TÉCNICO LISBOA

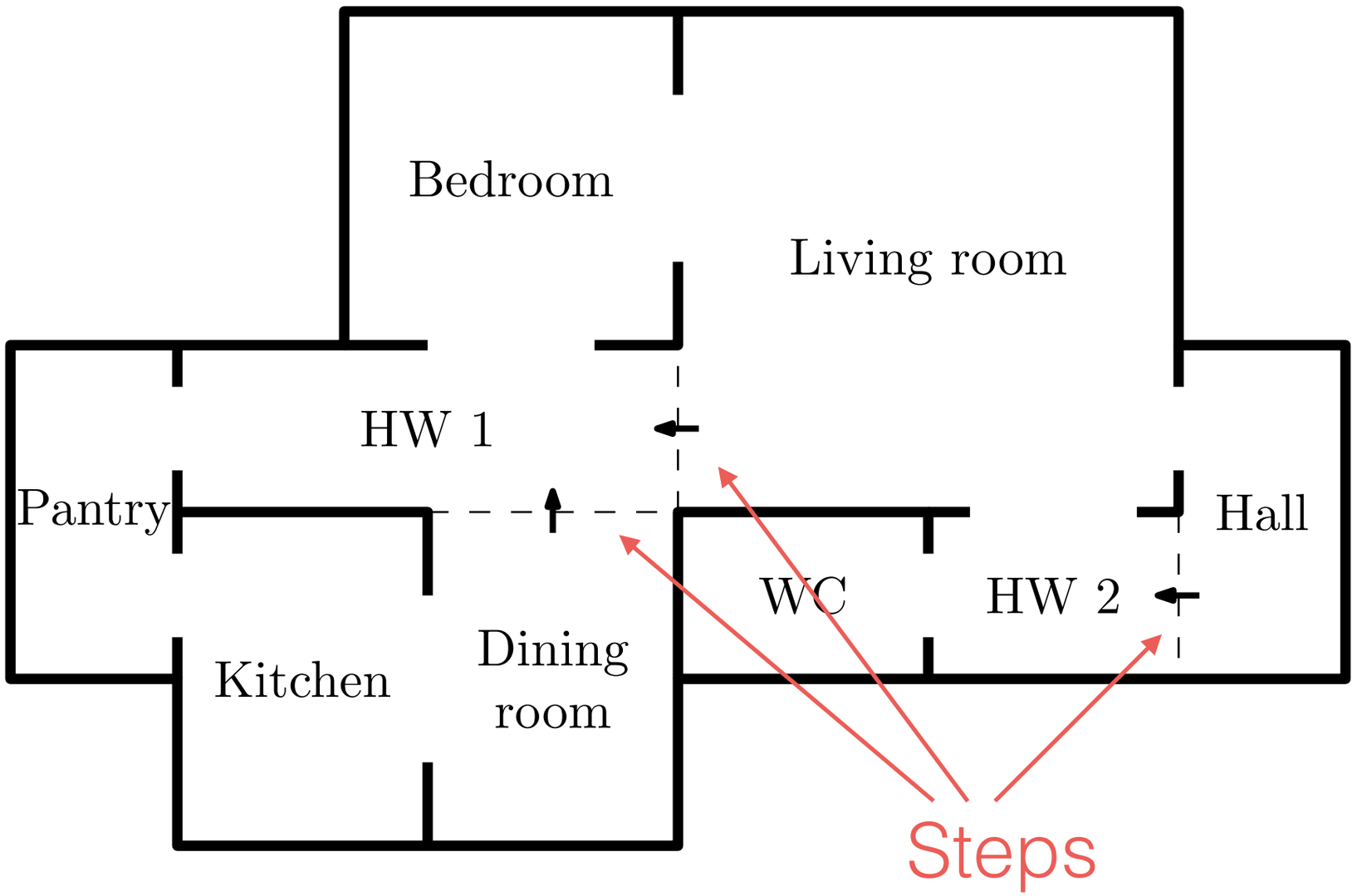
Planning, Learning and Decision Making

Lecture 9. Partially observable Markov decision problems

The household robot... again!

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

Household robot

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

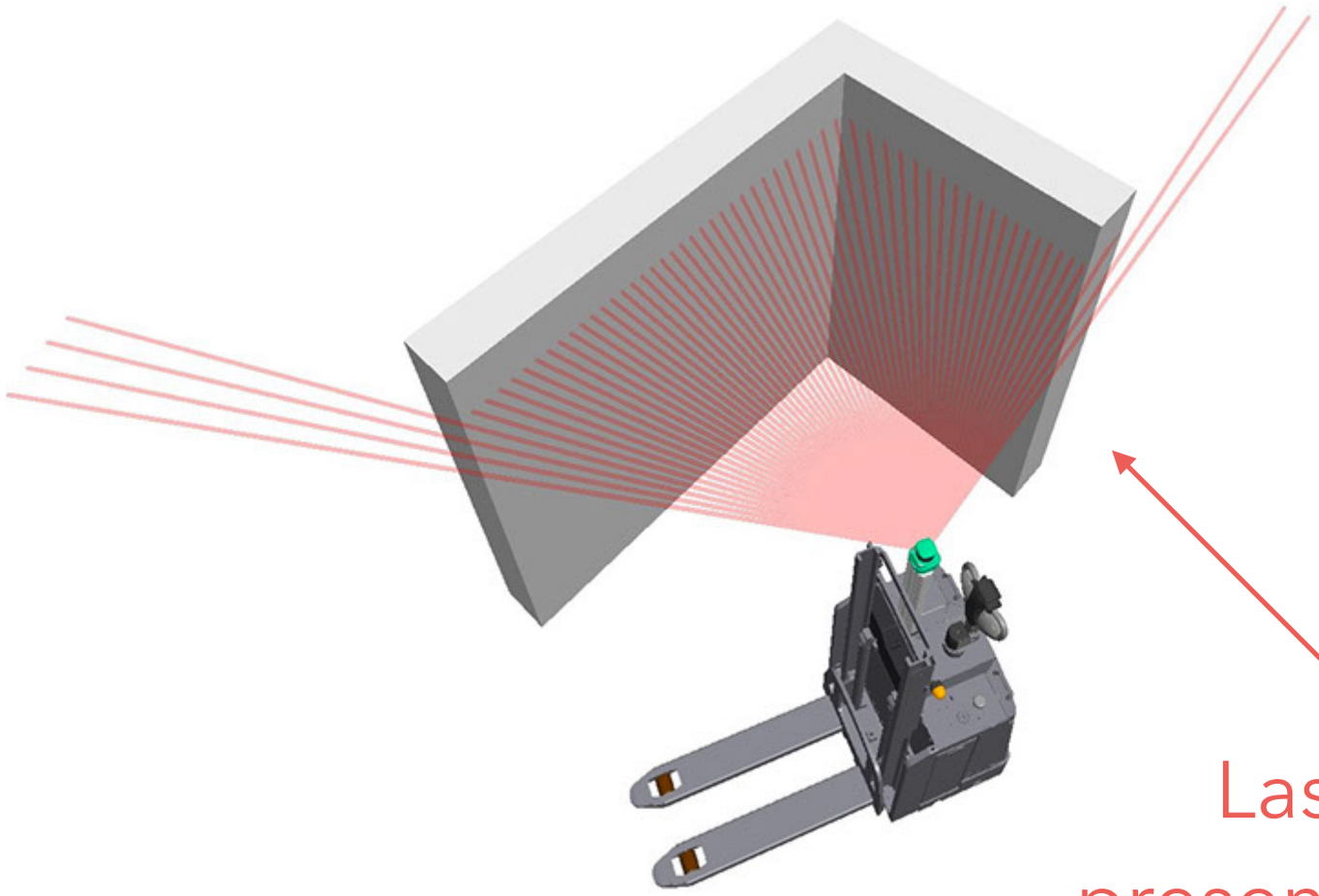
$$\mathscr{A} = \{\text{U(p), D(own), L(eft), R(ight), S(tay)}\}$$

Household robot

- Motions across a step fail with probability 0.4

Household robot

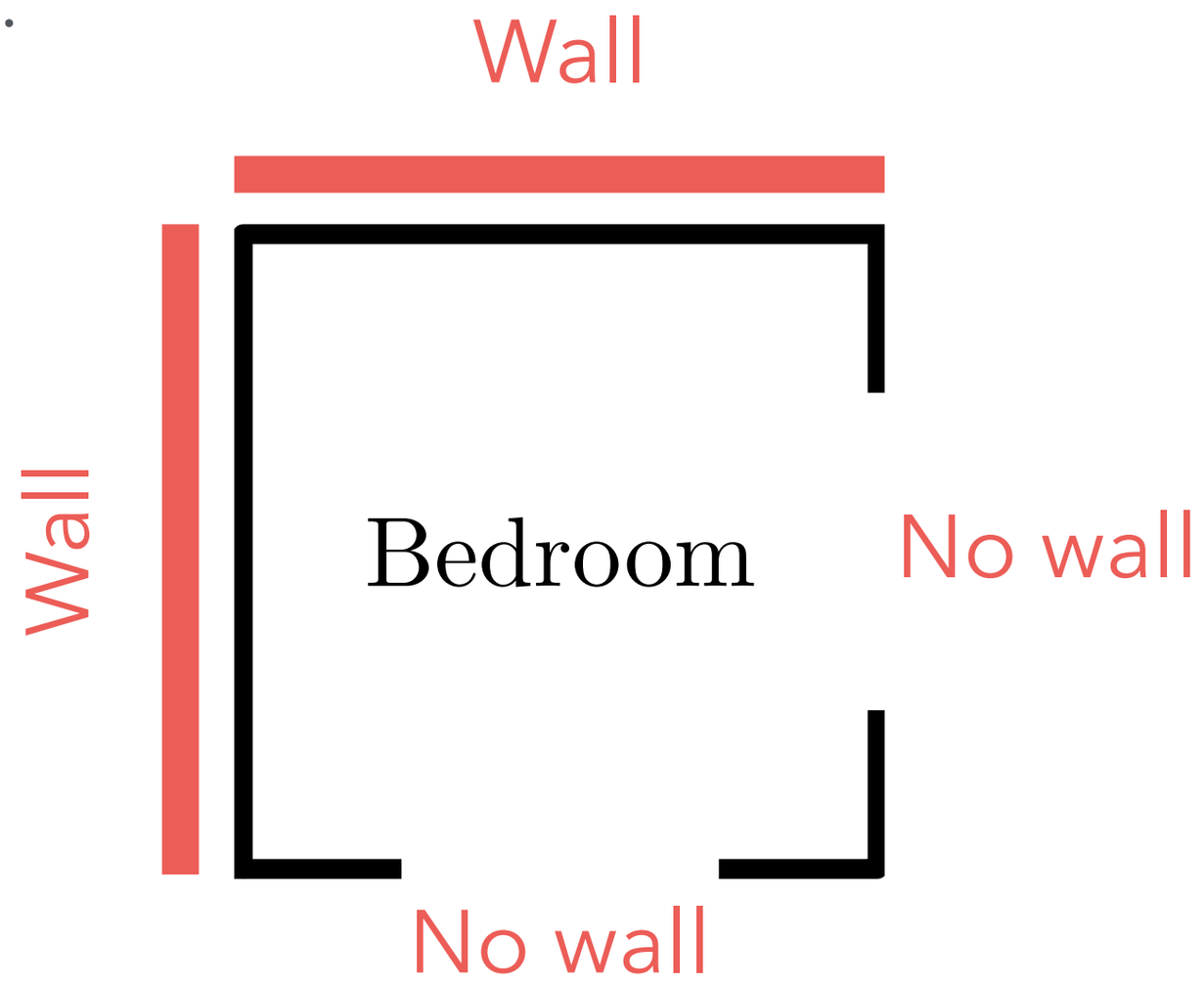
- Robot navigates using a laser



Laser detects presence or absence of walls

Household robot

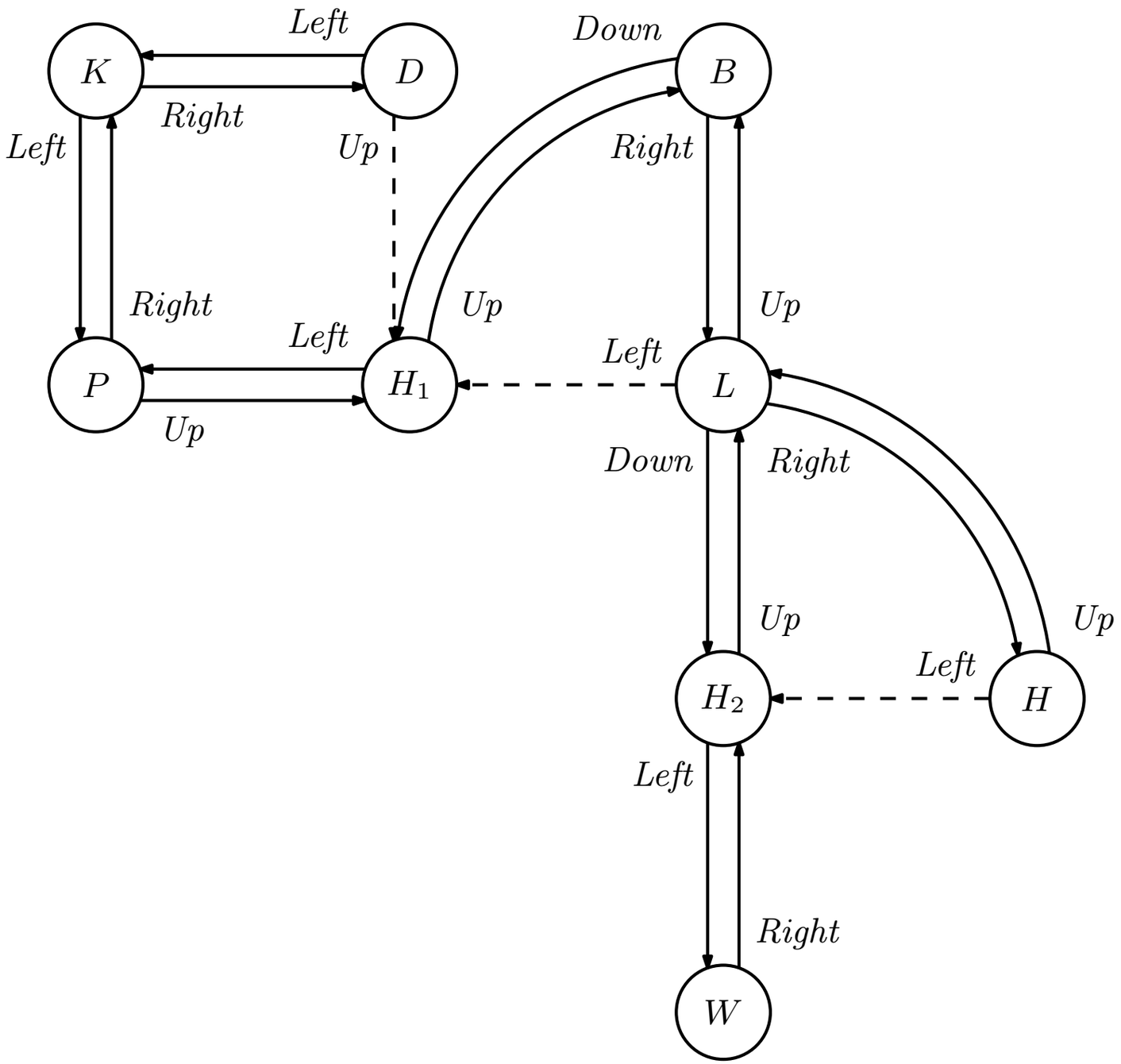
- For example:



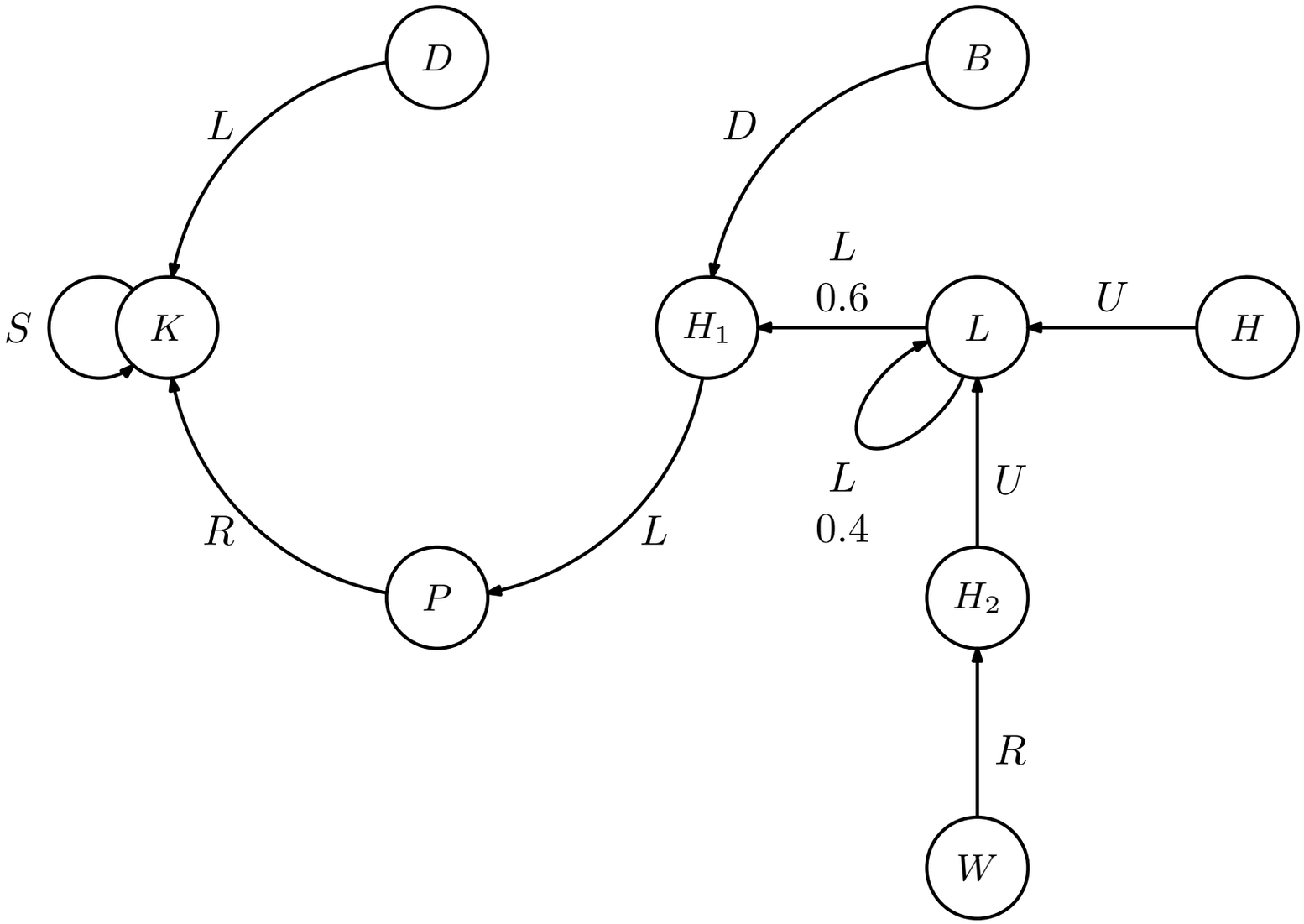
Household robot

- However, laser is not perfect
  - It fails to detect existing walls with 5% probability
  - It detects non-existing walls with 10% probability (in some situations with 20% probability)
  - Detection of a wall independent of adjacent walls

Movement of the robot



MDP solution



Unfortunately…

- At each step, what does the decision of the robot depend on?
  - Position of the robot

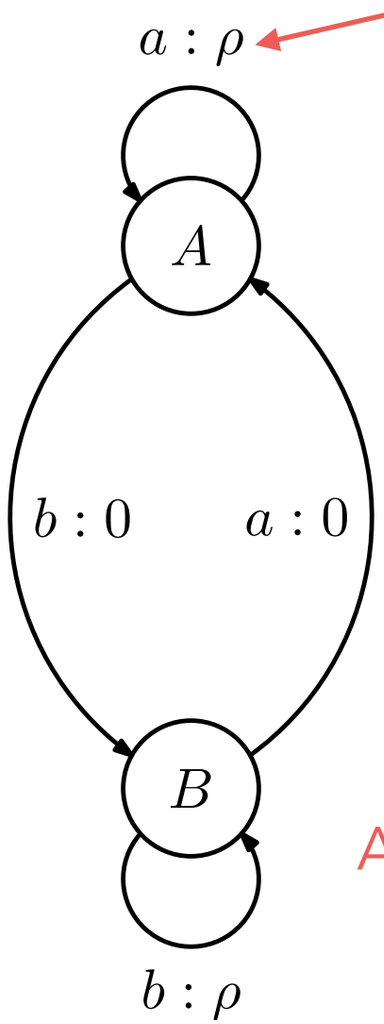
Position is not directly observable!

Decision must be based on observations!

The two-state nightmare

2-state problem

- Consider the following problem:

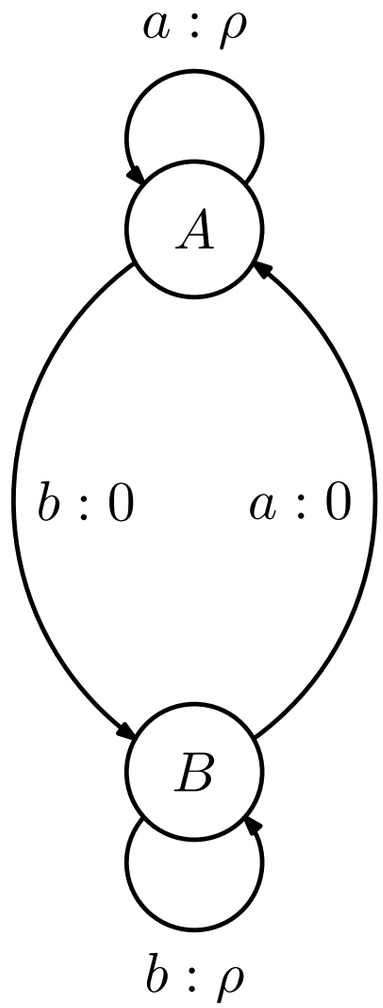


We allow arbitrary cost

All transitions occur with probability 1

2-state problem

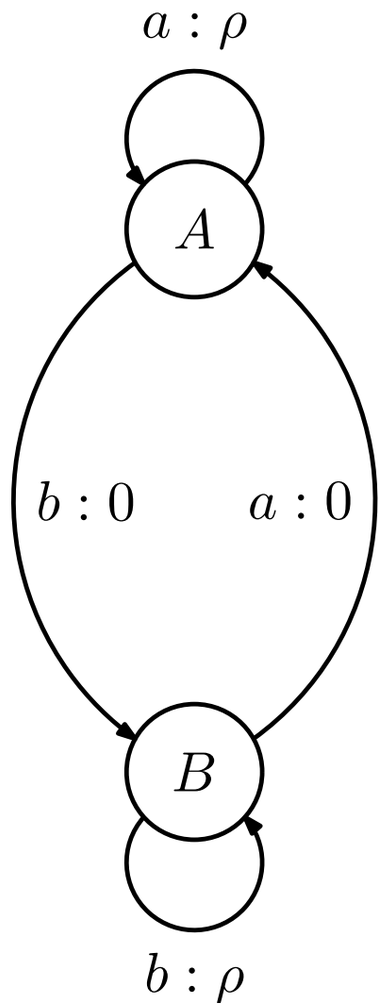
- What is the optimal policy?
  - Select action *b* in state *A*
  - Select action *a* in state *B*



2-state problem

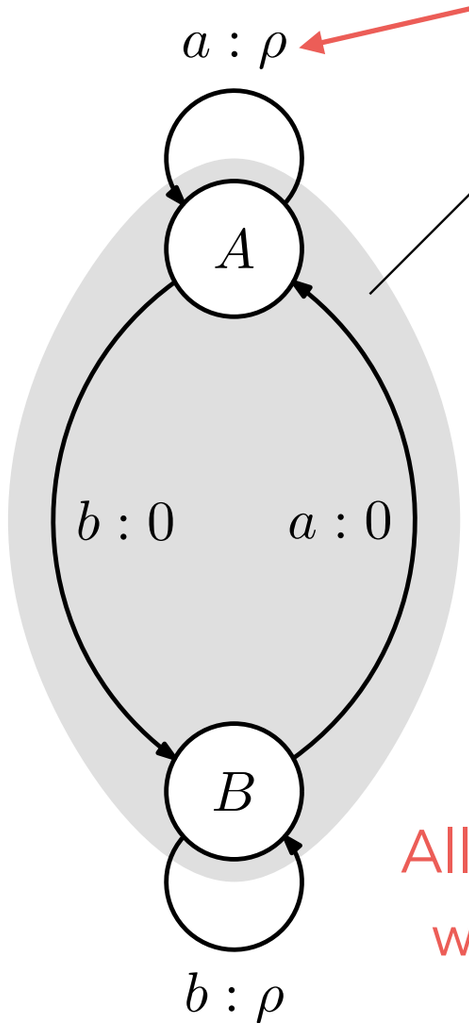
- What is the optimal cost-to-go?
  - Every step a cost of zero, so:

$$J^*(x) = 0$$



2-state problem, v. 2.0

- Consider the following problem:



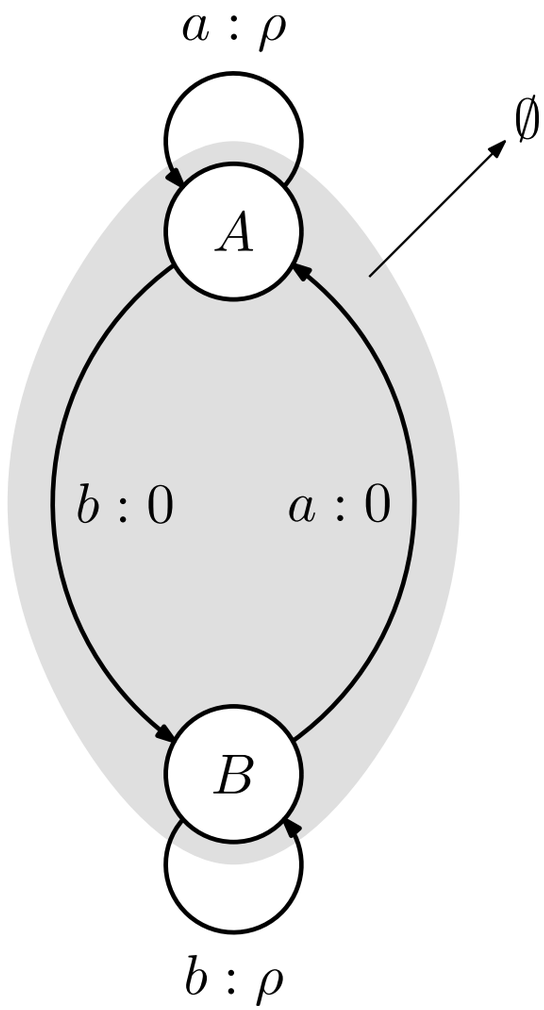
We allow arbitrary cost

No observations

All transitions occur with probability 1

2-state problem

- What is the optimal policy?
  - Not obvious…

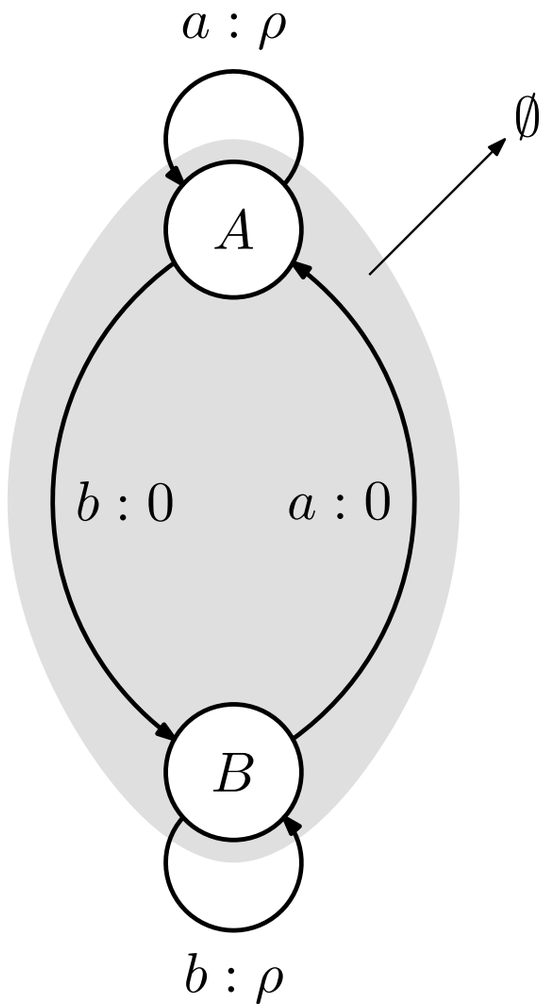


2-state problem

Tentative 1:

- Ignore partial observability
- Select actions deterministically
- "Memoryless policy"

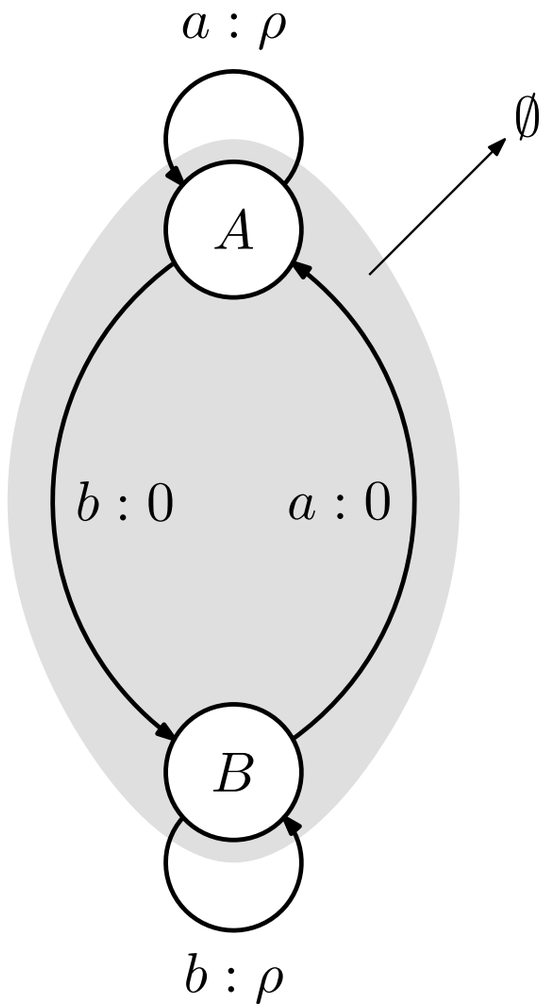
Decision must be constant!



2-state problem

- What is the cost-to-go?
  - Always select *a*
  - Best case: 0 followed by infinite ρs

$$J(x) = 0 + \gamma\rho + \gamma^2\rho + \dots$$
$$= \frac{\gamma\rho}{1-\gamma}$$



2-state problem

- Comparing with the optimal one:

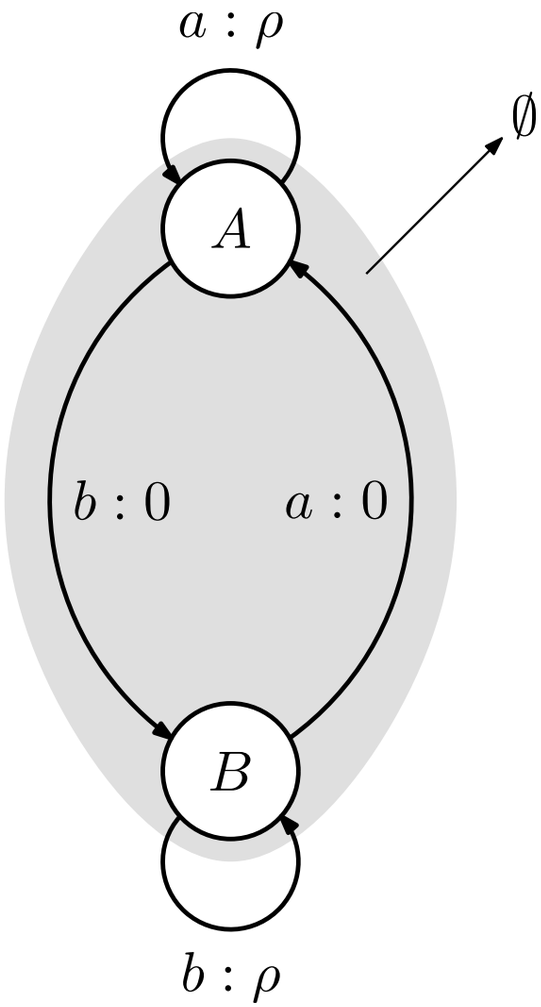
$$\frac{\gamma\rho}{1-\gamma} > 0$$

The best memoryless policy can be
arbitrarily worse than the best MDP policy!

2-state problem

Tentative 2:

- Ignore partial observability
- Select actions stochastically

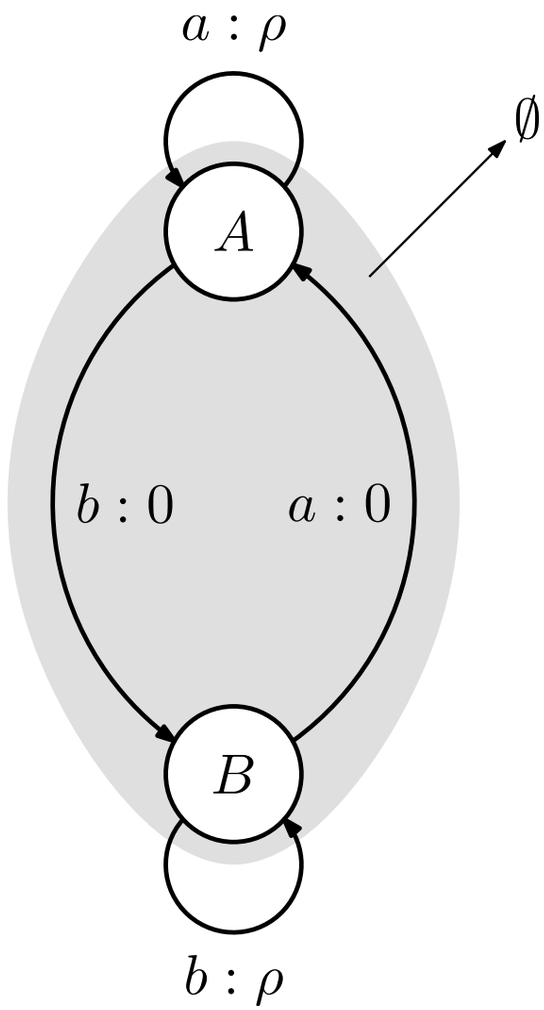


Select each action with probability 0.5

2-state problem

- What is the cost-to-go?
  - In each step incur an average cost of $\rho/2$
  - Therefore:

$$J(x) = \frac{\rho}{2} + \gamma\frac{\rho}{2} + \gamma^2\frac{\rho}{2} + \dots$$
$$= \frac{\rho}{2(1-\gamma)}$$



2-state problem

- Comparing with the deterministic one:

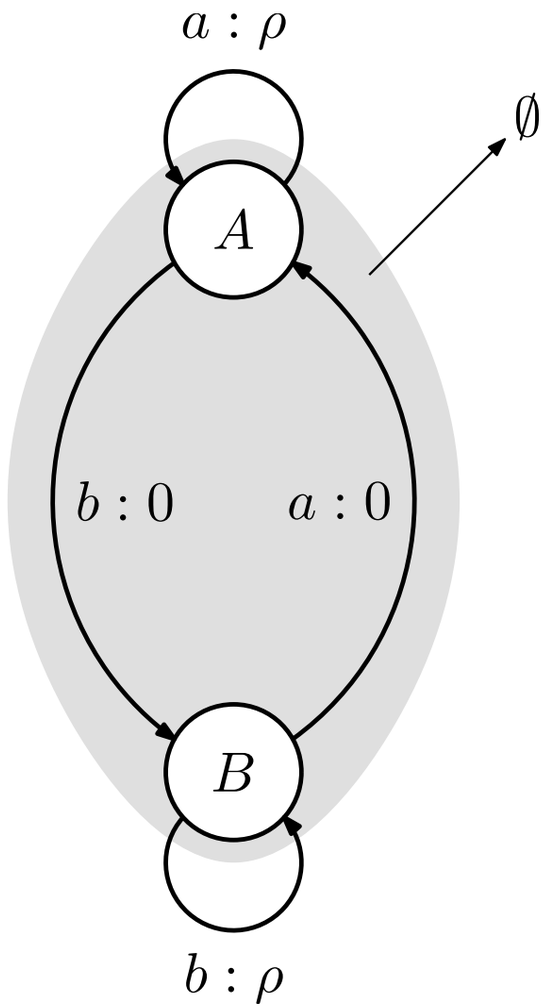
$$\frac{\rho}{2(1-\gamma)} < \frac{\gamma\rho}{1-\gamma} \qquad \text{if } \gamma > 0.5$$

The best deterministic policy can be
arbitrarily worse than the best stochastic policy!

2-state problem

Tentative 3:

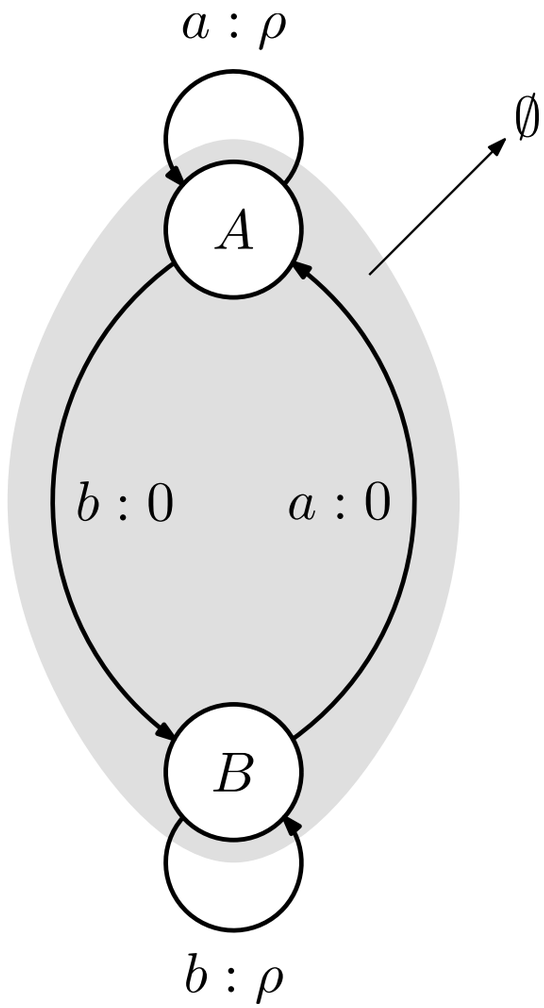
- Non-stationary policy:
    - Alternate action selection



2-state problem

- What is the cost-to-go?
  - Worst case: ρ followed by infinite 0s
  - Therefore:

$$J(x) = \rho + \gamma 0 + \gamma^2 0 + \dots$$
$$= \rho$$



2-state problem

- Comparing with the deterministic one:

$$\rho < \frac{\gamma\rho}{1-\gamma} \qquad \text{if } \gamma > 0.5$$

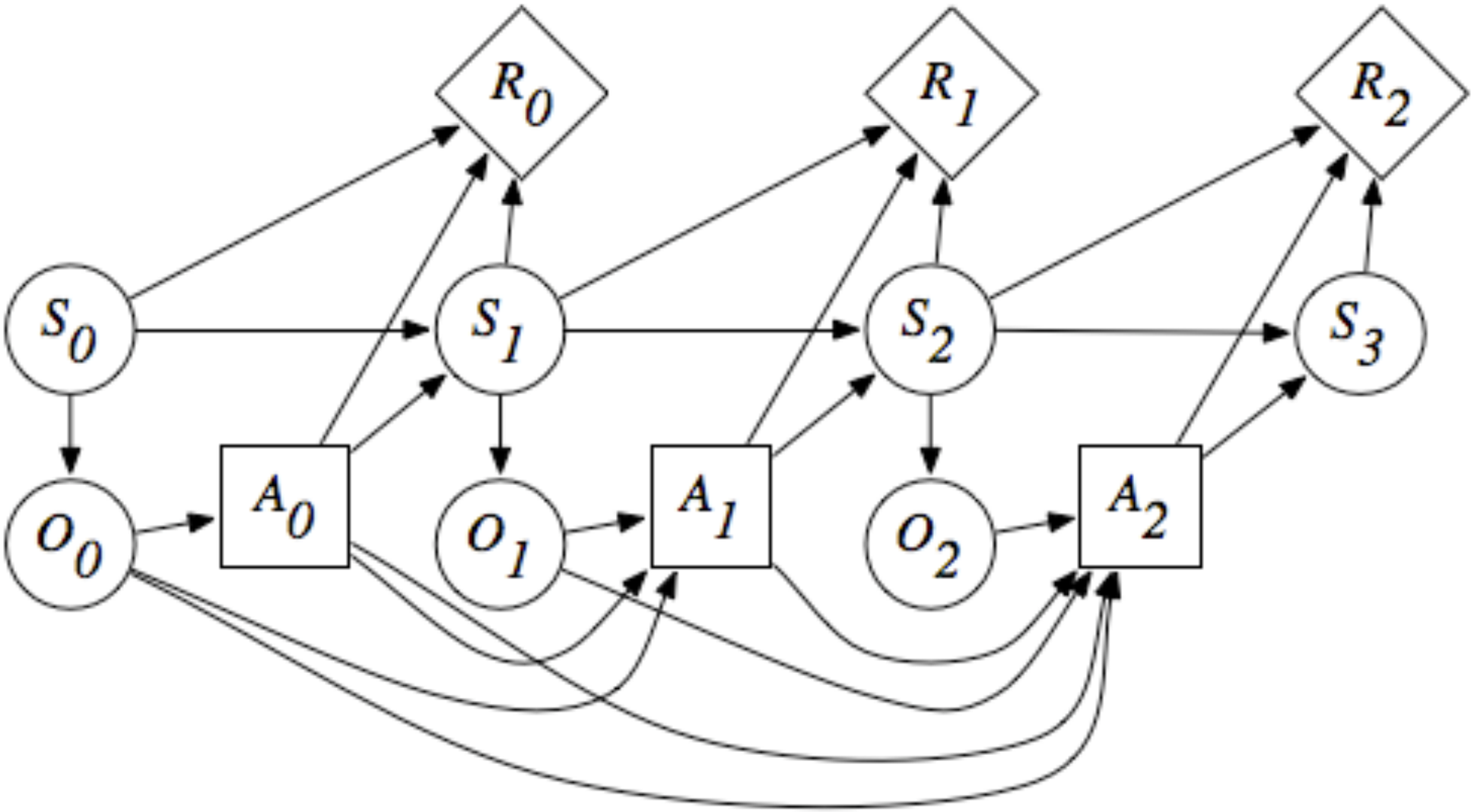
The best deterministic policy can be arbitrarily worse than the best non-stationary policy!

2-state problem

- Comparing with the stochastic one:

$$\rho < \frac{\rho}{2(1-\gamma)} \qquad \text{if } \gamma > 0.5$$

The best stochastic policy can be arbitrarily worse than the best non-stationary policy!



Partially observable MDPs

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**)

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

- 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.,

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

Additional assumptions:

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

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

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

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

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 $c(x, a) \in [0, 1]$

So far, everything looks like an MDP...

Observations

- Information that the agent actually *sees*
- We represent the observation at time t as z_t
- Set of possible observations is $\mathcal{Z}$ (finite)
- Observations depend on current state **and previous action**

Perception

- Describe how the observations depend on the state and the agent's actions
- We assume that observations depend only on the state and (previous) action

State-dependent observations

State-dependent observations

The state at instant t and action at instant t - 1 are enough to predict the observation at instant t:

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

Depends only on x_t and a_{t-1}

Additional assumptions:

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

Probability of observing *z* in *x* given *a*

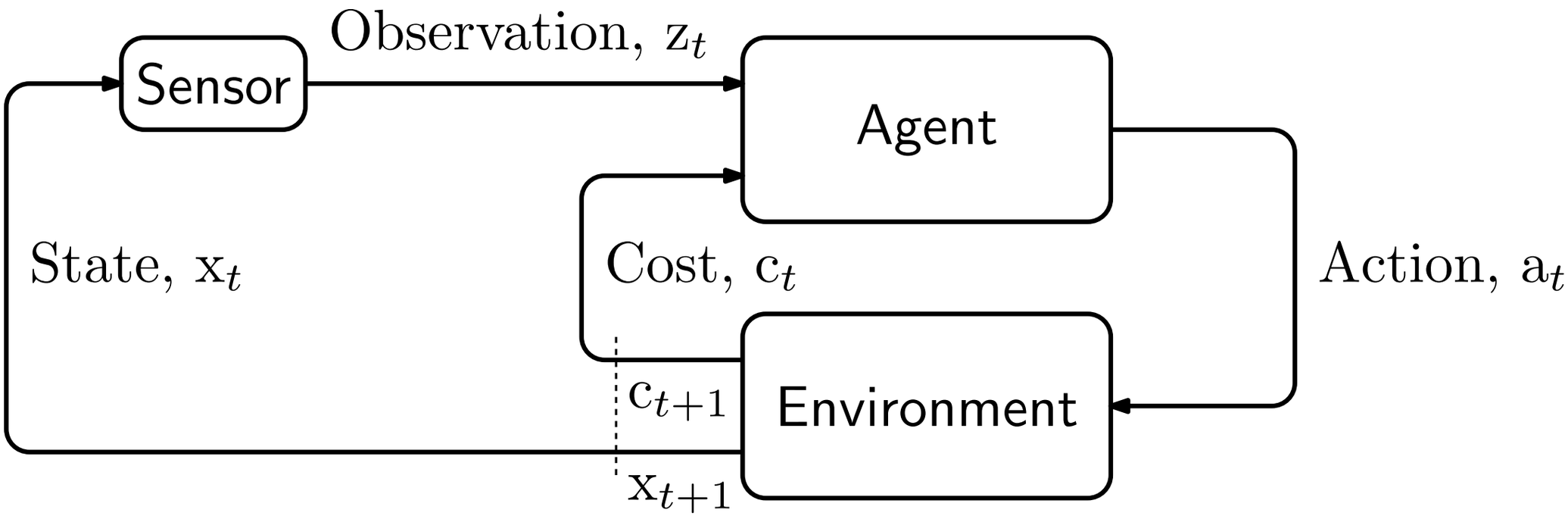
- For each action $a \in \mathscr{A}$, we store the observation probabilities in a **matrix $\mathbf{O}_a$**

$$[\mathbf{O}_a]_{xz} = \mathbb{P}\left[\mathrm{z}_t = z \mid \mathrm{x}_t = x, \mathrm{a}_{t-1} = a\right]$$

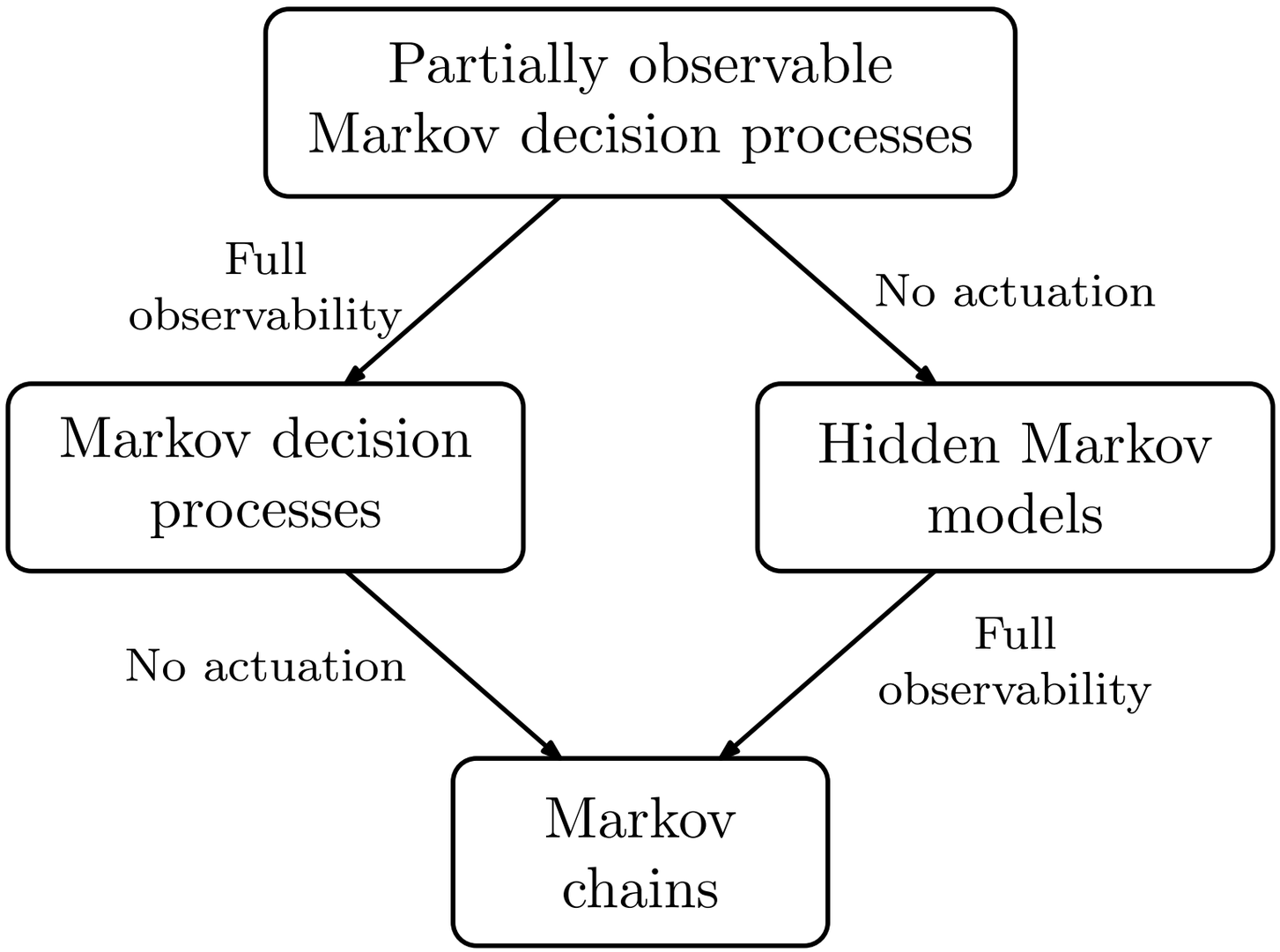
Partially observable MDP

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

Partially observable MDP



An overview


Partially observable
Markov decision processes
Full
observability
No actuation
Markov decision
processes
Hidden Markov
models
No actuation
Full
observability
Markov
chains

Decisions with POMDPs

History

- The **history** at time step *t*…
  - … is a random variable, h_t
  - … contains all that the agent **saw** up to time step *t*:

$$h_t = \{z_0, a_0, z_1, a_1, \ldots, z_{t-1}, a_{t-1}, z_t\}$$

- Set of *t*-length histories (histories up to time *t*) is denoted as $\mathscr{H}_t$

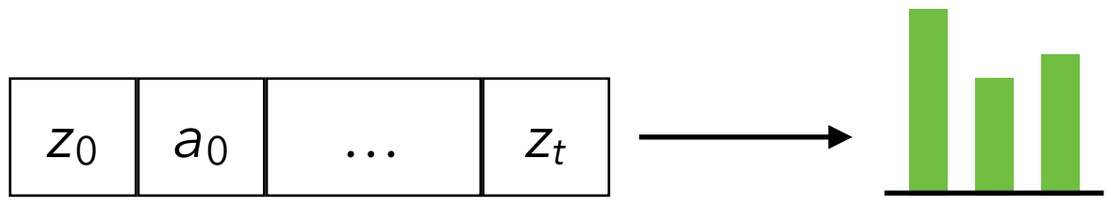
Policies

- A **policy** is a mapping $\pi : \mathscr{H} \to \Delta(\mathscr{A})$

Distributions over actions

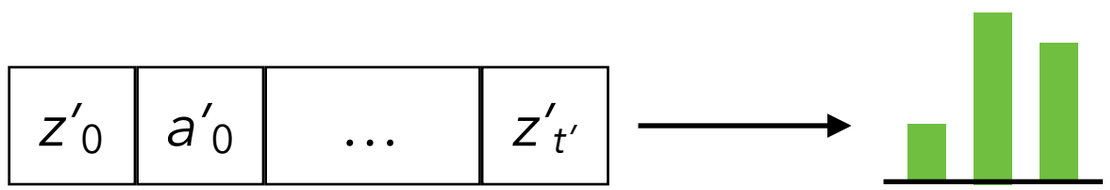
- $\pi(a \mid h)$ is a probability of selecting action a after observing history h

History h



Distribution $\pi(\cdot \mid h)$

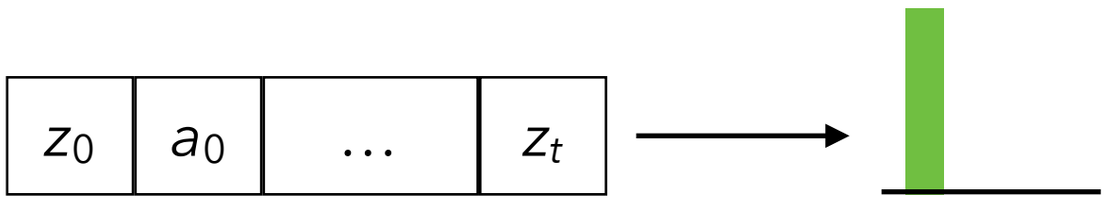
History h'



Distribution $\pi(\cdot \mid h')$

Types of policies

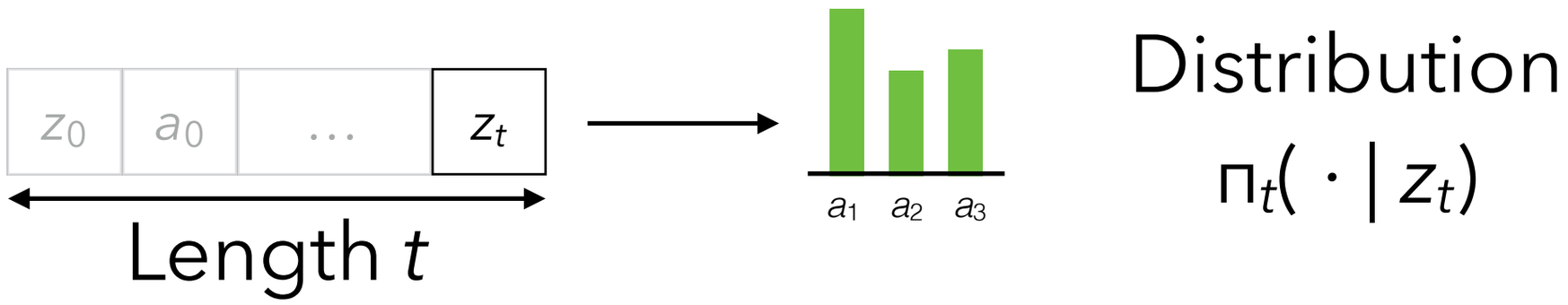
- **Deterministic**
  - … if there is one action that is selected with probability 1
  - We write $\pi(h)$ to denote such action



- **Stochastic**
  - … if it is not deterministic

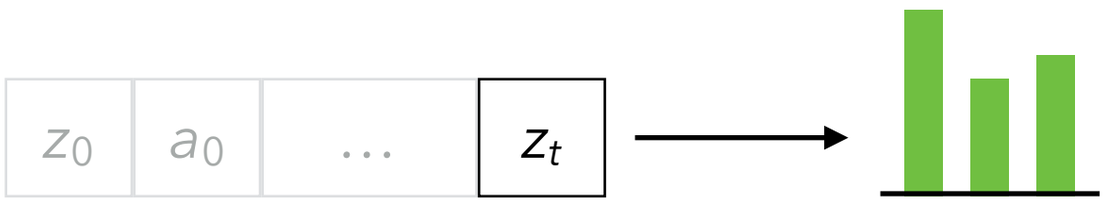
Types of policies

- **~~Markov~~ Memoryless**
  - … if the probability $\pi(a \mid h)$ depends only on the length of h and on its **last observation**
  - If $h = \{z_0, a_0, \ldots, z_t\}$, we write $\pi_t(a \mid z_t)$ to denote the probability $\pi(a \mid h)$



Types of policies

- **(Memoryless) Stationary**
  - … if it depends only on the last **observation** in h
  - If $h = \{z_0, a_0, \ldots, z_t\}$, we write $\pi(a \mid z_t)$ to denote the probability $\pi(a \mid h)$



Distribution
$\pi(\cdot \mid z_t)$

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:

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

Cost-to-go function

- Cost-to-go function:
  - Fix a policy, π
  - Deploy the agent according to some initial distribution $\boldsymbol{b}_0$
  - Let the agent go
  - Keep track of all costs to pay

Cost-to-go function

- How much will the agent pay?
  - Depends on the policy π
  - Depends on the initial distribution $\boldsymbol{b}_0$

$$J^{\pi}(\boldsymbol{b}) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}\mathrm{c}_{t} \;\middle|\; \mathrm{x}_0 \sim \boldsymbol{b}_0\right]$$

Discounted cost-to-go

Policy

Initial distribution

Examples

The tiger problem

The tiger problem

- You are a prisoner, trying to escape a dungeon
- At a point in your escape, you face two doors

The tiger problem

- Behind one of the doors (you don't know which) lies your freedom

The tiger problem

- Behind the other door (you don't know which) lies a fearsome tiger

The tiger problem

- You can try to open one of the doors or listen behind the doors
  - When you listen, you hear the tiger behind the correct door with 85% probability
  - You waste time and may be caught
- When you open a door, you either go free or die

The tiger problem

- You are cursed to keep repeating this forever…

The POMDP model

- Can you model this problem as a POMDP?

States

- What are the states?
  - Tiger on the left
  - Tiger on the right

Actions

- What are the actions?
  - Open left
  - Open right
  - Listen

Observations

- What are the observations?
  - Tiger left
  - Tiger right

Transition probabilities

- What are the transition probabilities?
  - Depend on the action
  - When listening, position of the tiger doesn't change

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

Transition probabilities

- What are the transition probabilities?
  - Depend on the action
  - When opening a door, you either die or go free
  - The world “resets” (it’s the curse!…)

$$\mathbf{P}_{OL} = \mathbf{P}_{OR} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

Observation probabilities

- What are the observation probabilities?
  - Depend on the action
  - When listening, you hear the tiger in the correct position with probability 0.85

$$\mathbf{O}_L = \begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix}$$

Observation probabilities

- What are the observation probabilities?
  - Depend on the action
  - When opening a door, you either die or go free
  - The world “resets”, but you hear nothing (model it as a random observation)

$$\mathbf{O}_{OL} = \mathbf{O}_{OR} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

Cost

- What is the cost function?
  - Maximum cost for dying
  - Minimum cost for going free
- **Depends on the action!**

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0.1 \\ 0 & 1 & 0.1 \end{bmatrix}$$

Small cost for listening