TÉCNICO LISBOA

Planning, Learning and Decision Making

Lecture 6. Markov decision problems

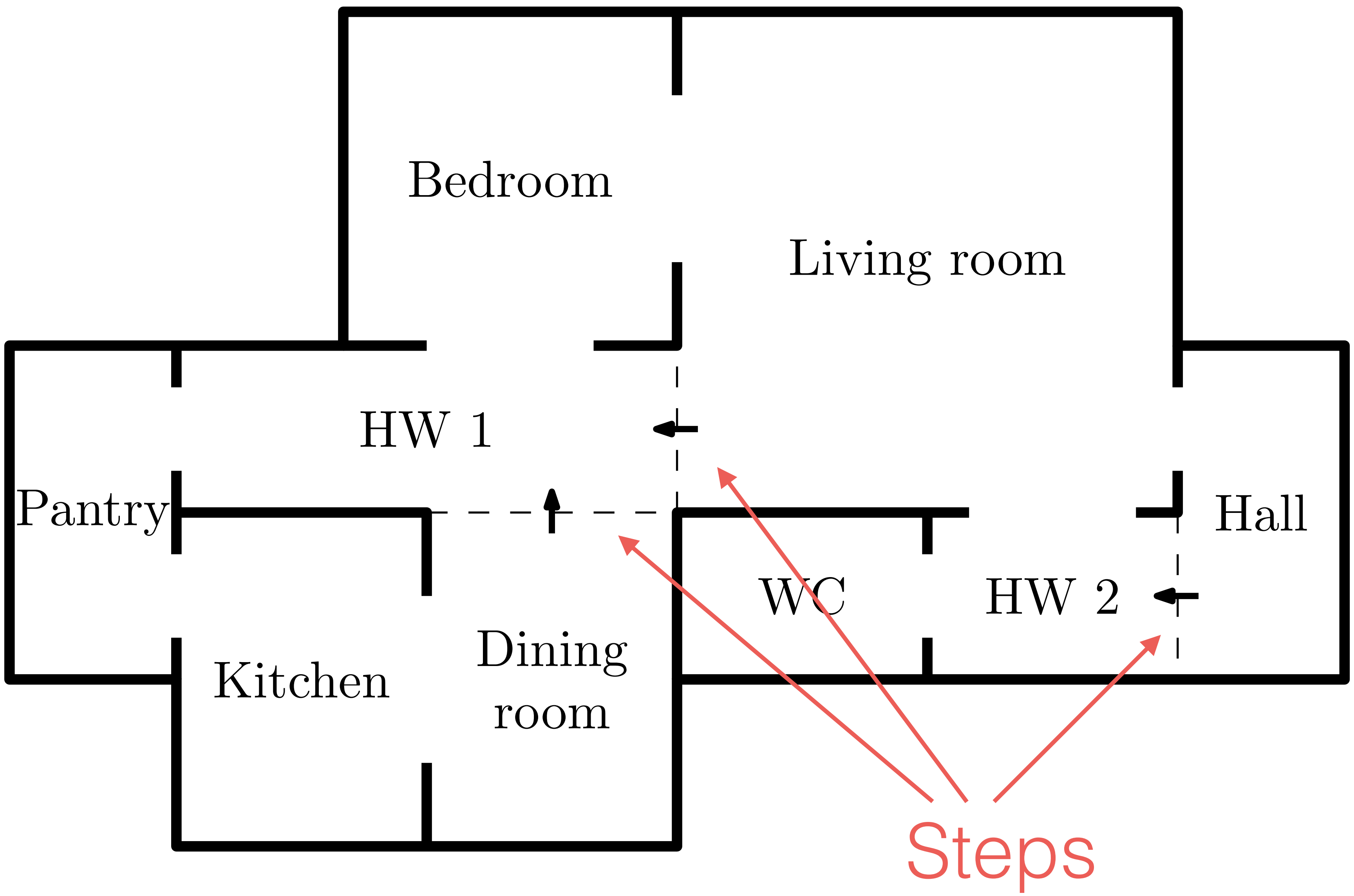
Sequential decision problems

TÉCNICO LISBOA

The household robot

Household robot

- Consider the household



Household robot

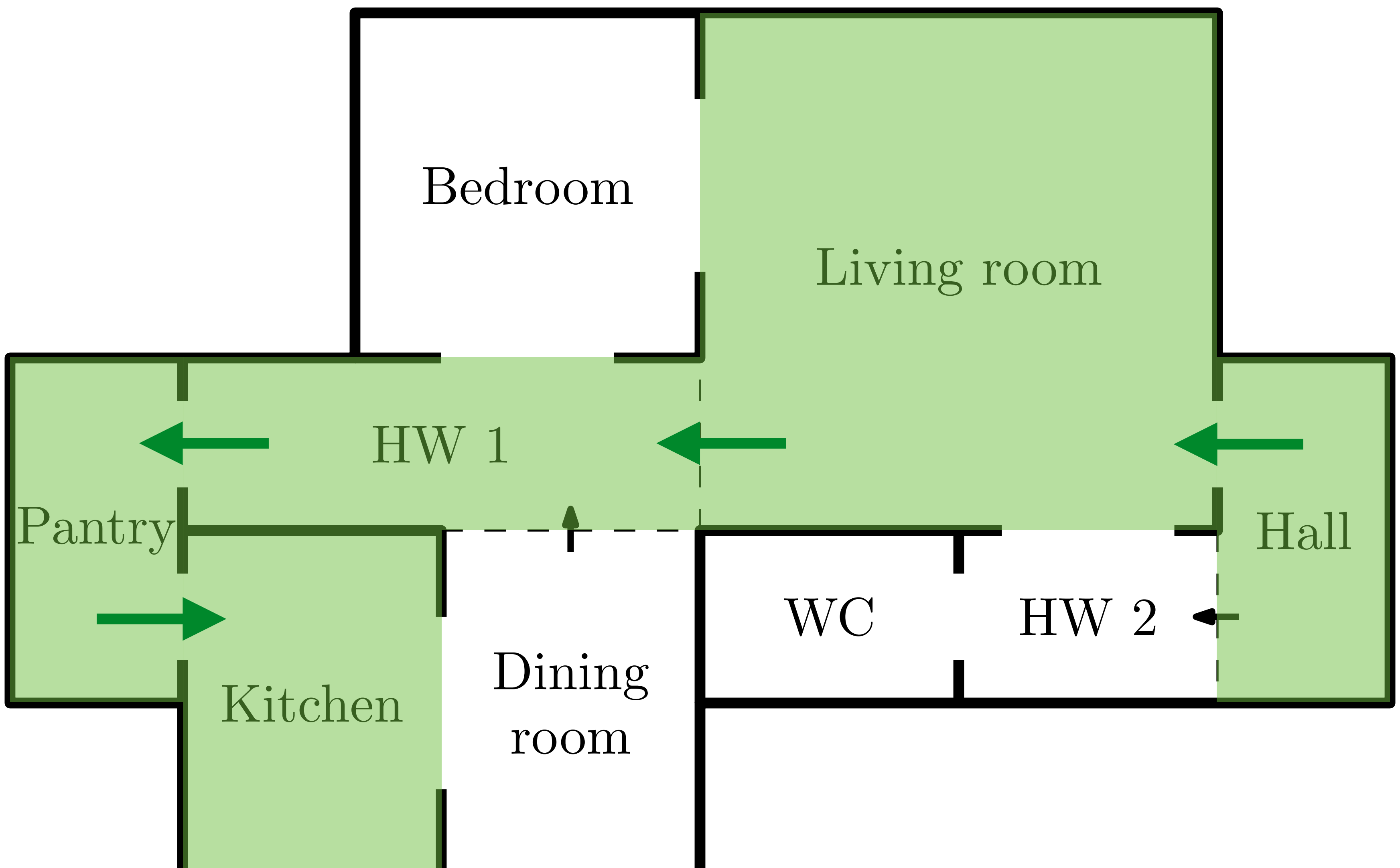
- Robot moves in the environment, assisting human users
- When at the Hall, receives a request from the Kitchen

A single decision

- We can model the problem as a single decision
- Robot must select among several paths

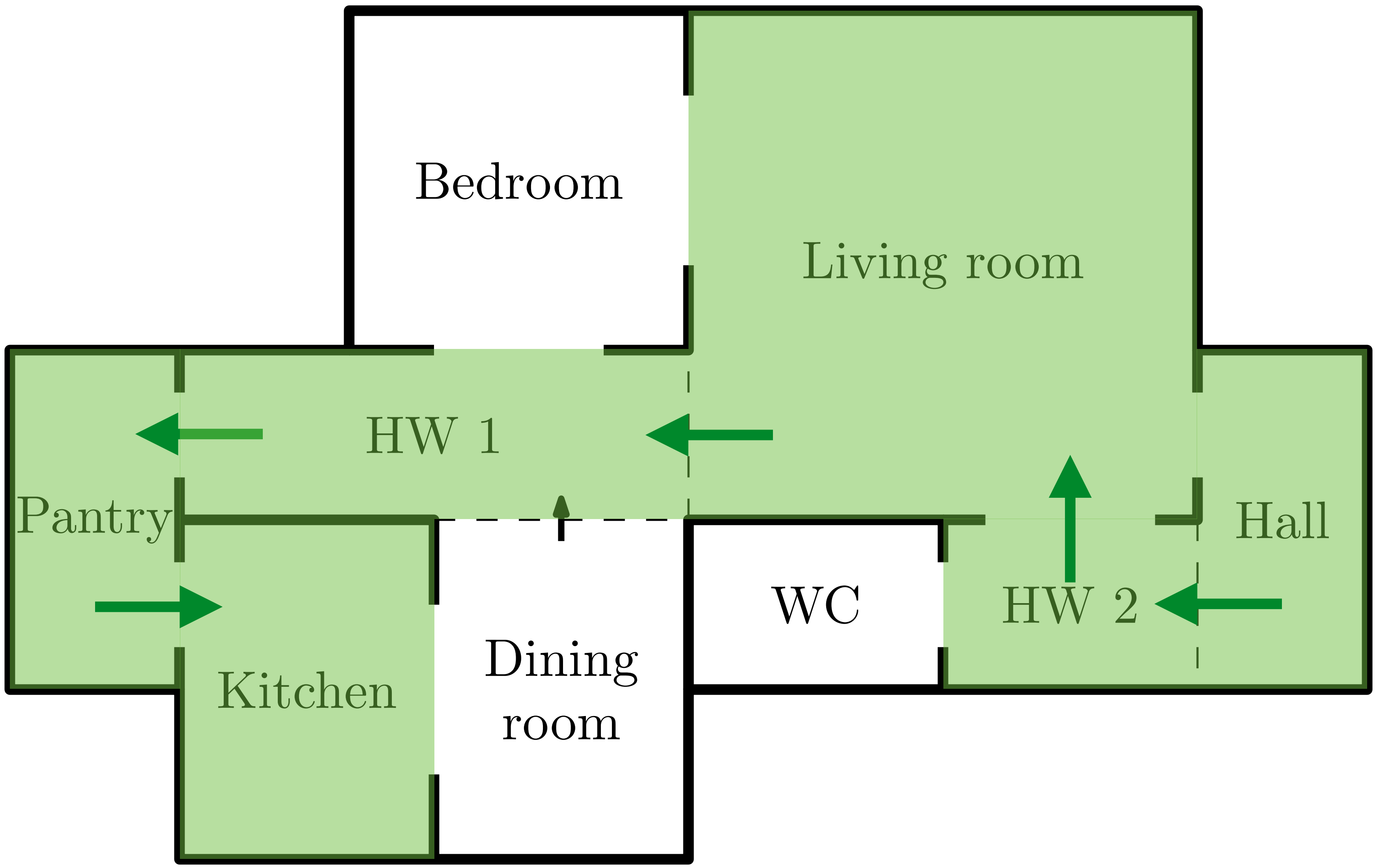
Path A

Hall → Living room → Hallway 1 → Pantry → Kitchen



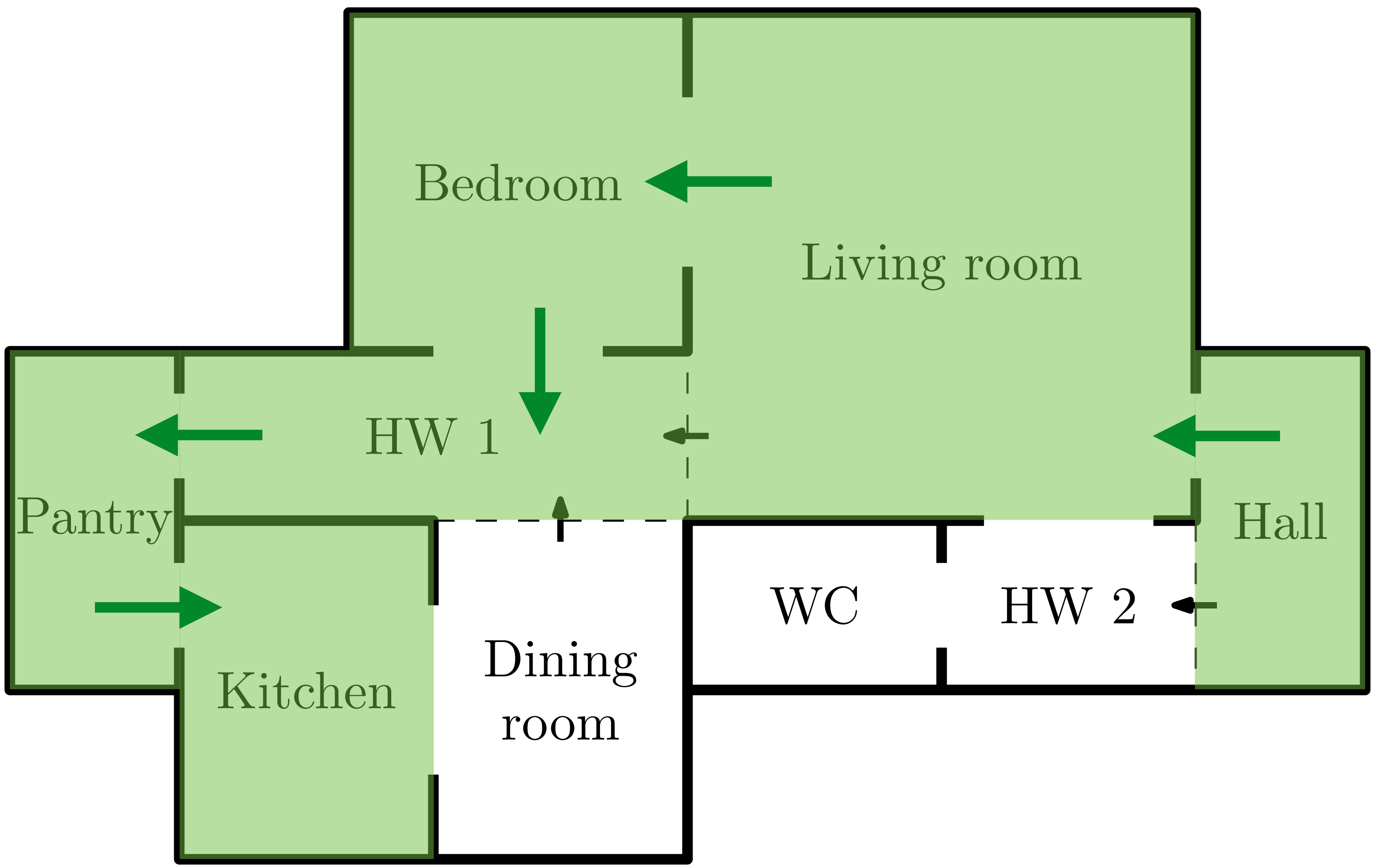
Path B

Hall → Hallway 2 → Living room → Hallway 1 → Pantry → Kitchen



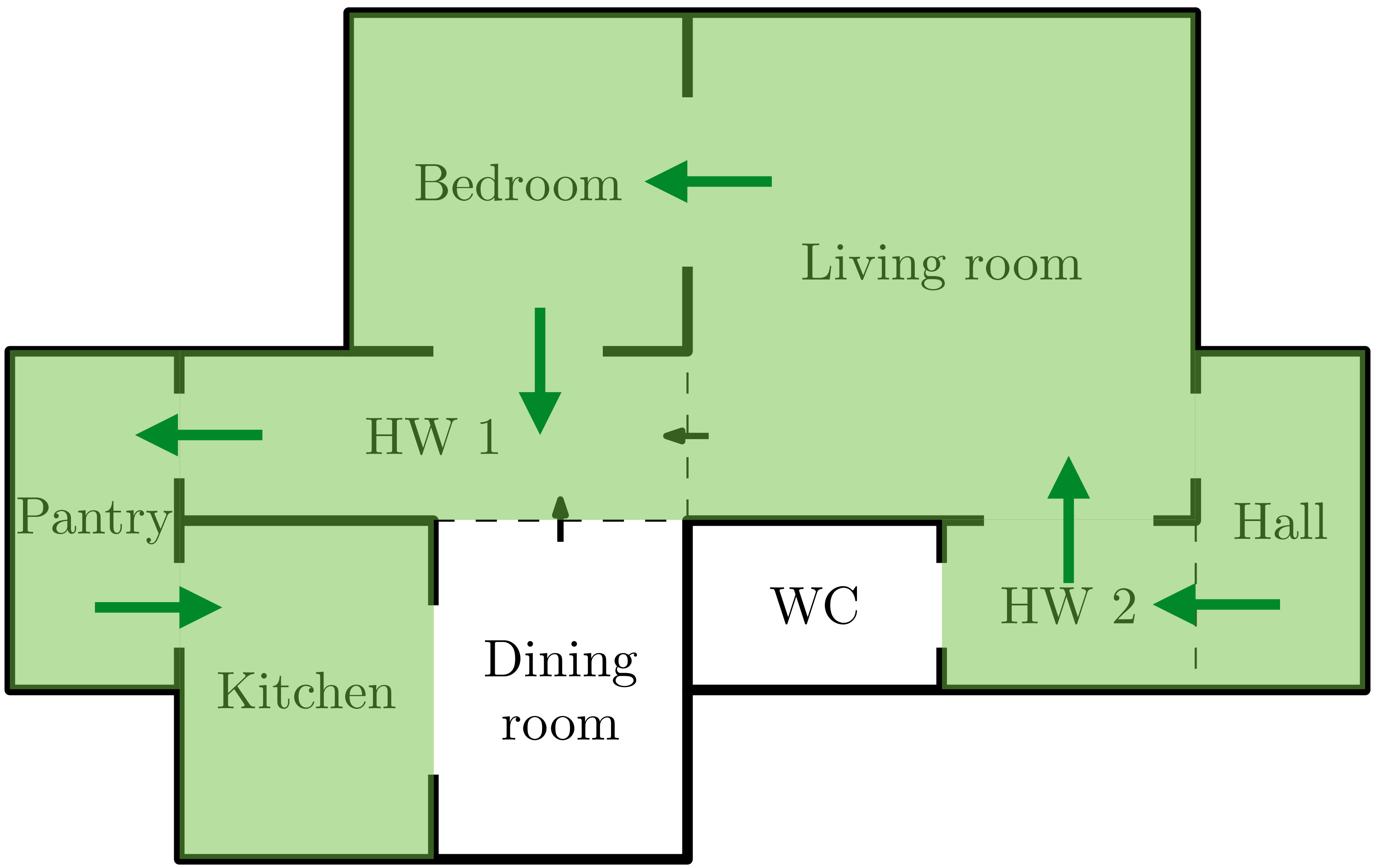
Path C

Hall → Living room → Bedroom → Hallway 1 → Pantry → Kitchen



Path D

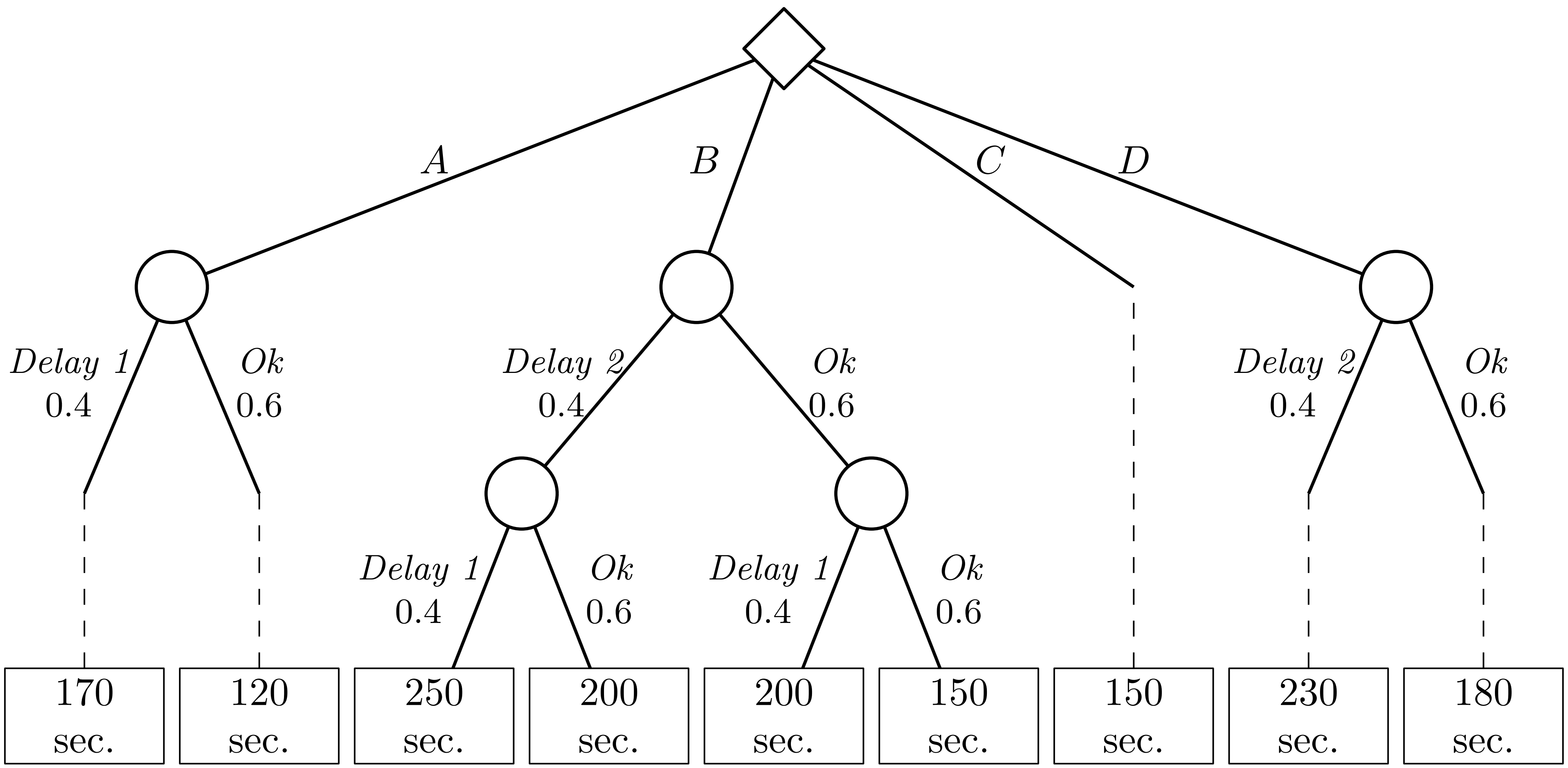
Hall → Hallway 2 → Living room → Bedroom → Hallway 1 → Pantry → Kitchen



A single decision

- Moving between two rooms takes around 30 seconds
- In steps, with a probability 0.4, it takes around 80 seconds

Decision tree



Q(A) = -140 Q(B) = -190 Q(C) = -150 Q(D) = -200

Observation n. 1

Costs vs. utility

- In many problems, we use **negative utilities**
- E.g., the student problem:
  - We used negative utilities to express loss in grades
- E.g., the robot problem:
  - We used negative utilities to express loss in time

Negative utility = cost

The notion of “goal”

- Cost (or utility) implicitly express the **goal** of the decision maker
- We are the **designers** of such goal: we provide the decision-maker/agent with a cost (or utility)
- The cost expresses **our own preferences** (as designers) regarding the behavior of the agent

Observation n. 2

Sequential problems

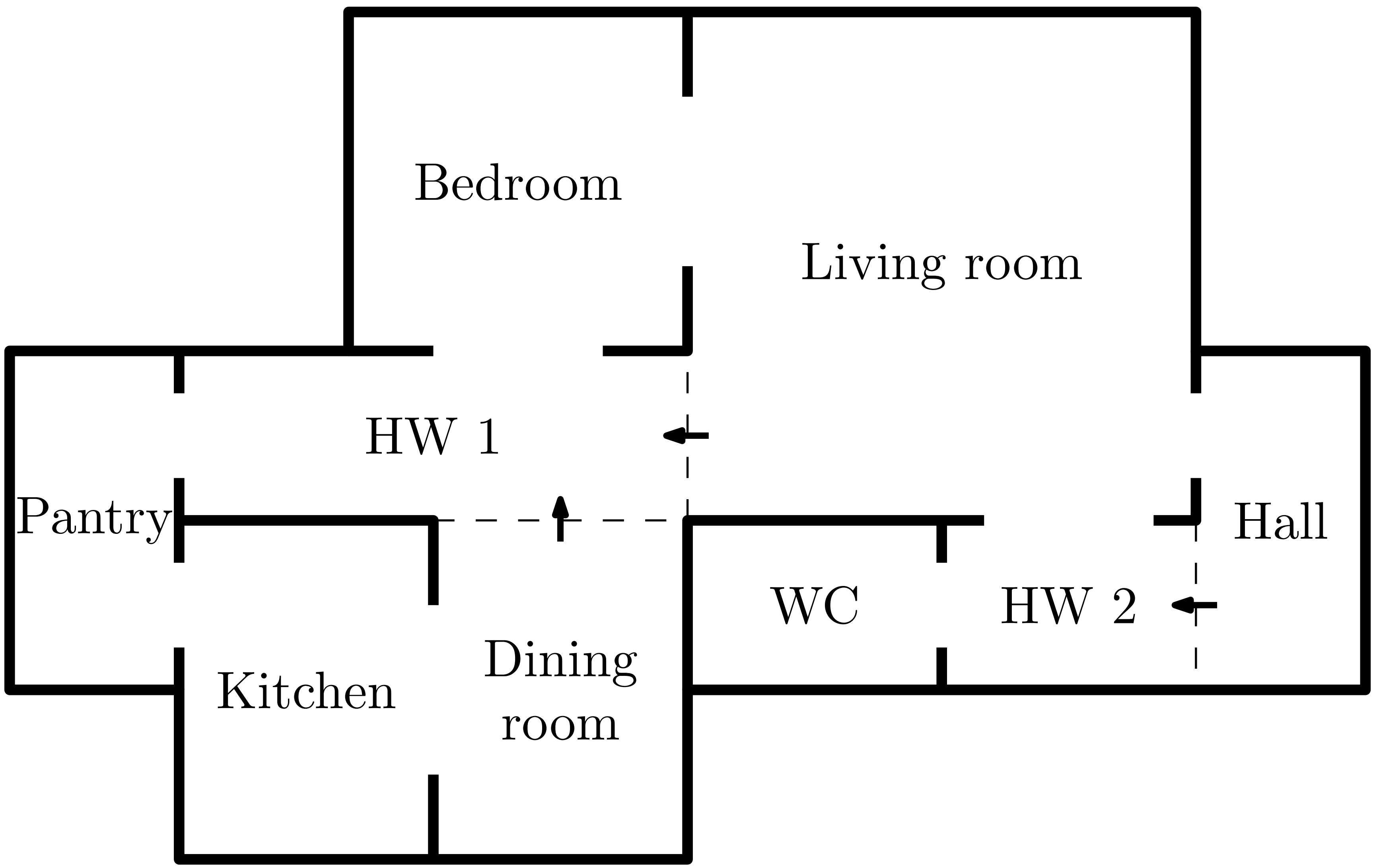
- Sequential problems (like the household robot) are poorly modeled by listing all sequences of actions

Sequence of decisions

The household robot (revisited)

Household robot

- Consider the household



Household robot

- Robot moves in the environment, assisting human users
- When at the Hall, receives a request from the Kitchen

One "movement",
one decision

Sequence of decisions

- At each step, the robot has available a set of actions:

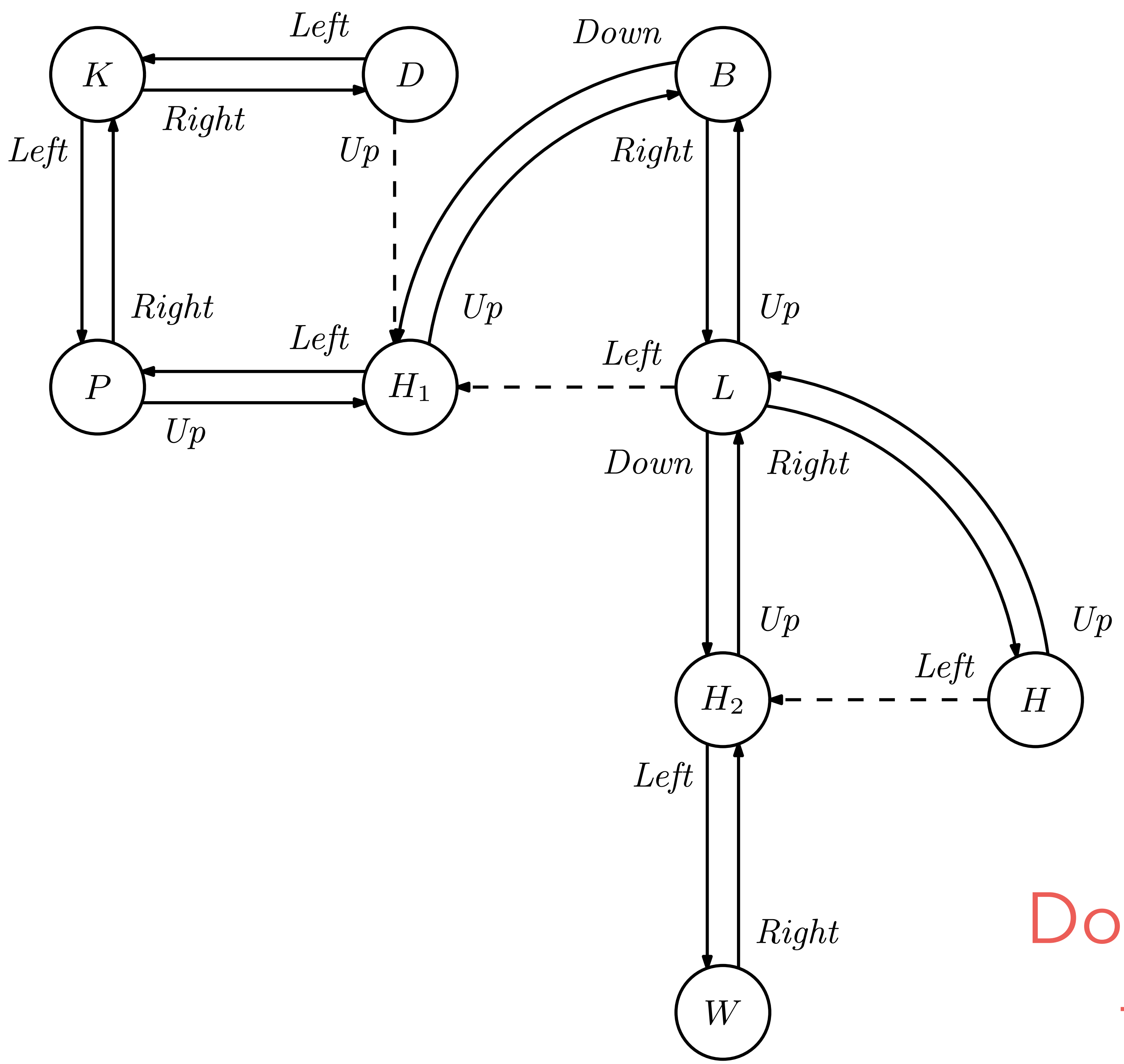
$$\mathscr{A} = \{\text{U(p), D(own), L(eft), R(ight), S(tay)}\}$$

Same symbol
as before

Sequence of decisions

- Motions across a step fail with probability 0.4

Movement of the robot



Does this look familiar?

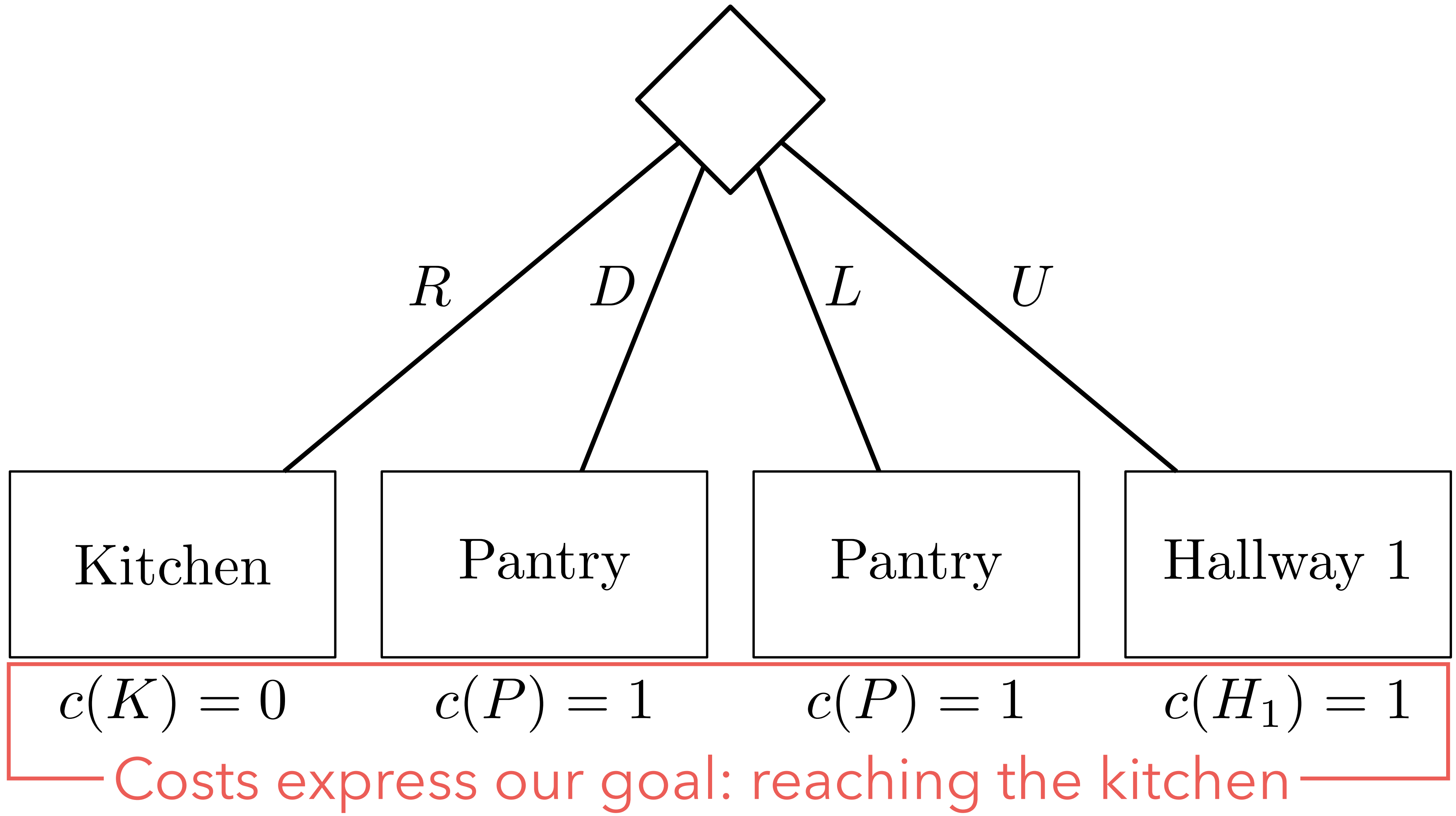
Sequence of decisions

- At each step, what does the decision of the robot depend on?
  - Position of the robot
  - Cost of outcome (1 whenever not in kitchen)

Why?

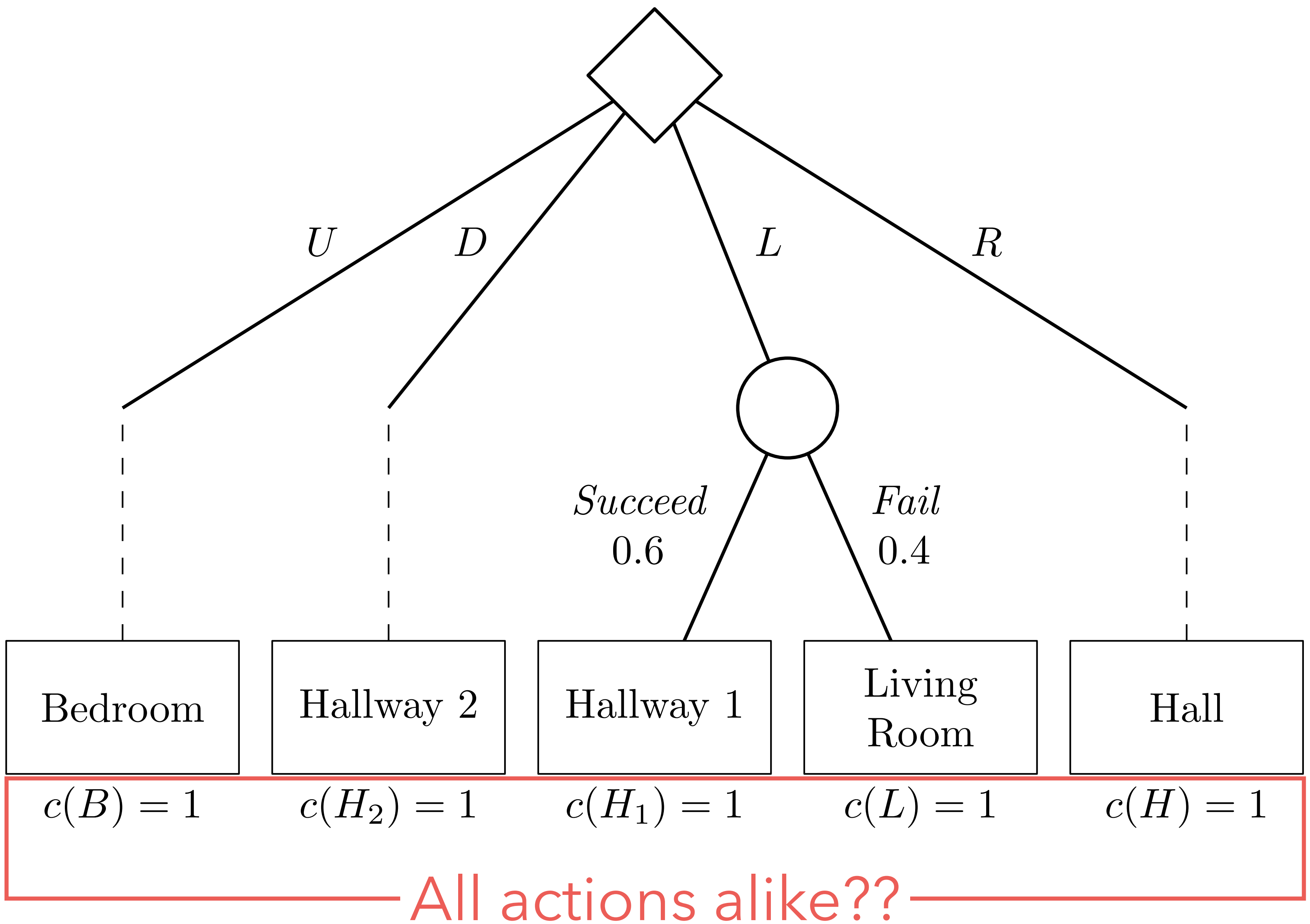
Example

- If the robot is at the Pantry…



Example

- If the robot is in the Living room…

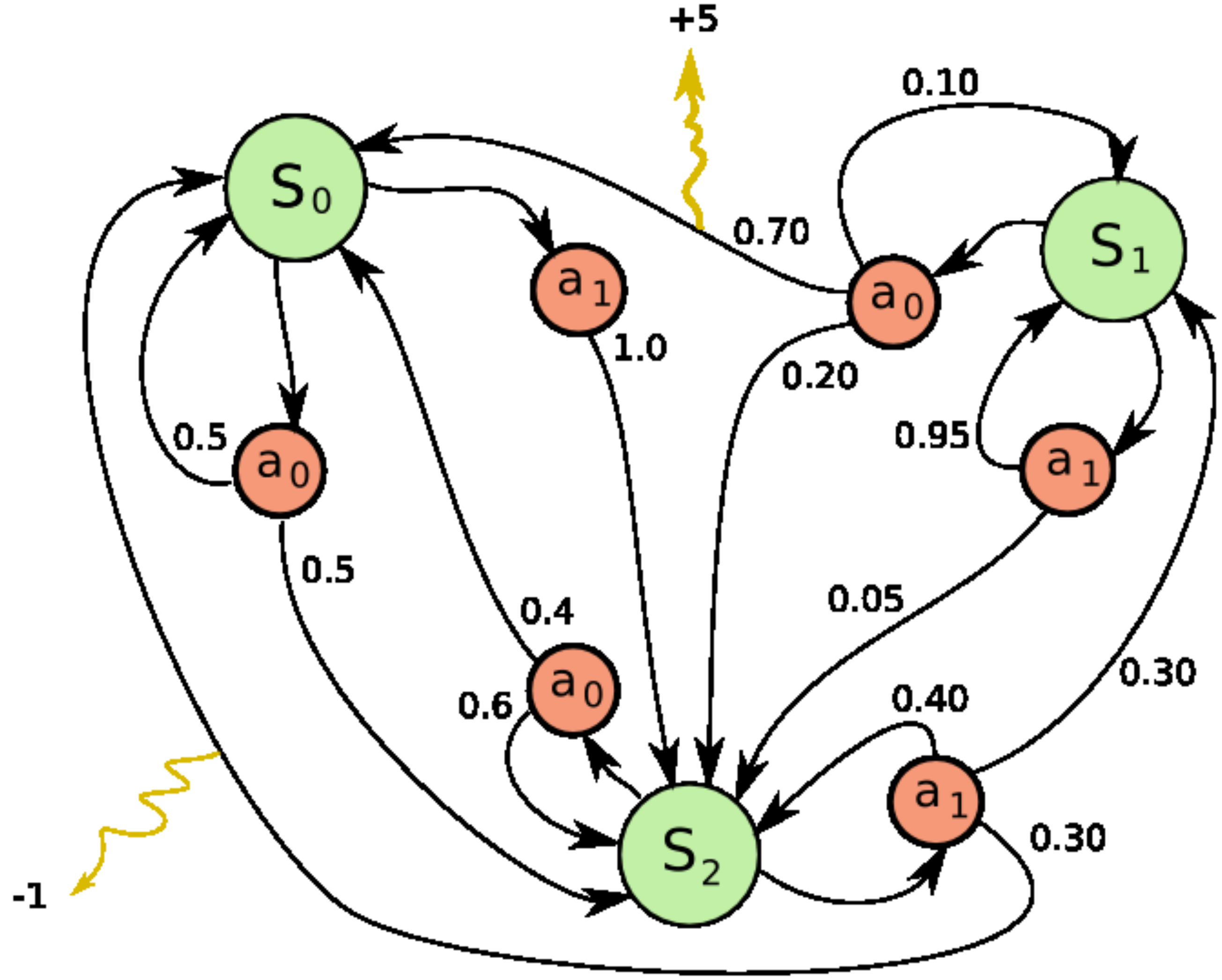


Immediate cost

- The cost used evaluates **instantaneously** the position/action of the robot
- It does not provide **long term** information
- We will call it the **immediate cost**

Two difficulties:

1. How to describe/model such a problem (in general)?
2. How to solve it (in general)?



Markov decision processes

What does the model need?

- Identify the **information** that the decision depends on

States

What does the model need?

- Identify the **actions** that the agent can take

Actions

What does the model need?

- Describe the action **outcomes**

Dynamics

What does the model need?

- Describe the **goal** of the agent

Costs

States

States

- Relevant information for decision making
- We represent the state at time t as x_t
- Set of possible states is $\mathcal{X}$ (finite, most of the time)
- Each step, the agent makes a decision (**decision epoch**)

Actions

Action

- Means by which the agent influences the "environment"
- We represent the action at time t as a_t
- Set of possible actions is $\mathscr{A}$ (finite)

Dynamics

Dynamics

- Describe how the state evolves as a consequence of the agent's actions
- We assume that it verifies the **Markov property**

Markov property

Key Property: Markov property

The state at instant $t + 1$ depends only on the state and action at time step t, i.e.,

$$\mathbb{P}\left[\mathrm{x}_{t+1} = y \mid \mathbf{x}_{0:t} = \boldsymbol{x}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] = \mathbb{P}\left[\mathrm{x}_{t+1} = y \mid \mathrm{x}_t = x_t, \mathrm{a}_t = a_t\right]$$

Controlled Markov chain

Additional assumptions:

- The probabilities $\boxed{\mathbb{P}\left[\mathrm{x}_{t+1} = y \mid \mathrm{x}_t = x, \mathrm{a}_t = a\right]}$ do not depend on *t*

  Transition probability from *x* to *y* given *a*

- For each action $a \in \mathcal{A}$, we store the transition probabilities in a **matrix P**$_a$

$$[\mathbf{P}_a]_{xy} = \mathbb{P}\left[\mathrm{x}_{t+1} = y \mid \mathrm{x}_t = x, \mathrm{a}_t = a\right]$$

Costs

Immediate costs

- Instantaneously evaluates **state and action**
- Represented as a function $c : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$
- For simplicity, we assume that $c(x, a) \in [0, 1]$

Markov decision process

- **Model** for sequential decision processes
- Described by:
  - State space, $\mathscr{X}$
  - Action space, $\mathscr{A}$
  - Transition probabilities, $\{\mathbf{P}_a, a \in \mathscr{A}\}$
  - Immediate cost function, $\mathbf{c}$

Useful notation

- Sometimes we write:
  - $\mathbf{P}(y \mid x, a)$ to denote $[\mathbf{P}_a]_{xy}$

$$\mathbf{P}_a = \begin{bmatrix} \mathbf{P}_a(x_1 \mid x_1) & \mathbf{P}_a(x_2 \mid x_1) & \dots & \mathbf{P}_a(x_N \mid x_1) \\ \mathbf{P}_a(x_1 \mid x_2) & \mathbf{P}_a(x_2 \mid x_2) & \dots & \mathbf{P}_a(x_N \mid x_2) \\ \vdots & & \ddots & \vdots \\ \mathbf{P}_a(x_1 \mid x_N) & \mathbf{P}_a(x_2 \mid x_N) & \dots & \mathbf{P}_a(x_N \mid x_N) \end{bmatrix}$$

Useful notation

- Sometimes we write:
  - **C** to denote the cost matrix, with $[\mathbf{C}]_{xa} = c(x, a)$

$$\mathbf{C} = \begin{bmatrix} c(x_1, a_1) & c(x_1, a_2) & \dots & c(x_1, a_M) \\ c(x_2, a_1) & c(x_2, a_2) & \dots & c(x_2, a_M) \\ \vdots & & \ddots & \vdots \\ c(x_N, a_1) & c(x_N, a_2) & \dots & c(x_N, a_M) \end{bmatrix}$$

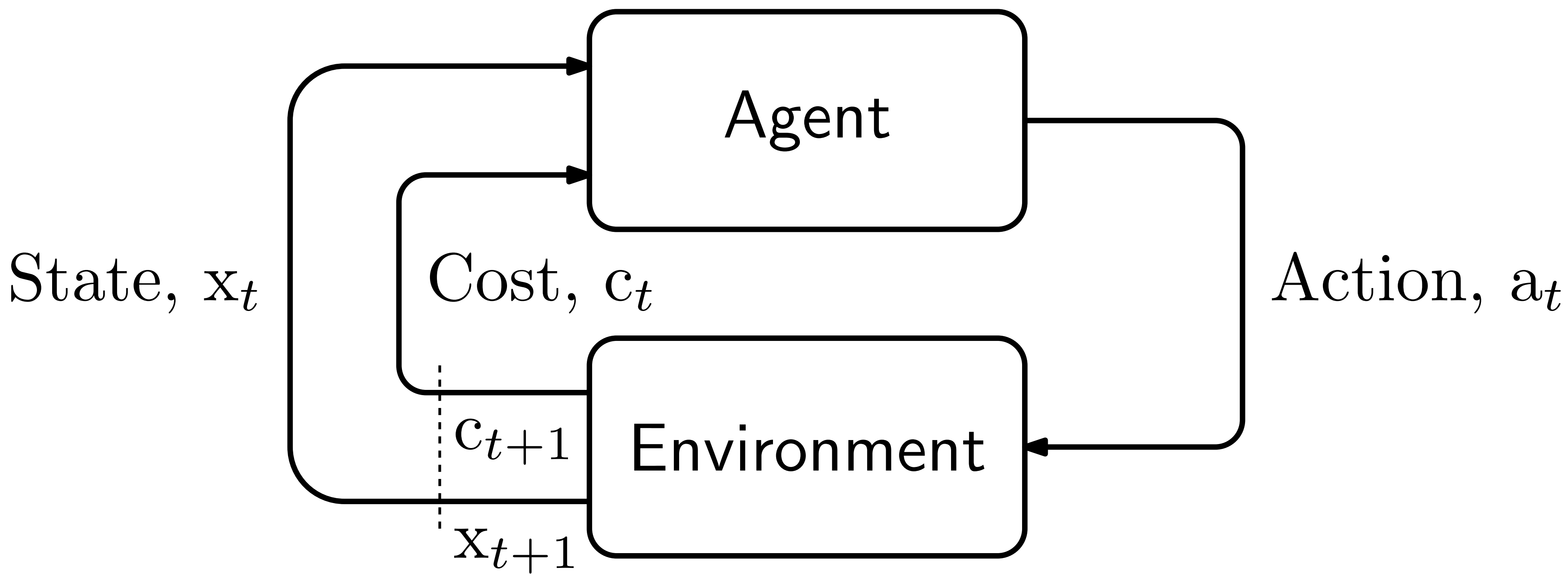
TÉCNICO LISBOA

Useful notation

- Sometimes we write:
  - $\mathbf{C}_{:,a}$ to denote the (column) vector with x component $c(x, a)$

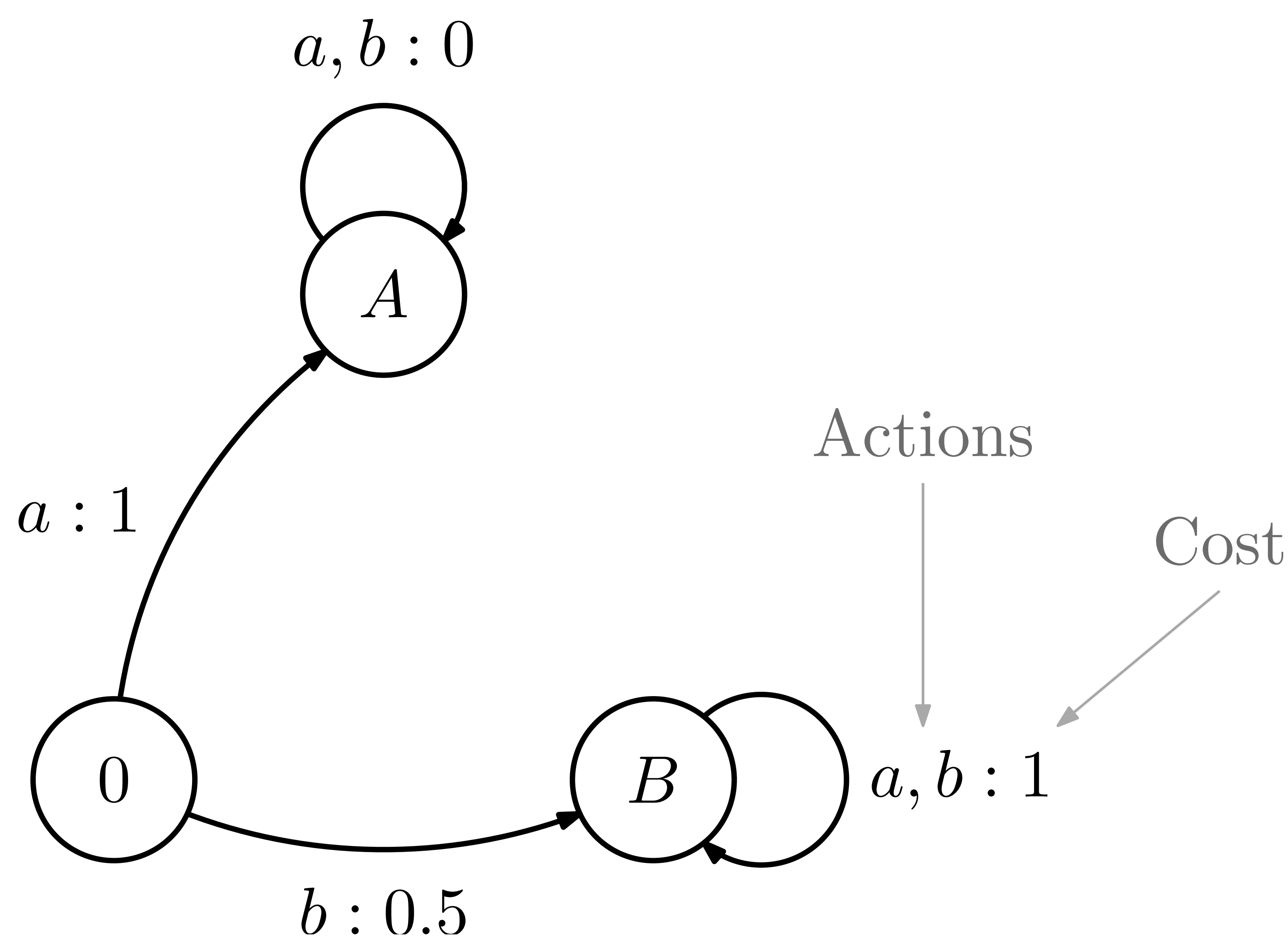
$$\begin{bmatrix} c(x_1, a_1) & \boxed{c(x_1, a_2)} & \dots & c(x_1, a_M) \\ c(x_2, a_1) & c(x_2, a_2) & \dots & c(x_2, a_M) \\ \vdots & & \ddots & \vdots \\ c(x_N, a_1) & c(x_N, a_2) & \dots & c(x_N, a_M) \end{bmatrix} \longrightarrow \mathbf{C}_{:,a} = \begin{bmatrix} c(x_1, a) \\ c(x_2, a) \\ \vdots \\ c(x_N, a) \end{bmatrix}$$

Markov decision process



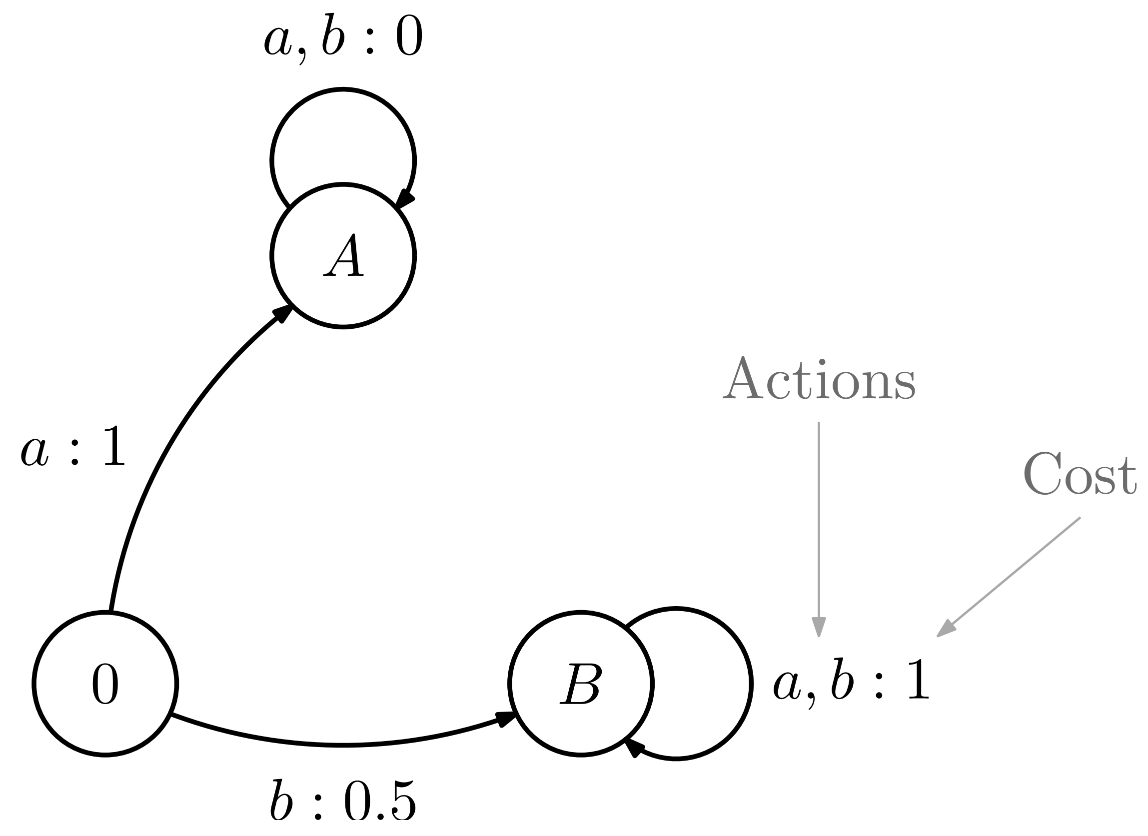
Examples

Example 1



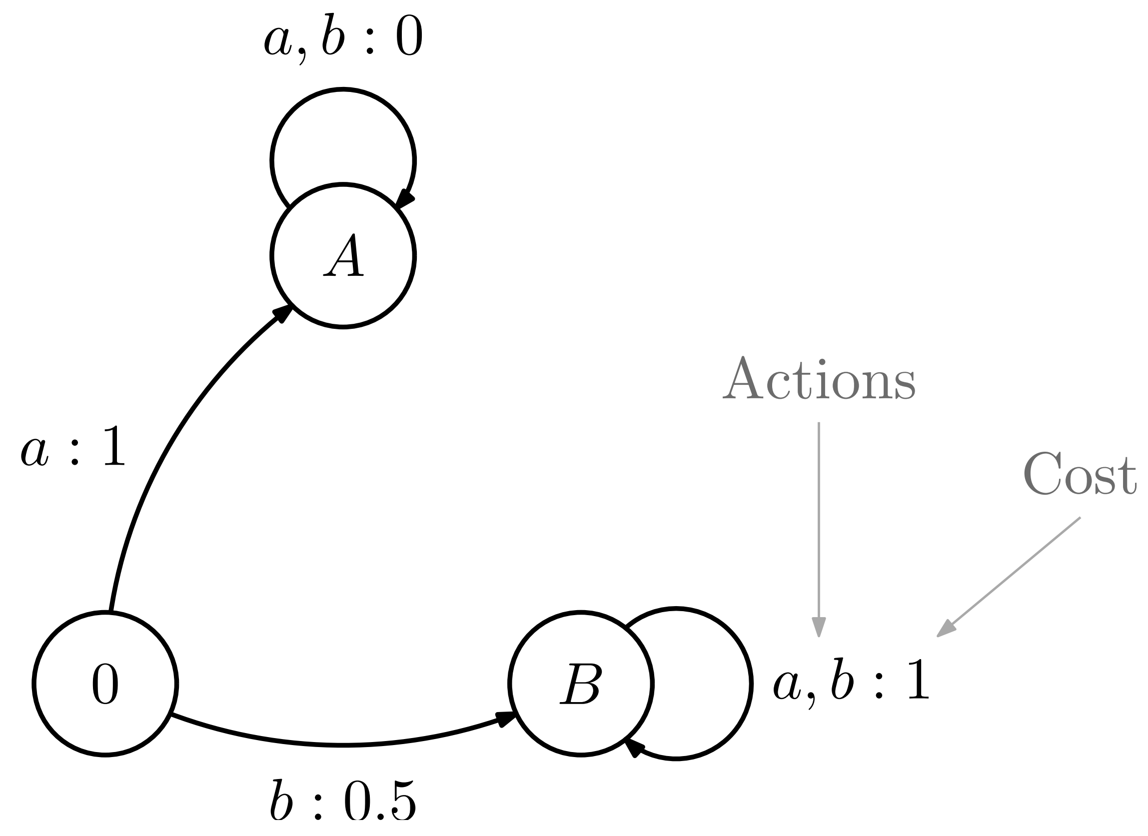
Model definition

- States:
  - $\mathscr{X} = \{0, A, B\}$



Model definition

- Actions:
  - $\mathscr{A} = \{a, b\}$

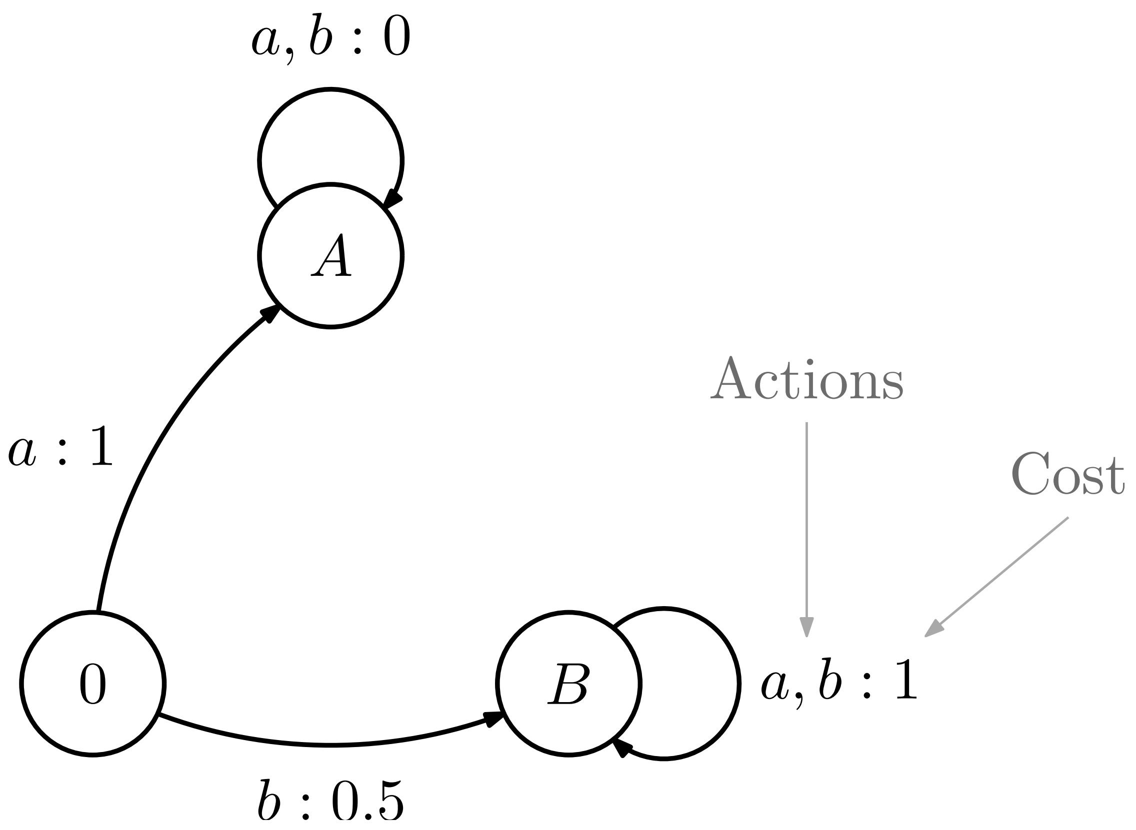


Model definition

- Transition probabilities:

$$\mathbf{P}_a = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

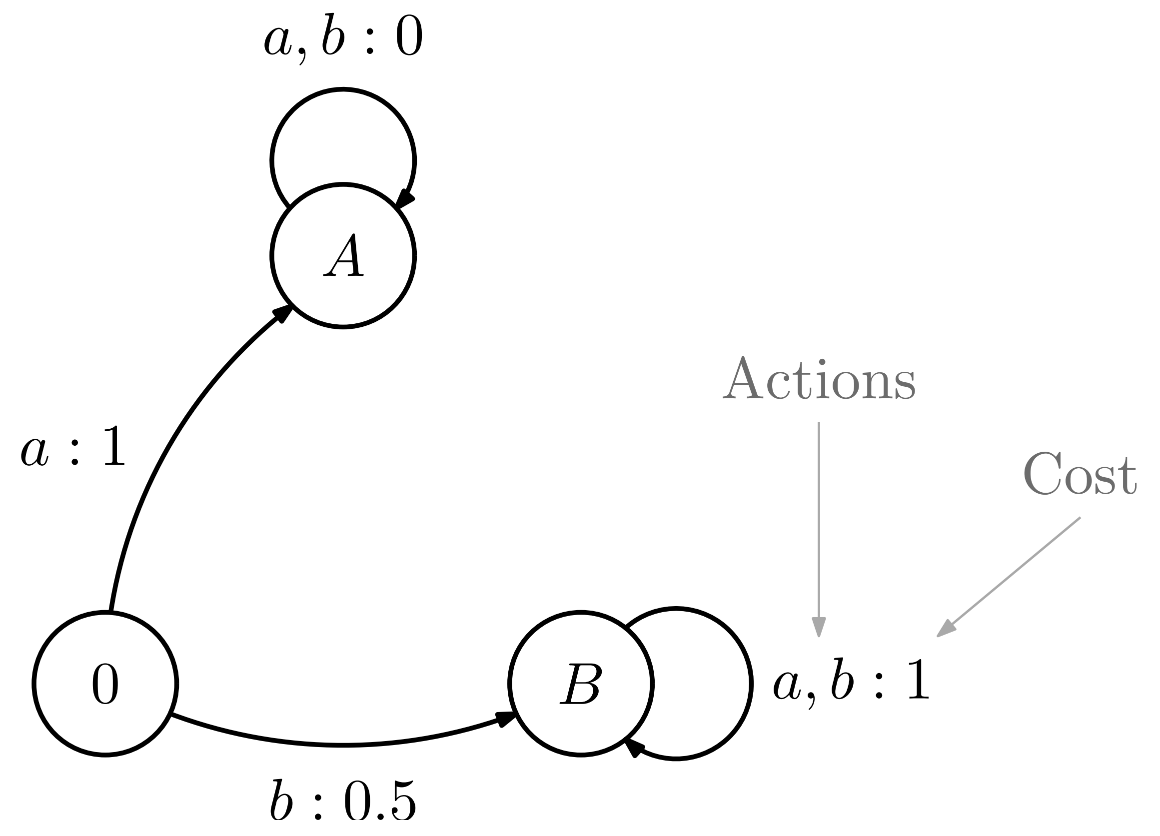
$$\mathbf{P}_b = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



Model definition

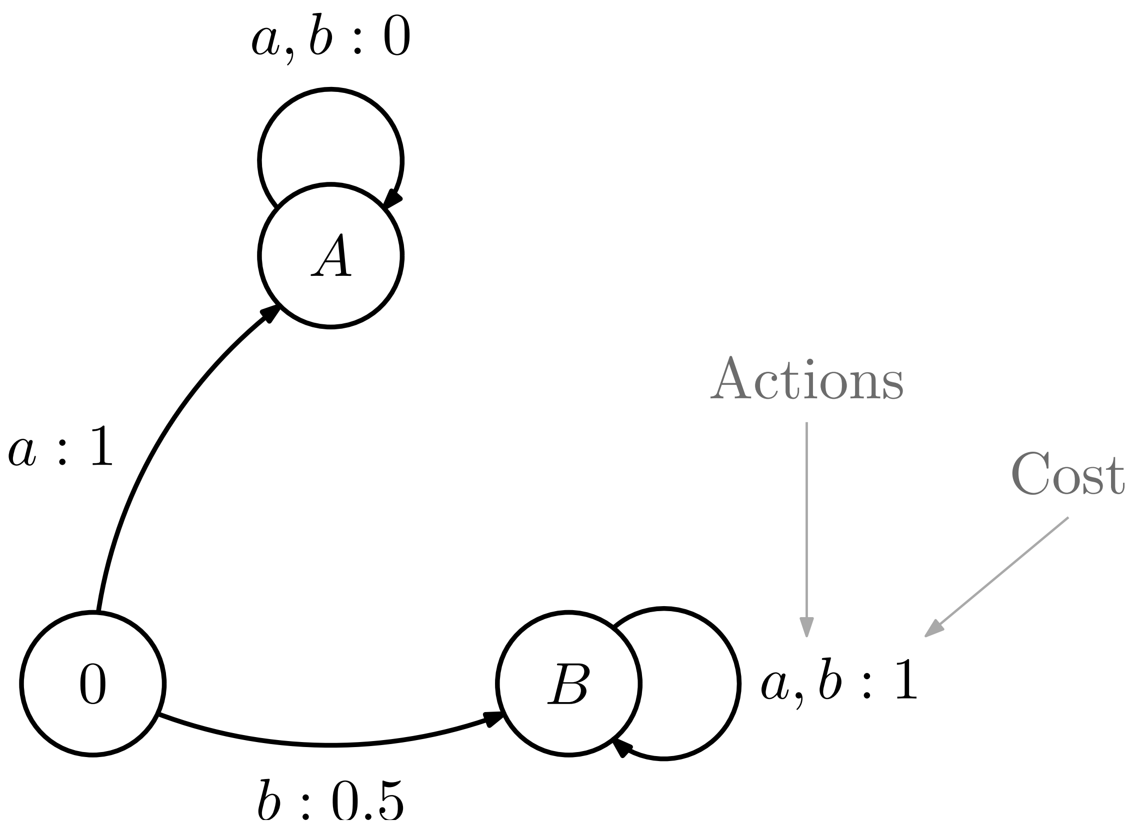
- Cost:

$$\mathbf{C} = \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$



What is the best decision?

- Depends on what "best" means
    - If single decision, then best is *b*
    - If multiple decisions, then best is *a*



Example 2

- A company wants to hire a computer engineer
- After initial trial, *N* candidates are selected for interview

Example 2

- Candidates are interviewed sequentially
- Order of the candidates for interview was selected randomly

TÉCNICO LISBOA

Example 2

- Manager must decide, after each interview, whether to hire or not (no second chances)
- Manager knows whether an interviewed candidate is the best so far
- If no candidate has been hired in the meantime, candidate N is necessarily hired

How to model this?

- What are the states?
  - What is relevant for the manager's decision?
    - Current candidate best so far or not
    - How many candidates have been interviewed/are missing
  - State-space:

Not best so far

- $\mathscr{X} = \{(\text{B}, 1), (\text{B}, 2), (\neg\text{B}, 2), \ldots, (\text{B}, N), (\neg\text{B}, N), \text{H}\}$

Best so far

Hired

How to model this?

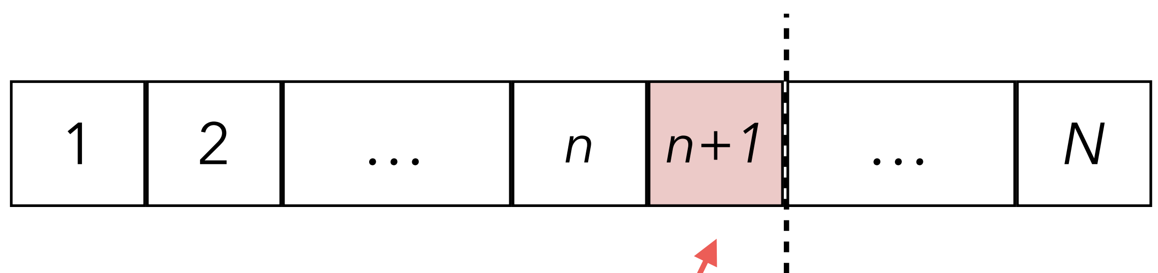
- What are the actions?
  - $\mathscr{A} = \{\text{H}, \neg\text{H}\}$

How to model this?

- Transition probabilities:
  - … tough!

How to model this?

- Transition probabilities:
  - What is the probability that the $(n + 1)$th candidate is the best so far?



Probability that the best among first $n+1$ candidates is candidate $n+1$?

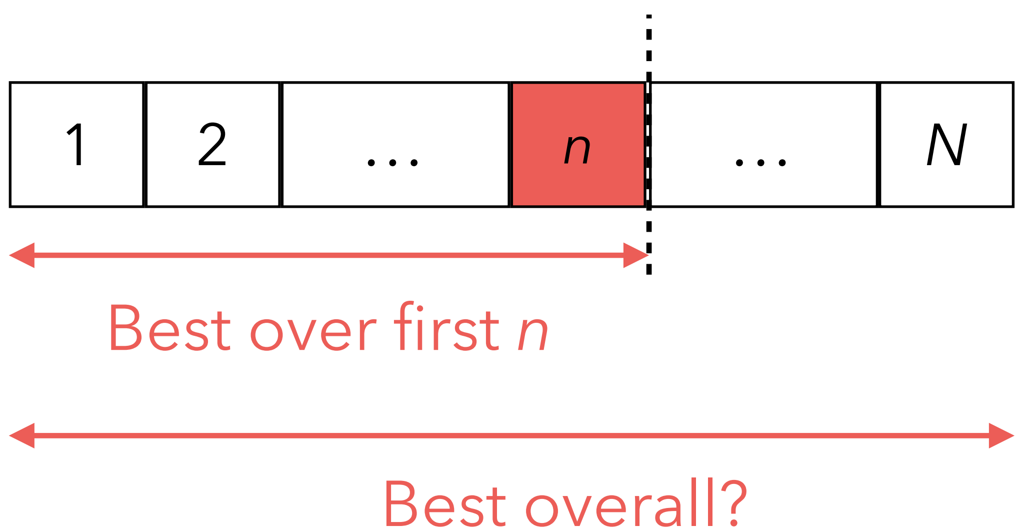
$\boxed{1 / (n + 1)}$

How to model this?

- Transition probabilities:
  - What is the probability that the $(n + 1)$th candidate is the best so far?
    - $1 / (n + 1)$
  - What's the probability that the $(n + 1)$th candidate is **not** the best so far?
    - $n / (n + 1)$

How to model this?

- Cost:
  - … hiring a guy who is not the best so far incurs maximum cost (clearly, that guy is not the best)
  - … what about hiring a guy who is the best so far after n interviews?
    - How likely is it that it is not the best overall?



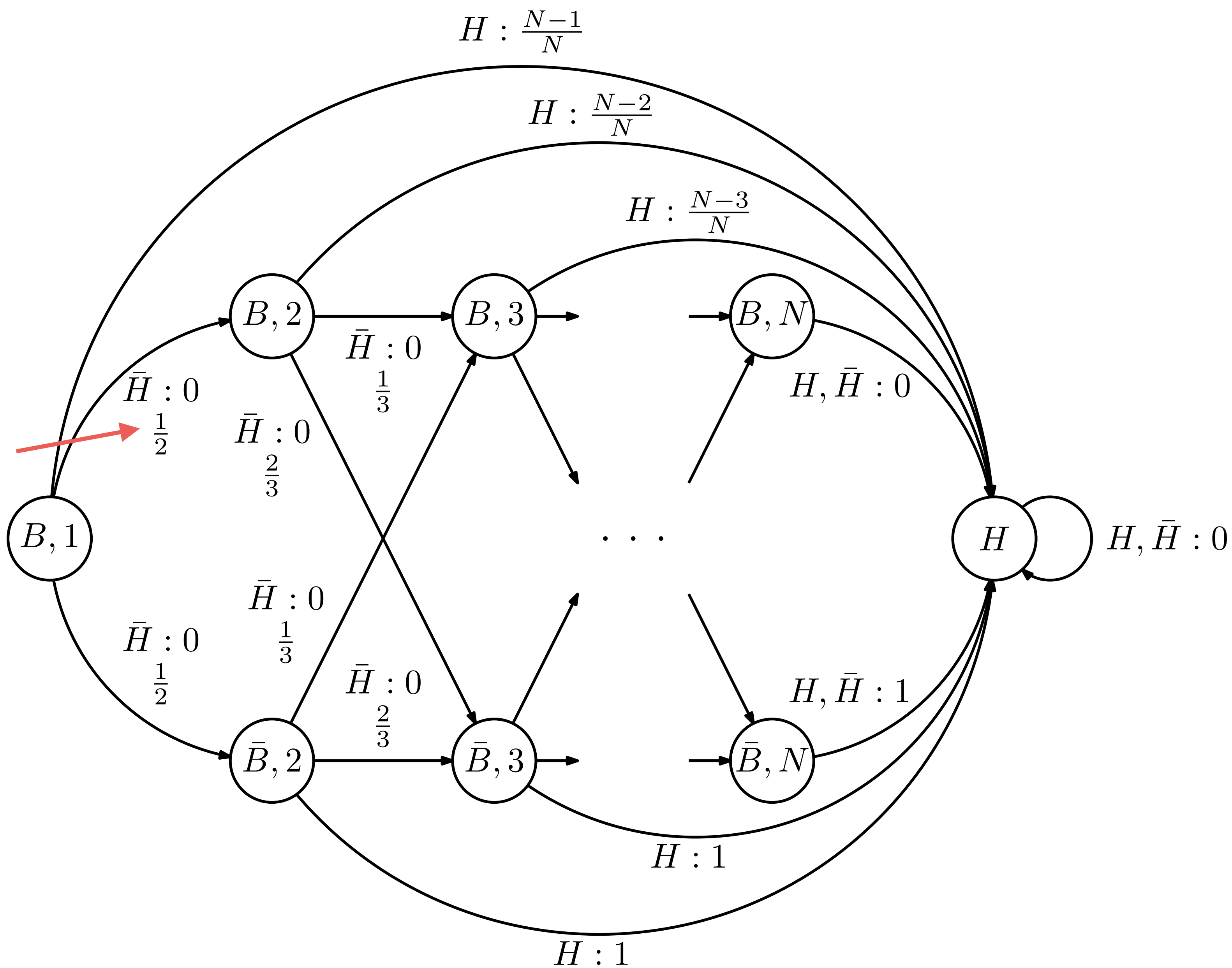
How to model this?

- Cost:
  - … hiring a guy who is not the best so far incurs maximum cost (clearly, that guy is not the best)
  - … what about hiring a guy who is the best so far after n interviews?
    - How likely is it that it is not the best overall?
      - $(N - n) / N$

How to model this?

- Putting everything together:

Probabilities



Decisions with Markov decision processes

Optimality?

- Given a **Markov decision process**, $(\mathscr{X}, \mathscr{A}, \{\mathbf{P}_a\}, c)$…
  - … what do we want to do?

Select the "best" actions

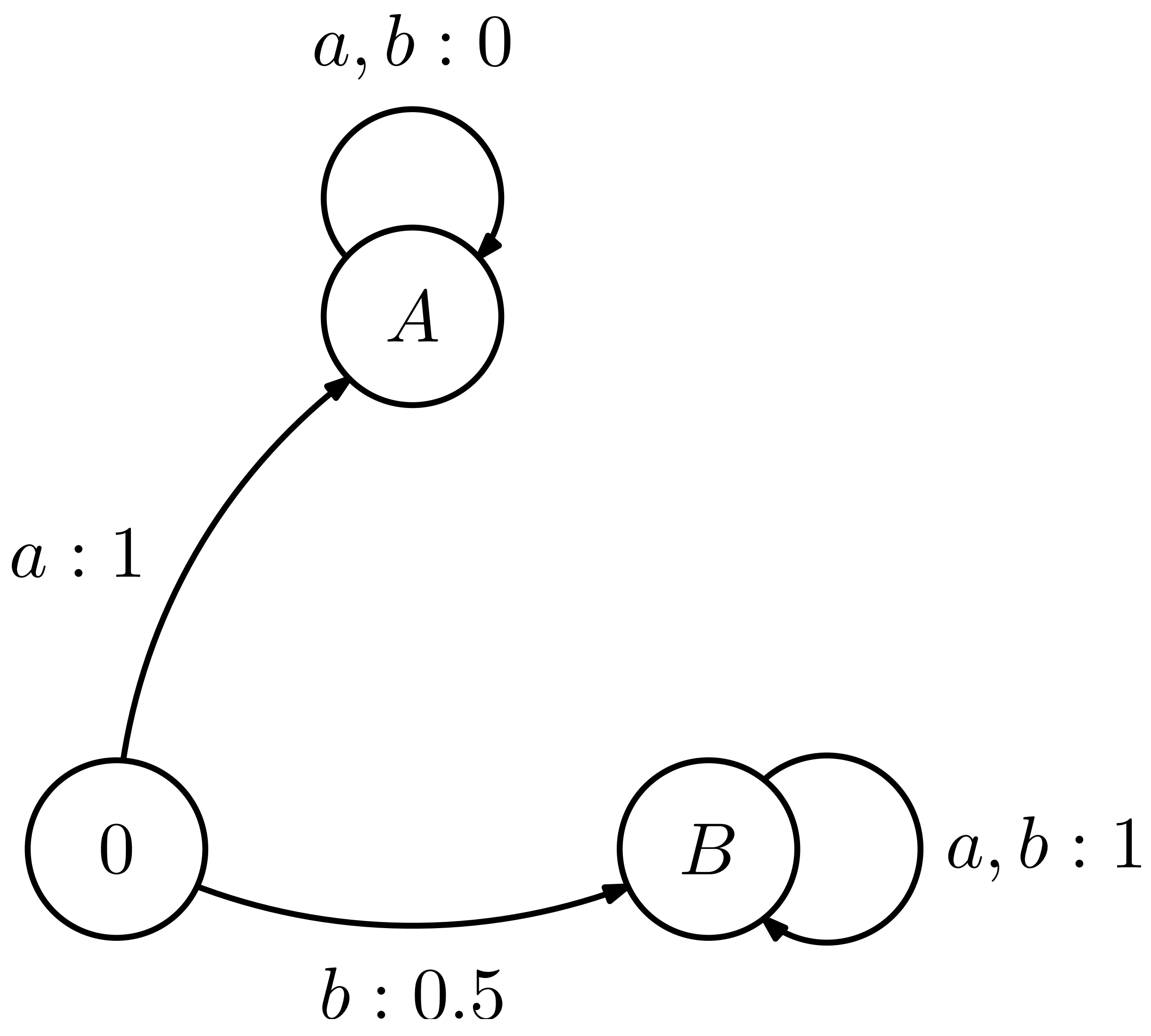
Optimality

- What are the “best” actions?
- We need a criterion to compare different **ways of selecting actions**

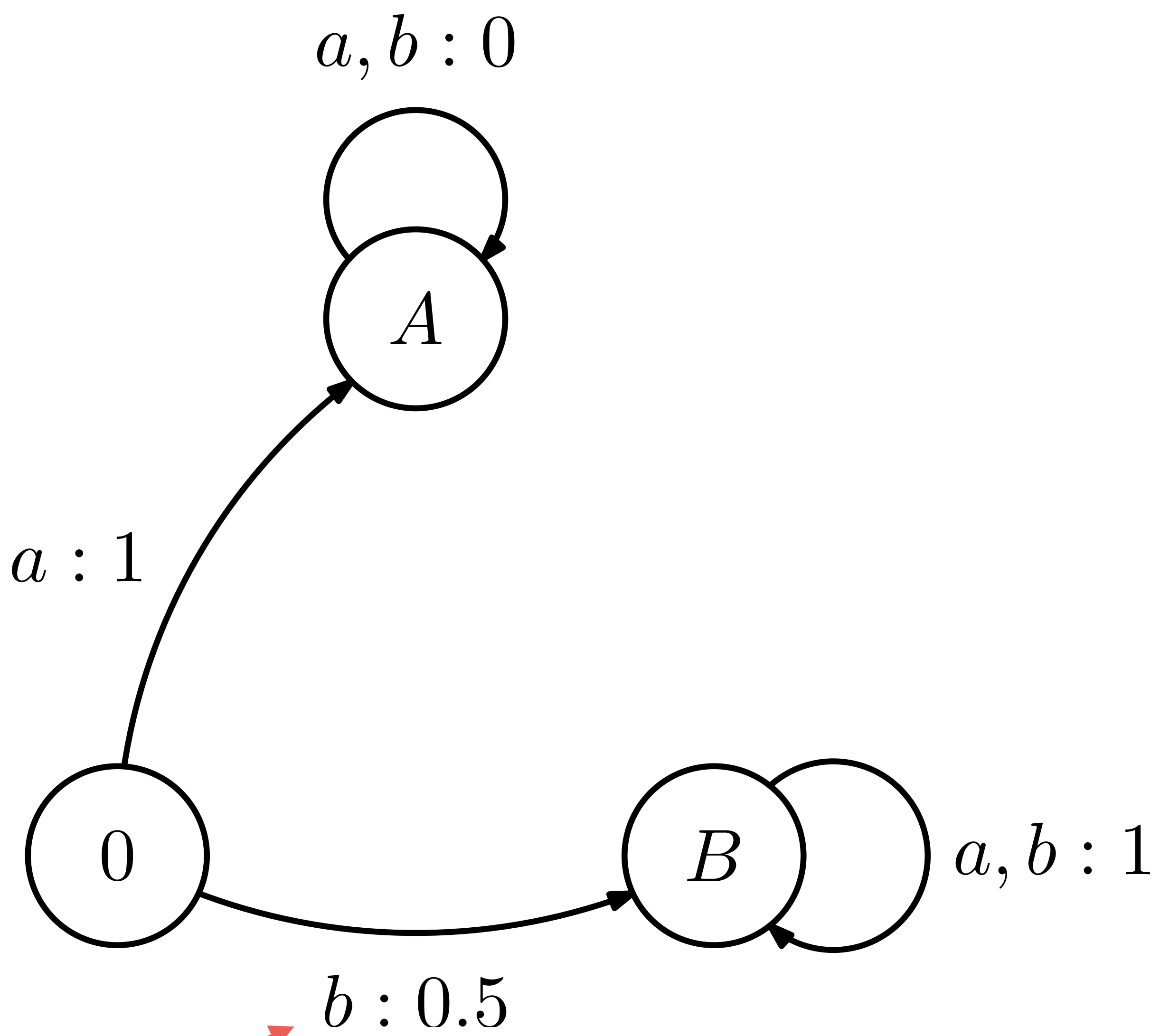
Optimality criterion

Example

What is the best action?



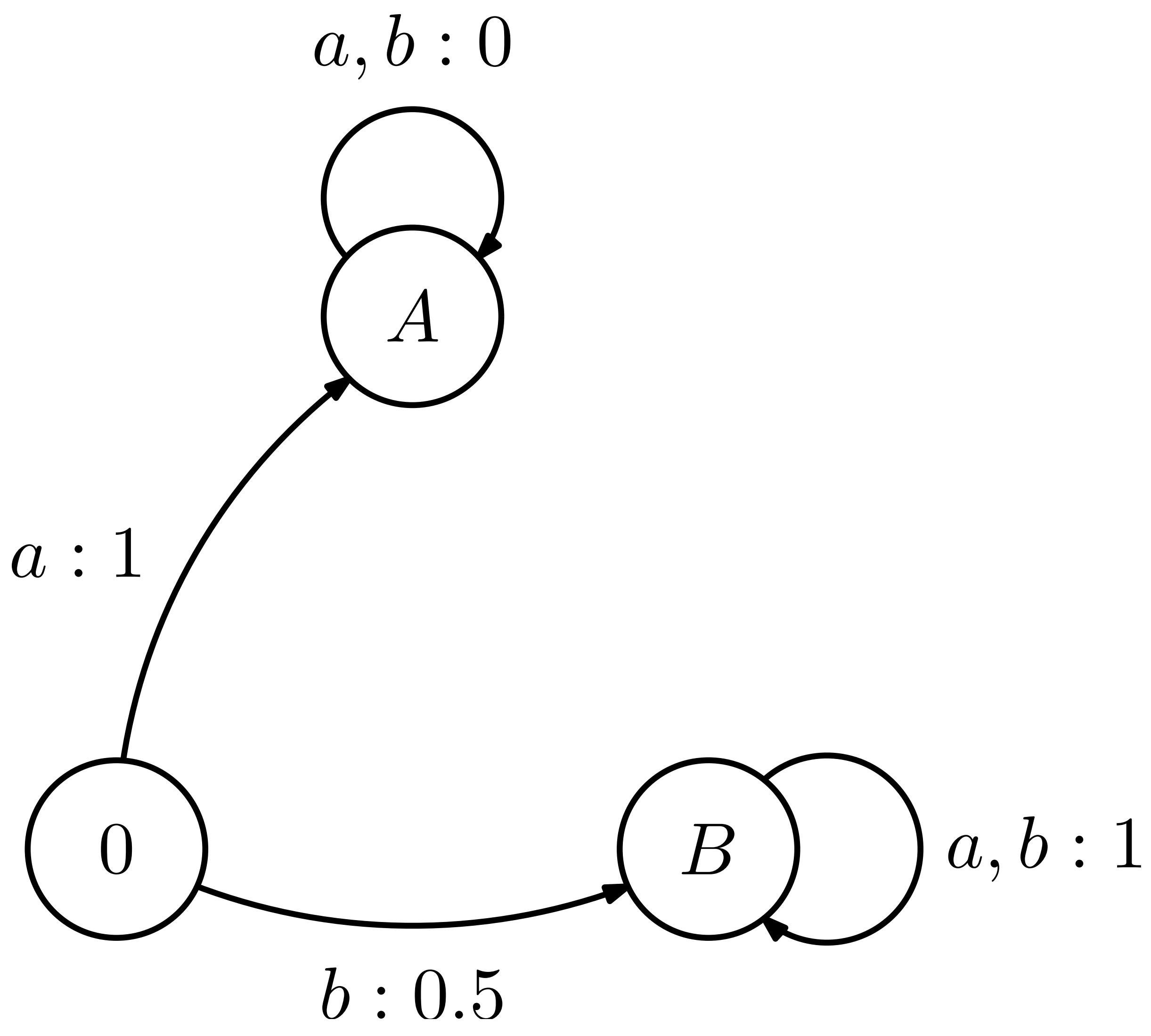
Example



If there is a single decision, *b* is the best!

Example

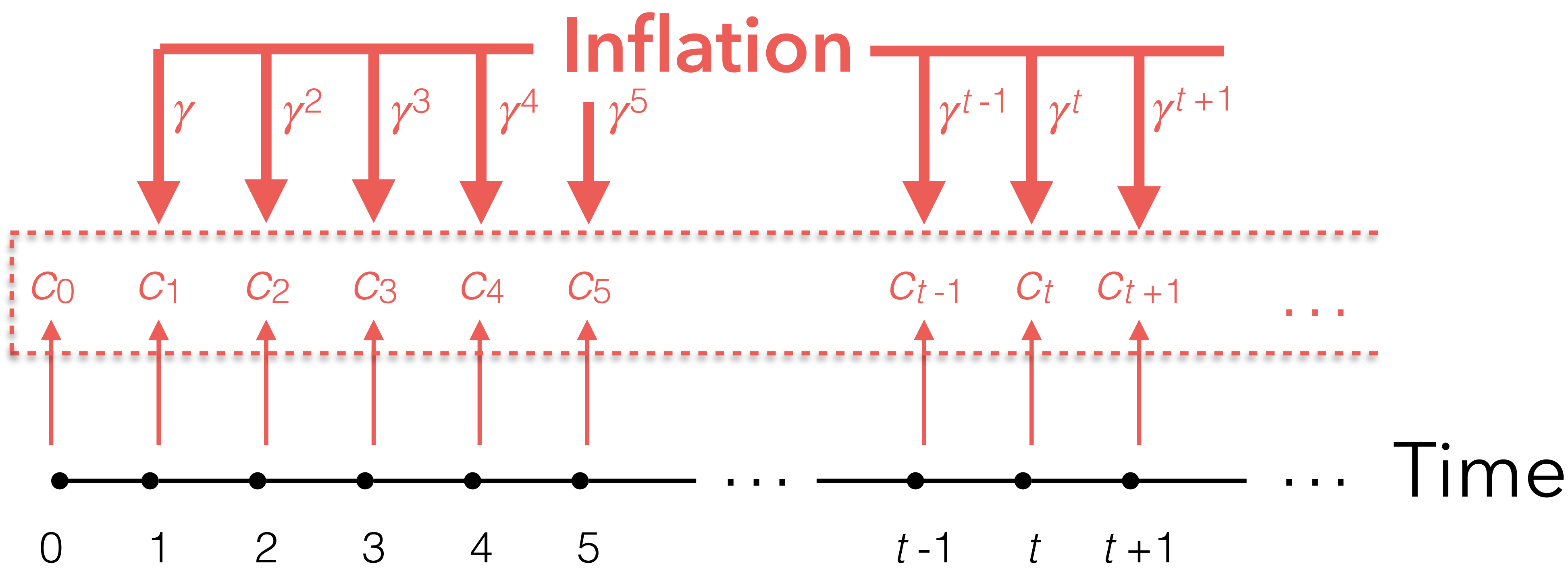
If there is more than one decision, *a* is the best!



Discounted cost-to-go

- Assumptions:
  - The agent lives forever (we don’t know n. of decisions)
  - There is an inflation rate (costs in the future are not as bad as costs now)
  - Agent wants to pay as little as possible

Discounted cost-to-go



Discounted cost-to-go

- Discounted cost-to-go:

Sum of all costs

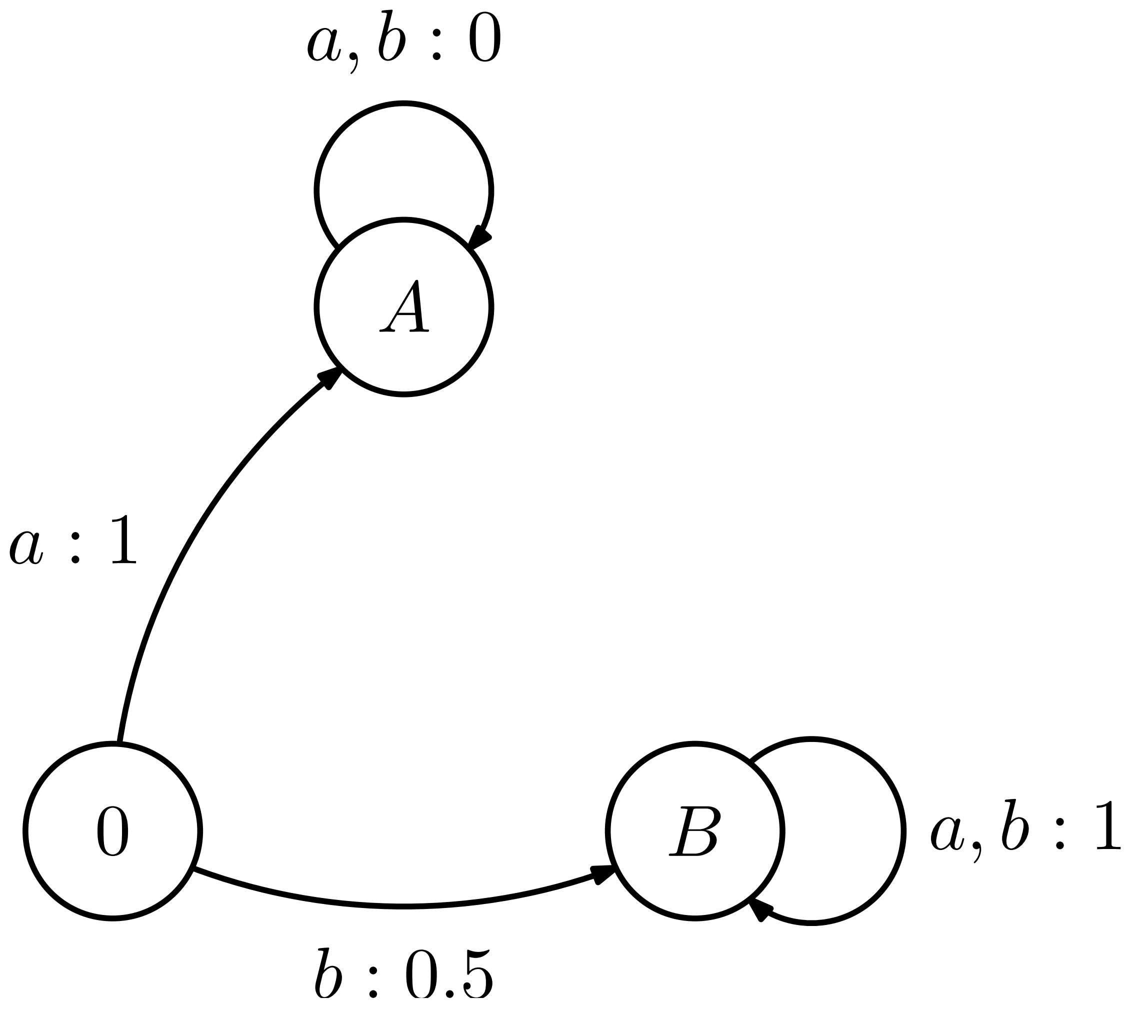
$$DC \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c_t\right]$$

Average

Discounted

Example

- What is the discounted cost-to-go if we always select b?
  - **It depends on where we start!**

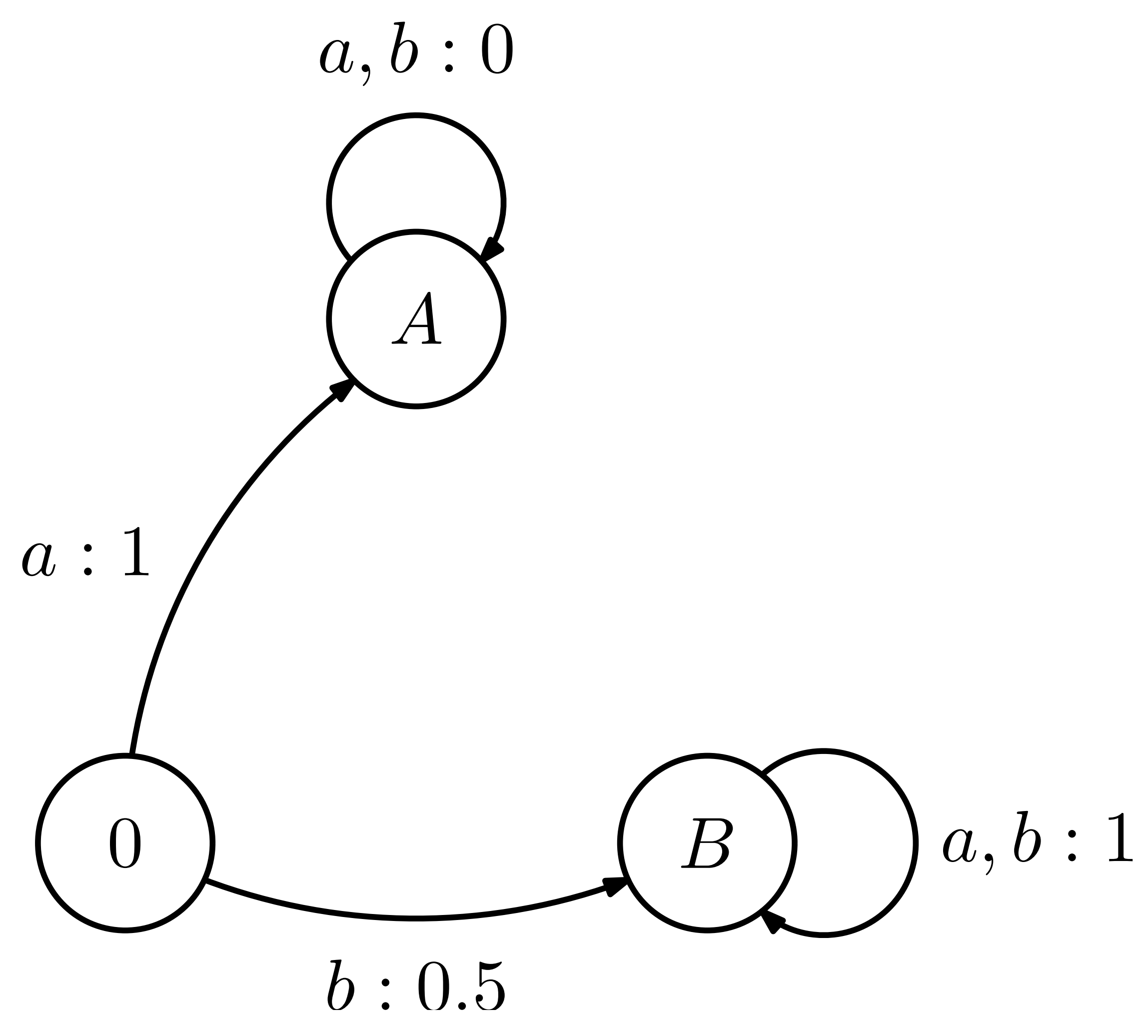


Example

- What if we start in *A*?

$$J(A) = 0 + \gamma 0 + \ldots = 0$$

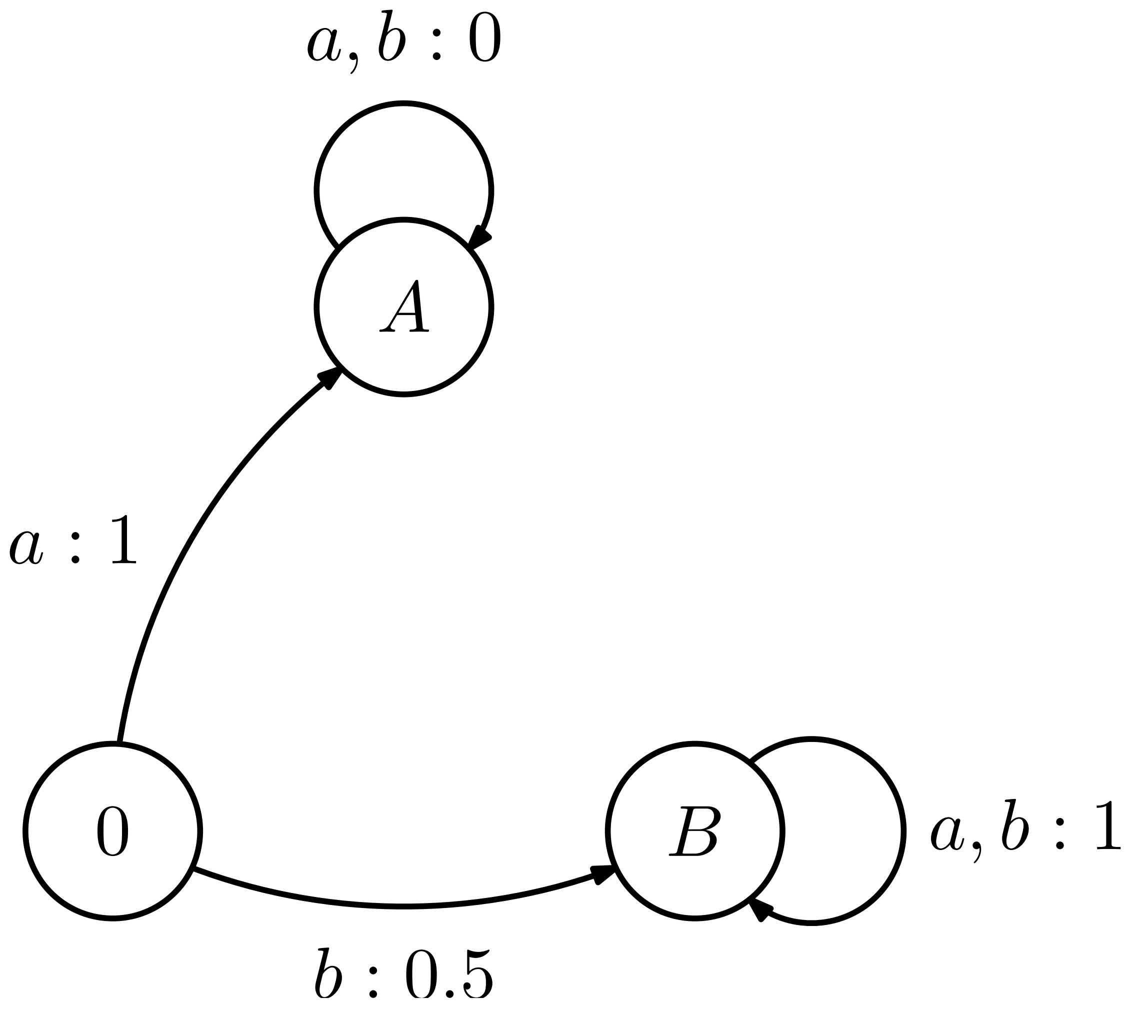
Cost-to-go
if we start
in *A*



Example

- What if we start in *B*?

$$J(B) = 1 + \gamma 1 + \dots$$
$$= \frac{1}{1-\gamma}$$



Example

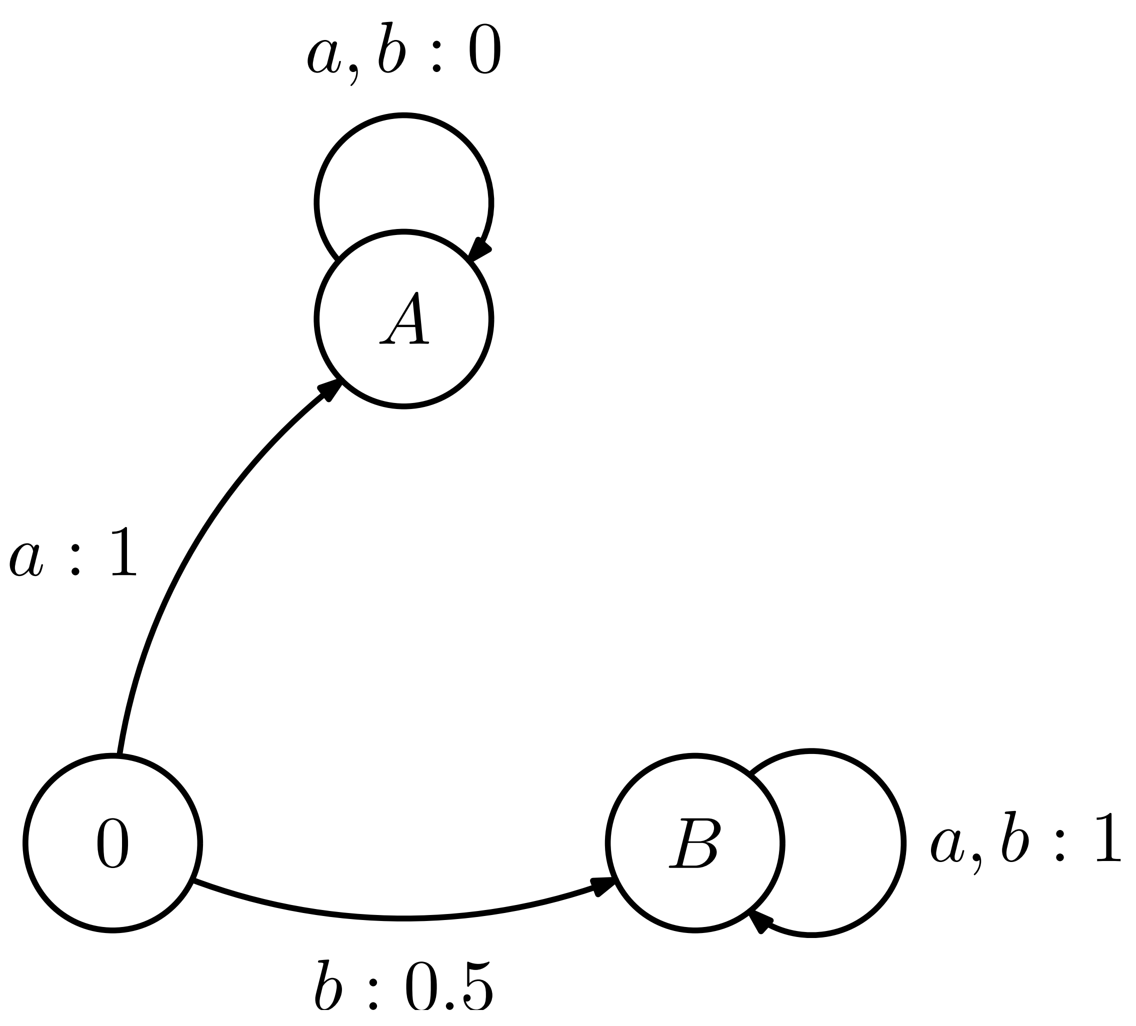
- What if we start in 0?

$$J(0) = 0.5 + \gamma 1 + \gamma^2 1 + \ldots$$

$$= 0.5 + \gamma \underbrace{\boxed{(1 + \gamma 1 + \ldots)}}_{J(B)}$$

$$= 0.5 + \gamma J(B)$$

$$= \frac{1}{2} \cdot \frac{1+\gamma}{1-\gamma}$$



Example

- What is the discounted cost-to-go if we always select *b*?

$$\boldsymbol{J} = \begin{bmatrix} \frac{1}{2} \cdot \frac{1+\gamma}{1-\gamma} \\ 0 \\ \frac{1}{1-\gamma} \end{bmatrix}$$