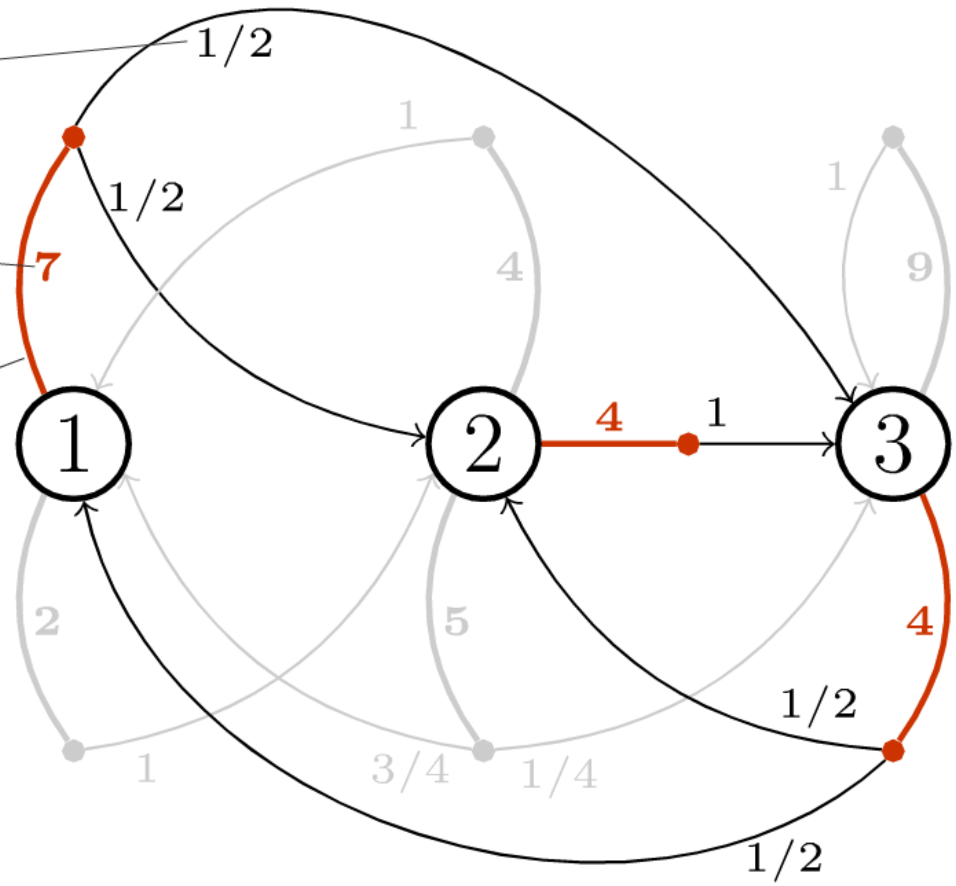



**Goal :** Find the optimal policy

$$\pi^* = \operatorname*{argmin}_{\pi} x^{\pi}$$

The answer depends on the chosen objective function :

- Total cost
- Average cost
- Discounted cost




**Goal :** Find the optimal policy

$$\pi^* = \operatorname*{argmin}_{\pi} x^{\pi}$$

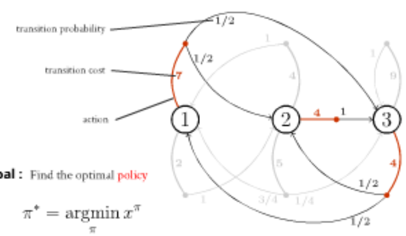
The answer depends on the chosen objective function :

- Total cost
- Average cost
- Discounted cost

# Markov Decision Processes

One action per state

In general


**Goal :** Find the optimal policy

$$\pi^* = \underset{\pi}{\operatorname{argmin}}\, x^\pi$$

The answer depends on the chosen objective function :

- Total cost
- Average cost
- Discounted cost

Resolution

## Linear Programming

## Value Iteration

## Policy Iteration

0. Choose an initial policy $\pi_0$

while $\pi_k \neq \pi_{k-1}$

1. Evaluate $\pi_k$ (Bellman)

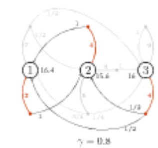
$$x^{\pi_k} = c^{\pi_k} + \gamma\, P^{\pi_k} x^{\pi_k}$$

2. Improve $\pi_k$

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmin}}\; c^\pi + \gamma\, P^\pi x^{\pi_k}$$

$k \leftarrow k+1$

end



# Linear Programming

## Value Iteration

## Policy Iteration

0. Choose an initial policy $\pi_0$

while $\pi_k \neq \pi_{k-1}$

1. Evaluate $\pi_k$ (Bellman)

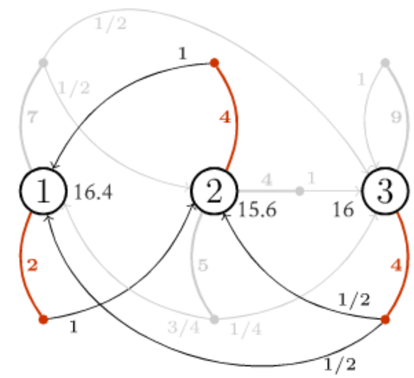
$$x^{\pi_k} = c^{\pi_k} + \gamma \, P^{\pi_k} x^{\pi_k}$$

2. Improve $\pi_k$

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmin}} \; c^{\pi} + \gamma \, P^{\pi} x^{\pi_k}$$

$k \leftarrow k+1$

end




$\gamma = 0.8$

# Policy Iteration

0. Choose an initial policy $\pi_0$

while $\pi_k \neq \pi_{k-1}$

1. Evaluate $\pi_k$ (Bellman)

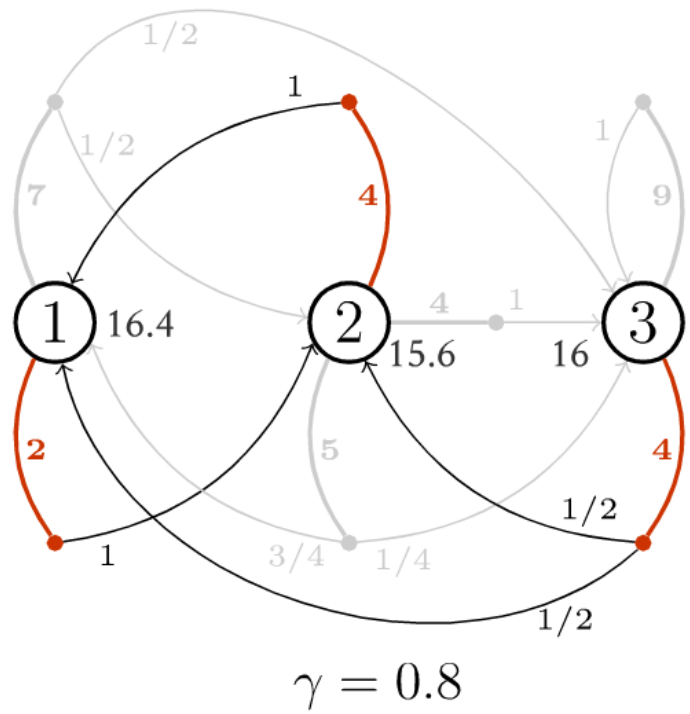
$$x^{\pi_k} = c^{\pi_k} + \gamma\, P^{\pi_k} x^{\pi_k}$$

2. Improve $\pi_k$

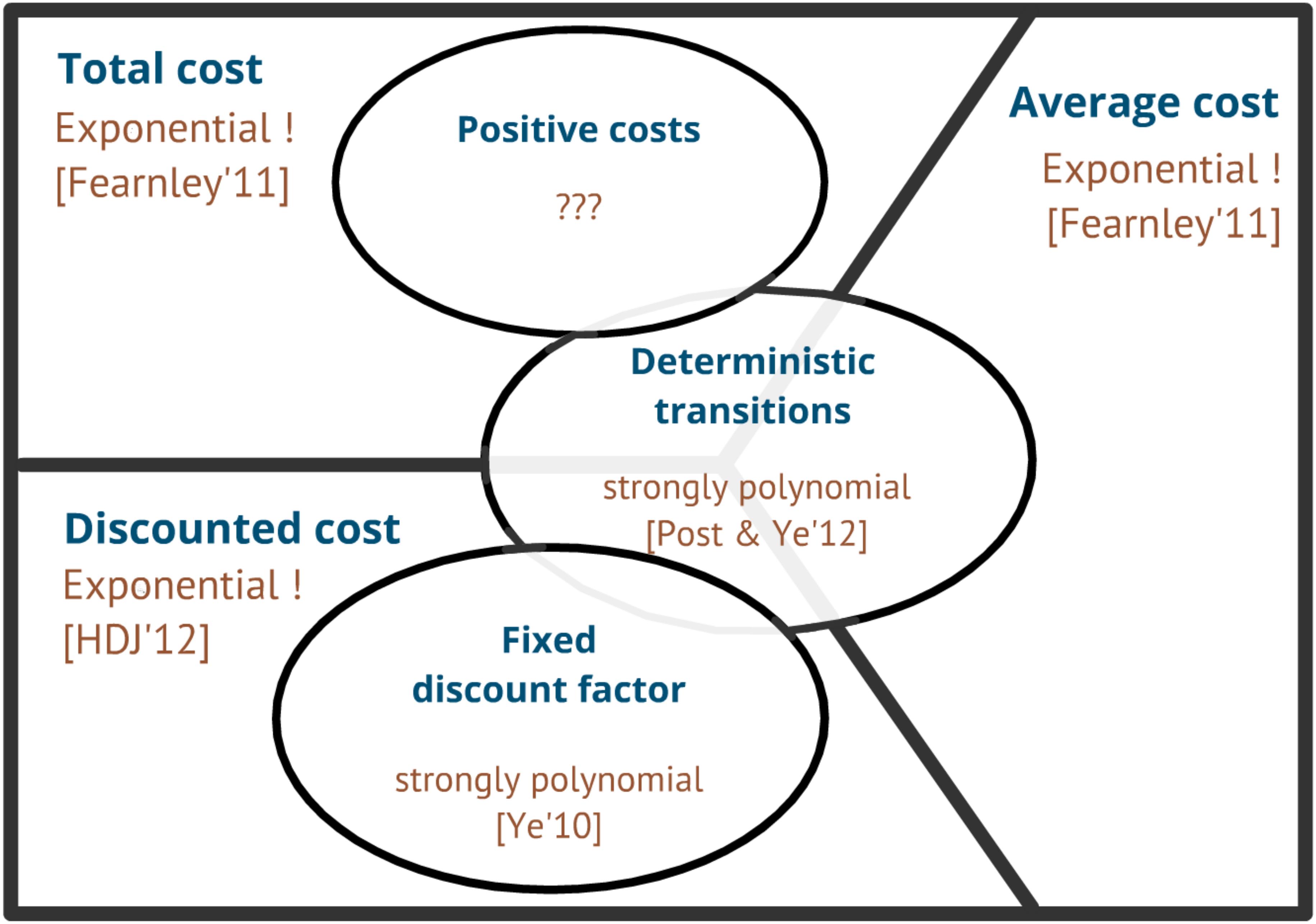
$$\pi_{k+1} = \underset{\pi}{\operatorname{argmin}}\; c^{\pi} + \gamma\, P^{\pi} x^{\pi_k}$$

$k \leftarrow k + 1$

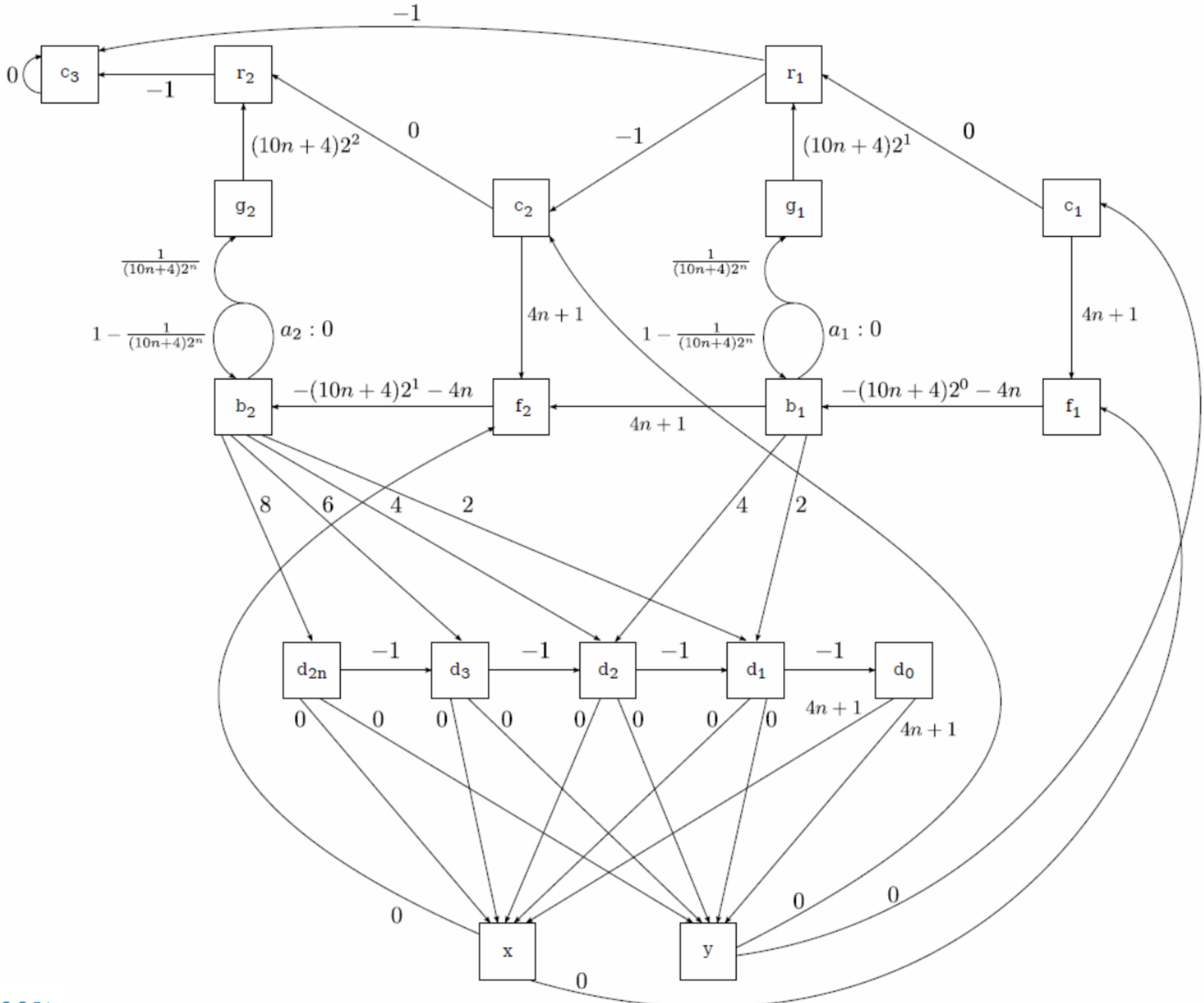
end



# Policy Iteration to solve Markov Decision Processes

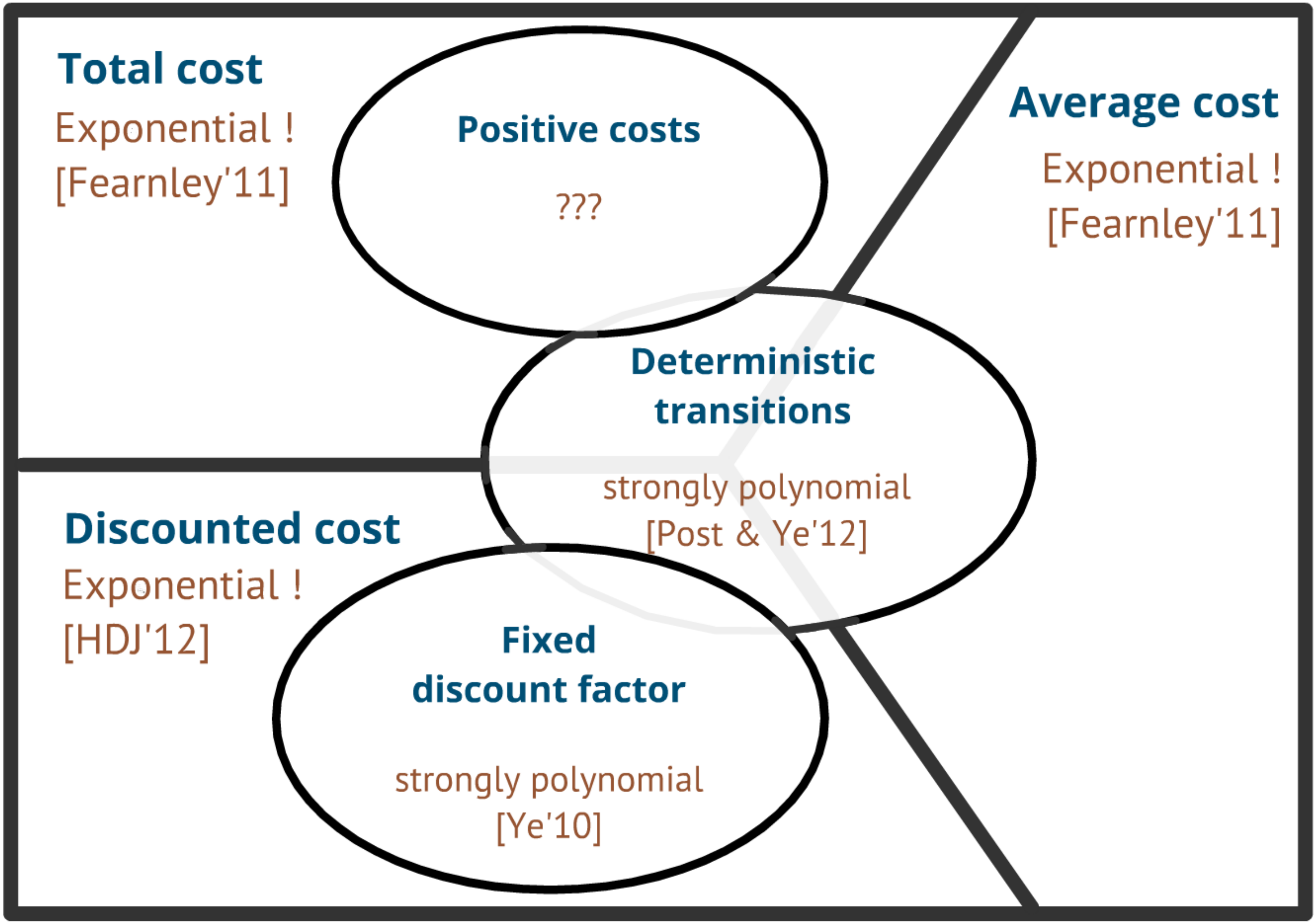



Prezi

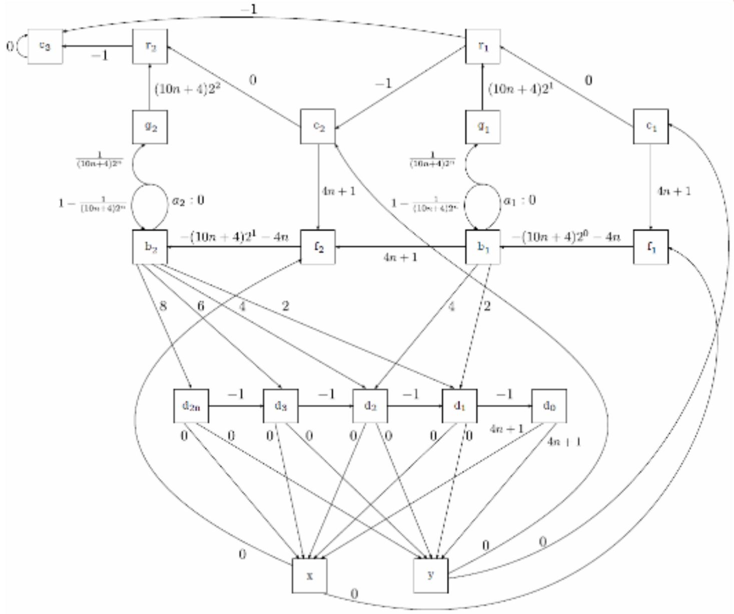


c3
r2
r1
g2
c2
g1
c1
b2
f2
b1
f1
d2n
d3
d2
d1
d0
x
y
−1
0
(10n + 4)2^2
(10n + 4)2^1
1/((10n+4)2^n)
1 − 1/((10n+4)2^n)
a2 : 0
a1 : 0
4n + 1
−(10n + 4)2^1 − 4n
−(10n + 4)2^0 − 4n
8
6
4
2

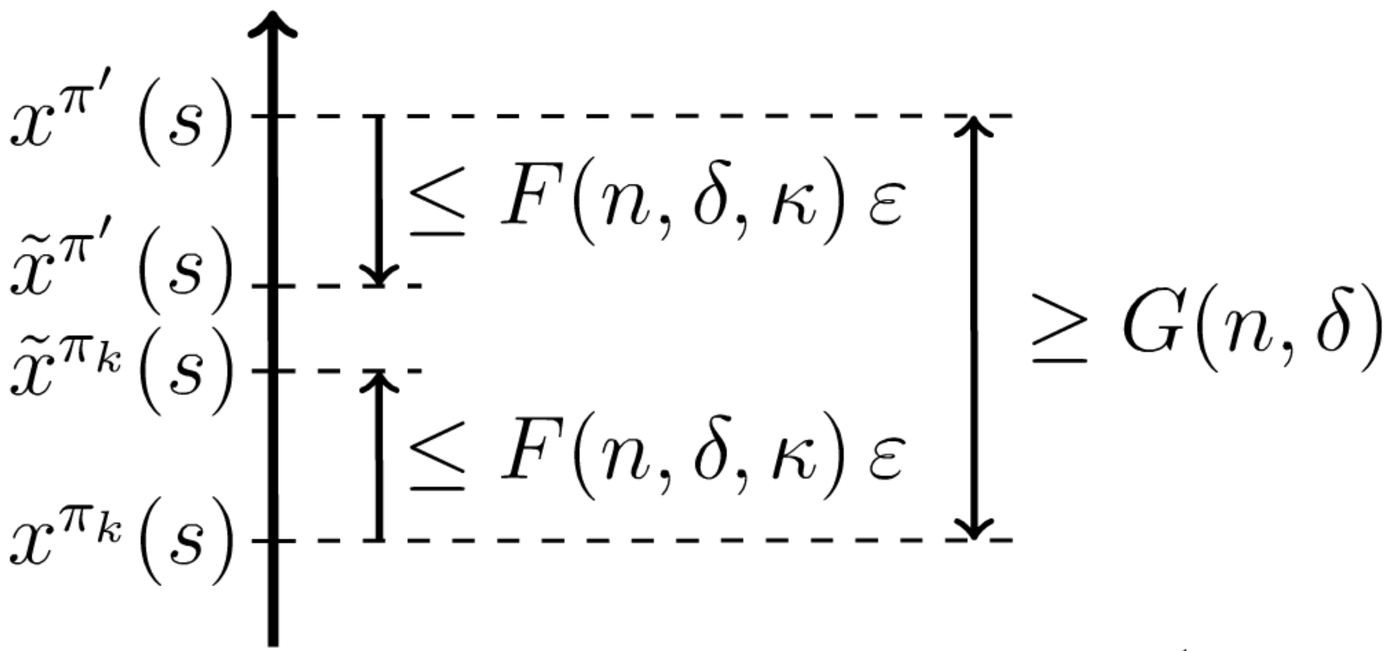
# Policy Iteration to solve Markov Decision Processes



We add discount $\gamma = 1 - \varepsilon$



→ How much perturbation?




→ OK for some $\varepsilon \sim \dfrac{1}{2^{q(n,\delta,\kappa)}}$

# Policy Iteration to solve Markov Decision Processes