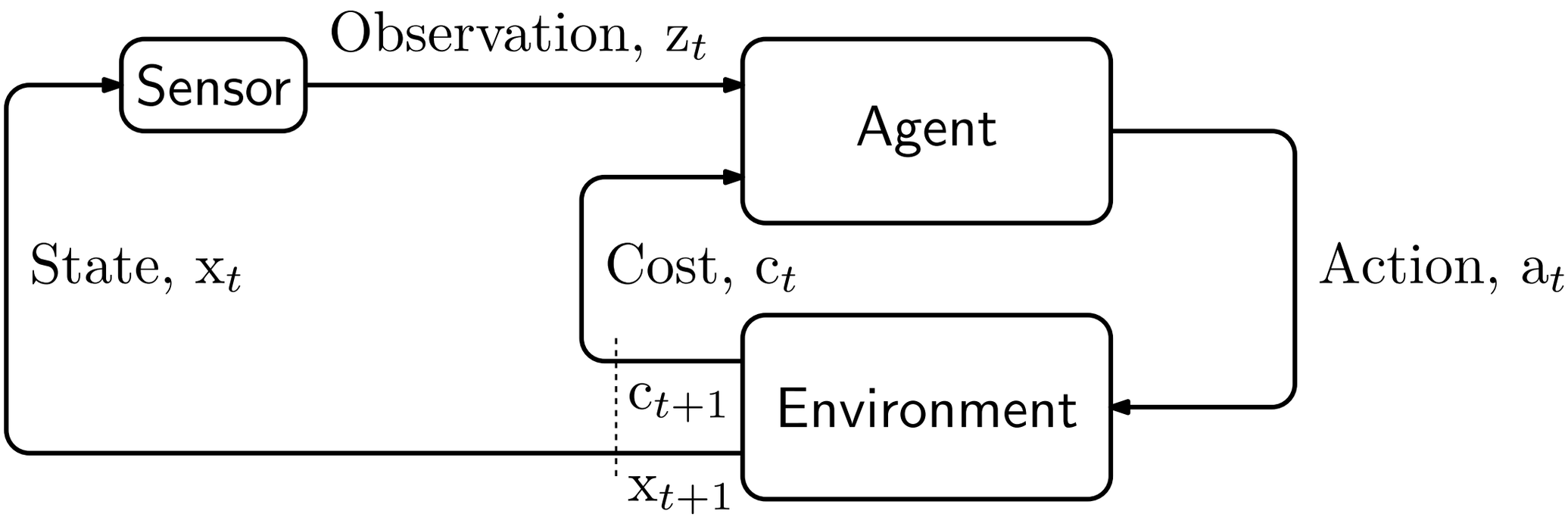
TÉCNICO LISBOA

Planning, Learning and Intelligent Decision Making

Lecture 7

PADInt 2024

Partially observable MDPs



But how do we select actions?

Challenges

- Deterministic memoryless policies are not good enough
- Optimal policy may need to keep track of the history...

An old trick

- At time t = 0, you don't know where the tiger is
- You execute "Listen" 2 times
- You observe "Right", "Right"
- How can you use this information?

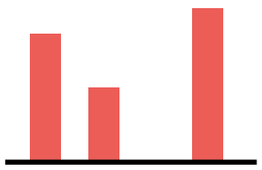
$$\mathbf{P}_{OL} = \mathbf{P}_{OR} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

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

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

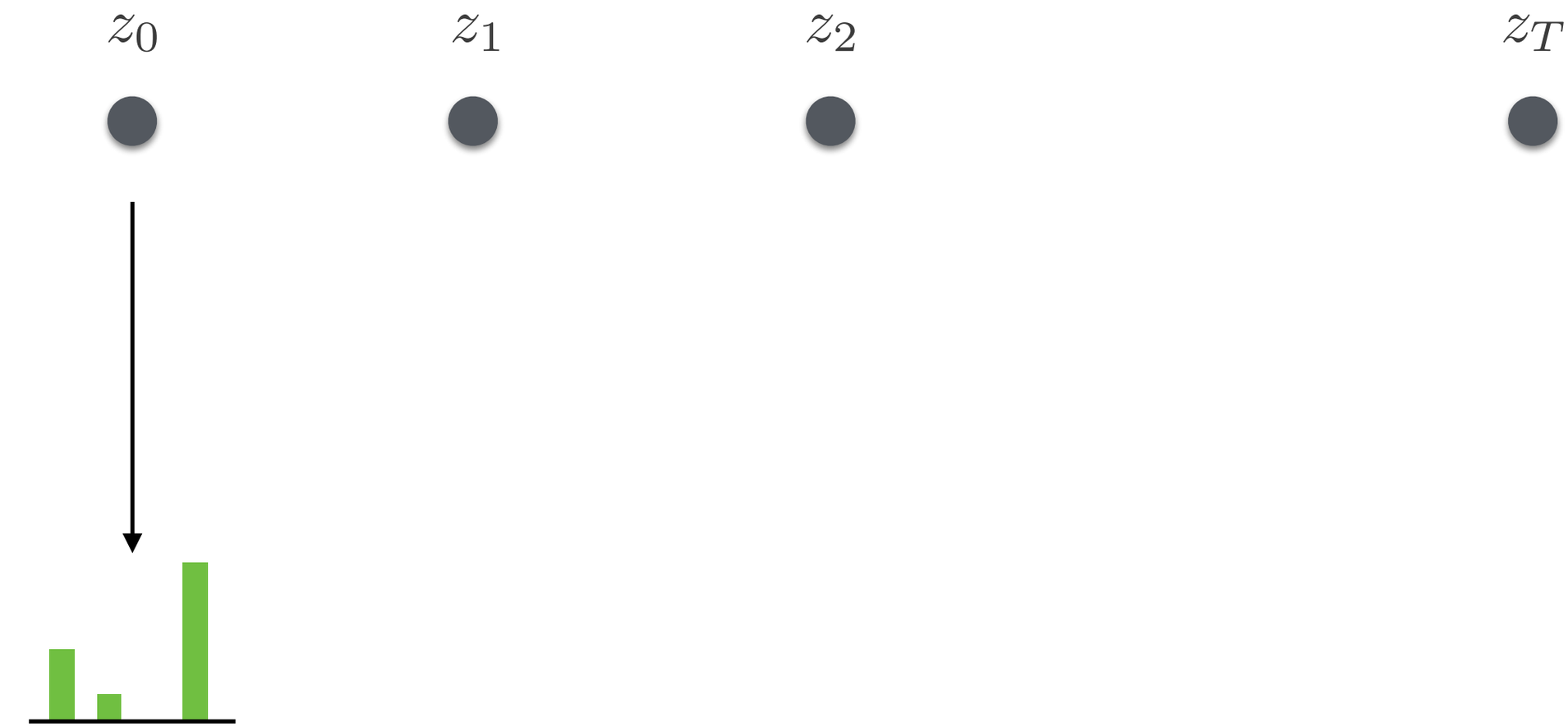
$$\mathbf{O}_{L} = \begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix}$$

An old trick



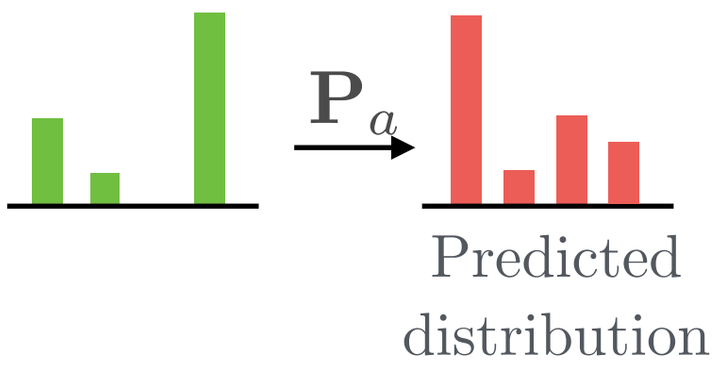
Initial distribution

An old trick

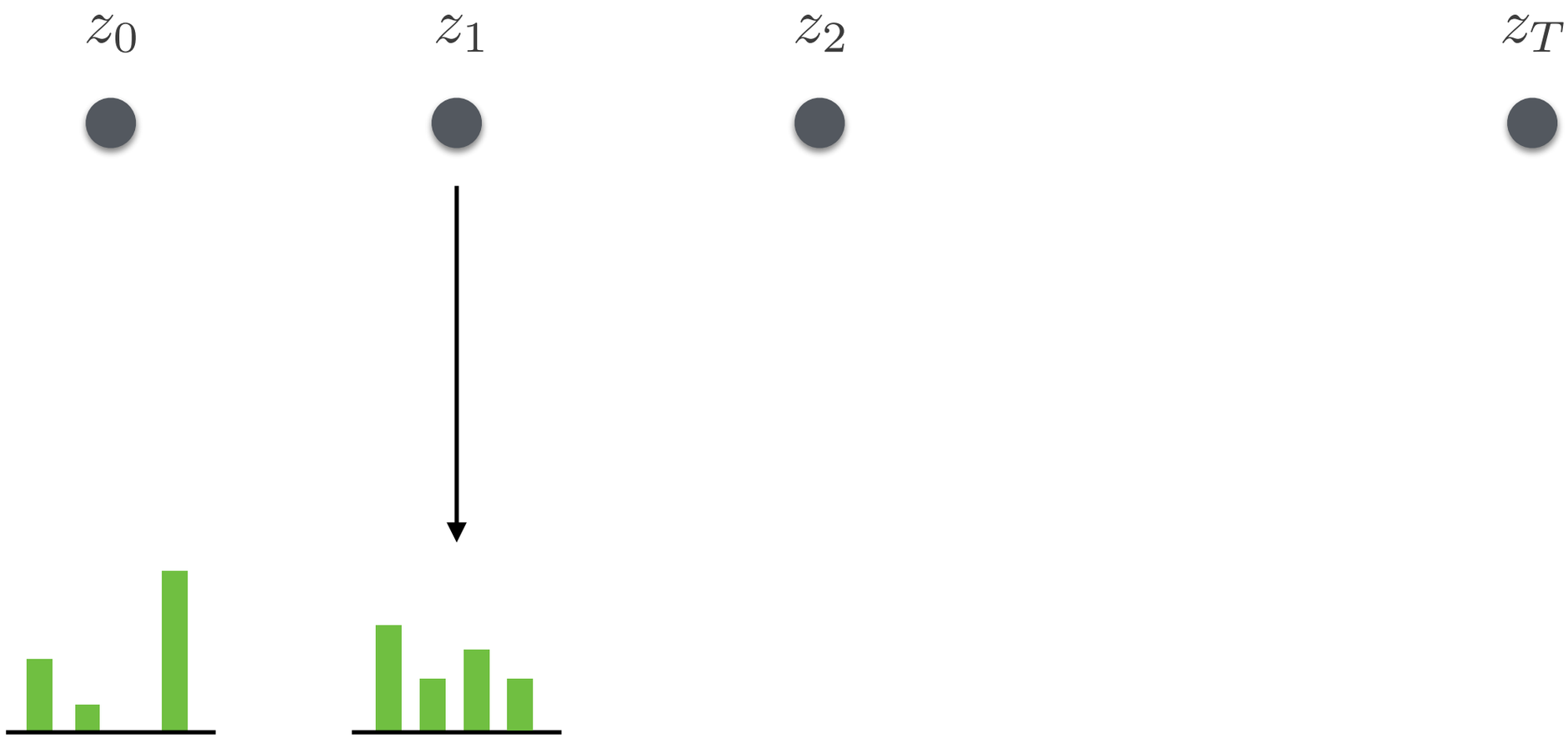


"Improved" distribution

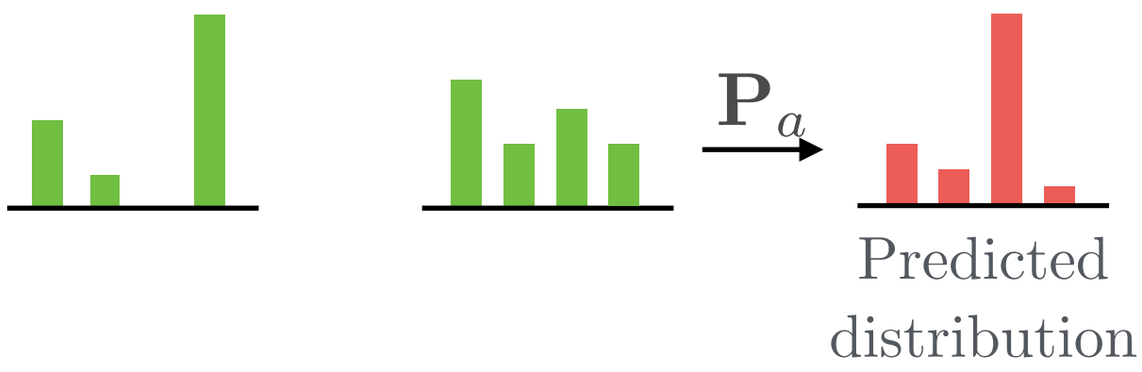
An old trick



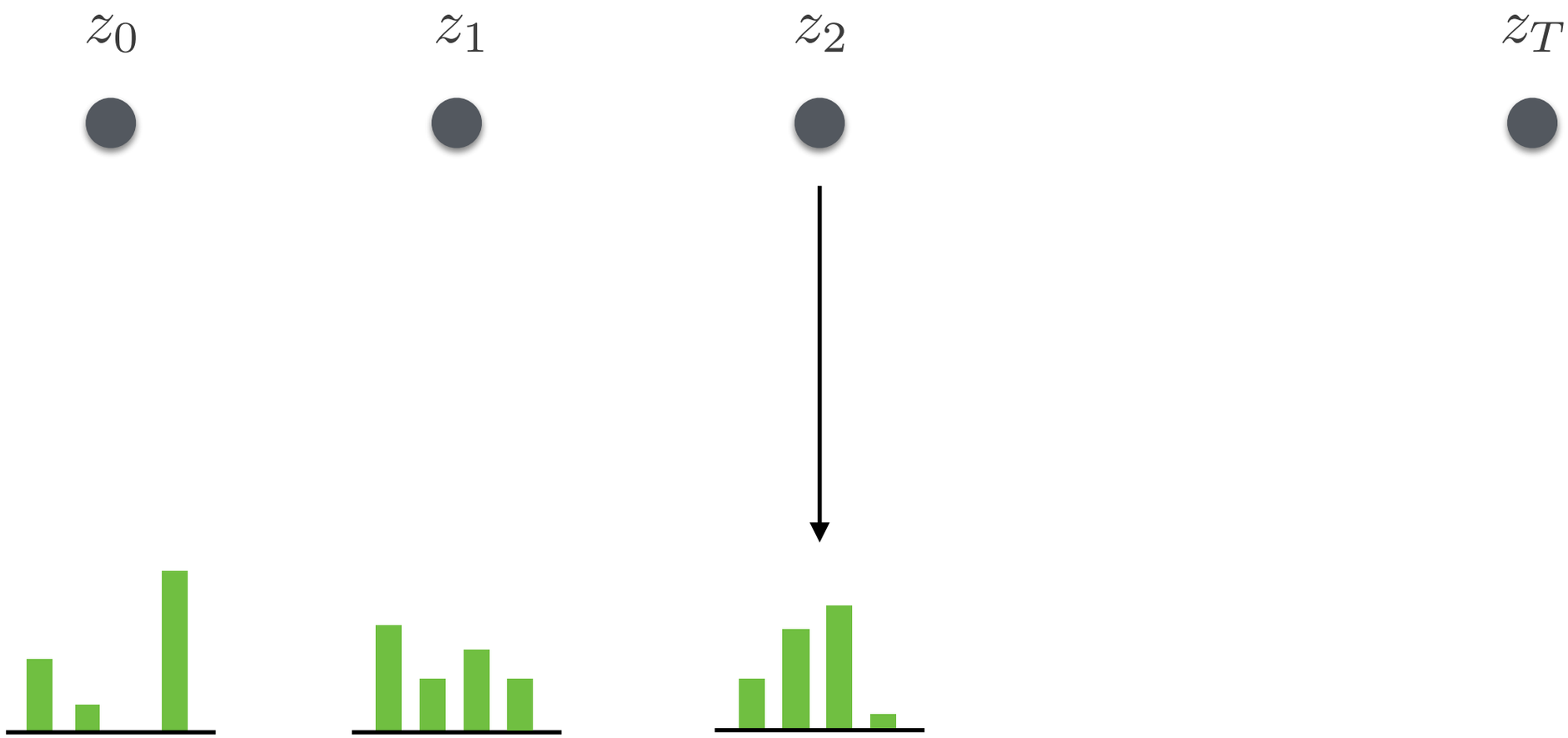
An old trick



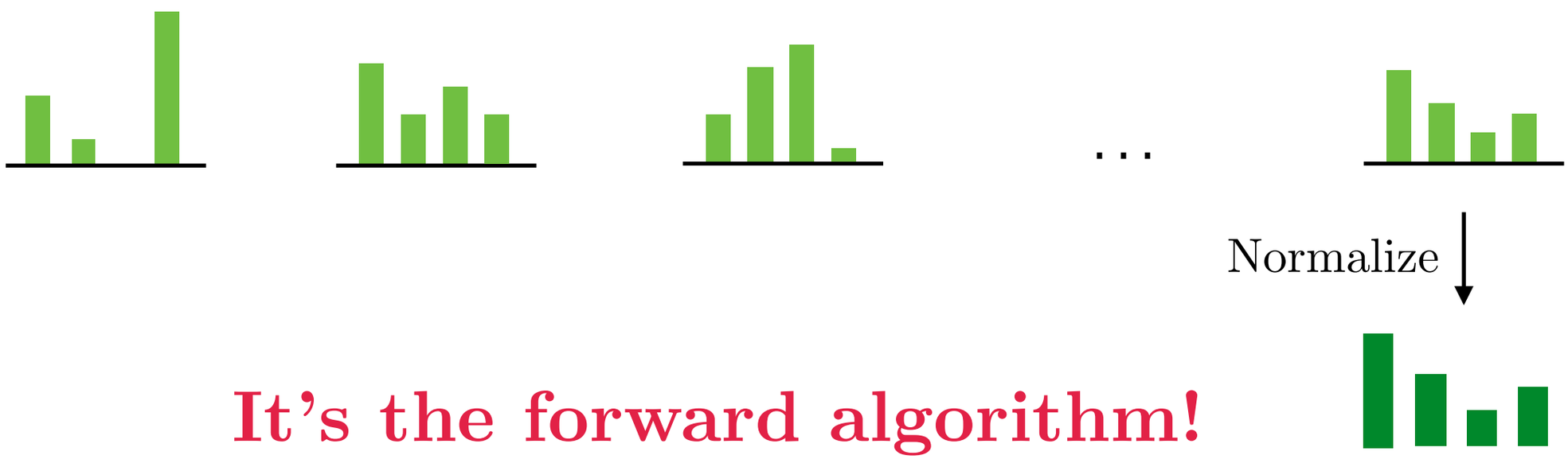
An old trick



An old trick



An old trick



It's the forward algorithm!

Let's do this:

- Initial distribution:

$$\boldsymbol{\alpha}_0 = \boldsymbol{\mu}_0 = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$$

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

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

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

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

Let's do this:

- Take action "Listen":

$$\hat{\boldsymbol{\alpha}}_1 = \boldsymbol{\alpha}_0^\top \mathbf{P}_L$$

$$= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$$

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

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

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

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

Let's do this:

- Consider observation "R":

$$\boldsymbol{\alpha}_1^\top = \hat{\boldsymbol{\alpha}}_1 \text{diag}(\mathbf{O}_L(R \mid \cdot))$$

$$= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.15 & 0 \\ 0 & 0.85 \end{bmatrix}$$

$$= \begin{bmatrix} 0.075 & 0.425 \end{bmatrix}$$

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

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

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

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

Let's do this:

- Again, action "Listen":

$$\hat{\boldsymbol{\alpha}}_2 = \boldsymbol{\alpha}_1^\top \mathbf{P}_L$$

$$= \begin{bmatrix} 0.075 & 0.425 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.075 & 0.425 \end{bmatrix}$$

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

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

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

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

Let's do this:

- ... and observation "R":

$$\boldsymbol{\alpha}_2^\top = \hat{\boldsymbol{\alpha}}_2 \text{diag}(\mathbf{O}_L(R \mid \cdot))$$

$$= \begin{bmatrix} 0.075 & 0.425 \end{bmatrix} \begin{bmatrix} 0.15 & 0 \\ 0 & 0.85 \end{bmatrix}$$

$$= \begin{bmatrix} 0.011 & 0.361 \end{bmatrix}$$

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

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

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

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

Let's do this:

- Finally, normalize:

$$\boldsymbol{\mu}_{2|0:2} = \frac{\boldsymbol{\alpha}_2^\top}{\|\boldsymbol{\alpha}_2\|_1}$$

$$= \frac{1}{0.373}\begin{bmatrix}0.011 & 0.361\end{bmatrix}$$

$$= \begin{bmatrix}0.03 & 0.97\end{bmatrix}$$

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

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

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

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

One extra observation?

- What if we make one extra observation?
  - This time, "L"?

Step

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

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

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

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

One extra observation?

- What if we make one extra observation?
  - This time, "L"?

$\boldsymbol{\mu}_{2|0:2} \rightarrow \boldsymbol{\mu}_{2|0:2}\mathbf{P}_L$

↓

Obs.

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

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

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

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

One extra observation?

- What if we make one extra observation?
  - This time, "L"?

$$\boldsymbol{\mu}_{2|0:2} \rightarrow \boldsymbol{\mu}_{2|0:2}\mathbf{P}_L$$

$$\rightarrow \boldsymbol{\mu}_{2|0:2}\mathbf{P}_L \text{diag}(\mathbf{O}_L(L \mid \cdot))$$

↓

Norm.

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

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

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

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

One extra observation?

- What if we make one extra observation?
  - This time, "L"?

$$\boldsymbol{\mu}_{2|0:2} \rightarrow \boldsymbol{\mu}_{2|0:2}\mathbf{P}_L$$

$$\rightarrow \boldsymbol{\mu}_{2|0:2}\mathbf{P}_L \text{diag}(\mathbf{O}_L(L \mid \cdot))$$

$$\rightarrow \frac{\boldsymbol{\mu}_{2|0:2}\mathbf{P}_L \text{diag}(\mathbf{O}_L(L \mid \cdot))}{\|\boldsymbol{\mu}_{2|0:2}\mathbf{P}_L \text{diag}(\mathbf{O}_L(L \mid \cdot))\|_1}$$

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

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

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

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

One extra observation?

- What if we make one extra observation?
  - This time, "L"?

$$\boldsymbol{\mu}_{3|0:3} = \begin{bmatrix} 0.15 & 0.85 \end{bmatrix}$$

Probability that tiger is on the right

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

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

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

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

The belief

- We call the distribution $\boldsymbol{\mu}_{t|0:t}$ the **belief** at time t
  - We will denote it as $\boldsymbol{b}_t$
  - $\boldsymbol{b}_t$ is a distribution over $\mathcal{X}$
  - $\boldsymbol{b}_t(x)$ is the agent's belief that $\mathrm{x}_t = x$, given all the history, i.e.,

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

The belief

- We can update the belief using the previous equation
  - ... after executing action a...
  - ... after making observation z...

$$\boldsymbol{b}_{t+1} = \frac{\boldsymbol{b}_t \mathbf{P}_a \text{diag}(\mathbf{O}_a(z \mid \cdot))}{\|\boldsymbol{b}_t \mathbf{P}_a \text{diag}(\mathbf{O}_a(z \mid \cdot))\|_1}$$

or, component-wise, ...

$$\boldsymbol{b}_{t+1}(x') = \frac{\sum_{x \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x')}{\sum_{x, x'' \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_a(x'' \mid x) \mathbf{O}_a(z \mid x'')}$$

Belief update

The belief

- We can update the belief using the previous equation
  - ... after executing action a...
  - ... after making observation z...

$$\boldsymbol{b}_{t+1} = \frac{\boldsymbol{b}_t \mathbf{P}_a \mathrm{diag}(\mathbf{O}_a(z \mid \cdot))}{\|\boldsymbol{b}_t \mathbf{P}_a \mathrm{diag}(\mathbf{O}_a(z \mid \cdot))\|_1}$$

or, component-wise, ...

$$\boldsymbol{b}_{t+1}(x') = \frac{\sum_{x \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x')}{\sum_{x, x'' \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_a(x'' \mid x) \mathbf{O}_a(z \mid x'')} \quad \mathbf{B}(\boldsymbol{b}_t, z, a)$$

The belief

- The belief at time-step $t+1$ depends only on:

$$\boldsymbol{b}_{t+1}(x') = \frac{\sum_{x \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_{a_t}(x' \mid x) \mathbf{O}_{a_t}(z_{t+1} \mid x')}{\sum_{x,x'' \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_{a_t}(x'' \mid x) \mathbf{O}_{a_t}(z_{t+1} \mid x'')}$$

The belief at time step t

The action at time step t

The observation at time step $t+1$

... only quantities from time t

The belief

- The belief at time-step $t+1$ depends only on:

$$\boldsymbol{b}_{t+1}(x') = \frac{\sum_{x \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_{a_t}(x' \mid x) \mathbf{O}_{a_t}(z_{t+1} \mid x')}{\sum_{x,x'' \in \mathcal{X}} \boldsymbol{b}_t(x) \mathbf{P}_{a_t}(x'' \mid x) \mathbf{O}_{a_t}(z_{t+1} \mid x'')}$$

  - ... the belief at time t
  - ... the action at time t
  - ... the observation at time $t+1$

- The belief at time t **summarizes the history** up to time t!

The belief is Markov!

Partially observable MDP

- 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, $\mathbf{c}$

Belief MDP

- Described by:
  - State space, $\mathcal{B}$ — Set of all beliefs
  - Action space, $\mathcal{A}$ — Same set of actions
  - Transition probabilities $\mathbf{P}_B$ (from the belief update)
  - Immediate cost function c_B

$$c_B(\boldsymbol{b}, a) = \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

Optimality

Cost-to-go function

- **We can adapt MDP results to POMDPs through belief-MDPs**
- Optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[c_B(\boldsymbol{b}, a) + \gamma \sum_{\boldsymbol{b}'} \mathbf{P}_B(\boldsymbol{b}' \mid \boldsymbol{b}, a) J^*(\boldsymbol{b}') \right]$$

$$c_B(\boldsymbol{b}, a) = \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

Cost-to-go function

- **We can adapt MDP results to POMDPs through belief-MDPs**
- Optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[c_B(\boldsymbol{b}, a) + \gamma \sum_{\boldsymbol{b}'} \mathbf{P}_B(\boldsymbol{b}' \mid \boldsymbol{b}, a) J^*(\boldsymbol{b}') \right]$$

$$\mathbf{P}_B(\boldsymbol{b}' \mid \boldsymbol{b}, a) = \sum_{z \in \mathcal{Z}} \sum_{x, x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x, a) \mathbf{O}_a(z \mid x', a) \mathbb{I}[\boldsymbol{b}' = \mathbf{B}(\boldsymbol{b}, z, a)]$$

Cost-to-go function

- **We can adapt MDP results to POMDPs through belief-MDPs**
- Optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \underbrace{\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \Big[c(x,a) }_{c_B(\boldsymbol{b},a)} + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*(\mathbf{B}(\boldsymbol{b}, z, a)) \Big]$$

Cost-to-go function

- **We can adapt MDP results to POMDPs through belief-MDPs**
- Optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*(\mathbf{B}(\boldsymbol{b}, z, a)) \right]$$

$$\mathbf{P}_B(\boldsymbol{b}' \mid \boldsymbol{b}, a)$$

Cost-to-go function

- **We can adapt MDP results to POMDPs through belief-MDPs**
- Optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*(\mathbf{B}(\boldsymbol{b}, z, a) \right]$$

$\mathbf{T}$

- Operator $\mathbf{T}$ transforms arbitrary Js into new Js:

$$(\mathbf{T}J)(\boldsymbol{b}) = \min_a \left[c_B(\boldsymbol{b}, a) + \gamma (\mathbf{P}_B(\boldsymbol{b}, a) J)(\boldsymbol{b}') \right]$$

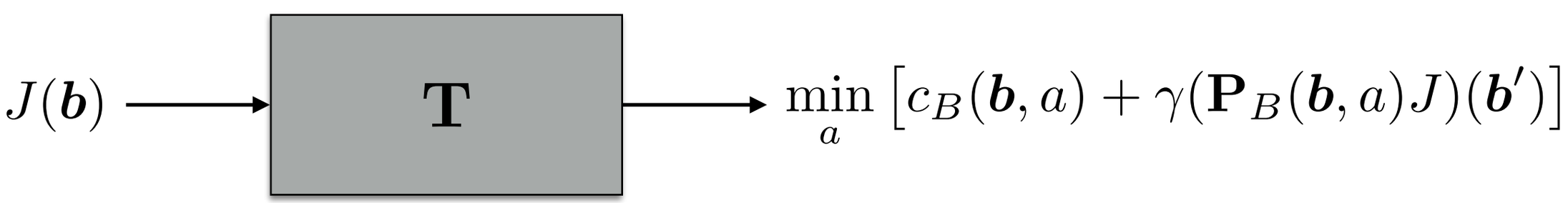
Cost-to-go function

- **We can adapt MDP results to POMDPs through belief-MDPs**
- Optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*(\mathbf{B}(\boldsymbol{b}, z, a) \right]$$

$\mathbf{T}$

- Operator $\mathbf{T}$ transforms arbitrary Js into new Js:



… however…

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[c_B(\boldsymbol{b}, a) + \gamma \sum_{\boldsymbol{b}'} \mathbf{P}_B(\boldsymbol{b}' \mid \boldsymbol{b}, a) J^*(\boldsymbol{b}') \right]$$

How do we represent this?

Representing J*

- J^* is a function defined in $\mathbb{R}^n$ (belief)
- We cannot represent it explicitly
- Therefore,

Input: Belief MDP $\mathcal{M} = (\mathcal{B}, \mathcal{A}, \{\mathbf{P}_B(a)\}, c_B, \gamma)$

Input: ...erance ϵ

1: Initi... $k = 0$

2: Initialize $J_0(\boldsymbol{b})$ ... 0, for all $\boldsymbol{b}$

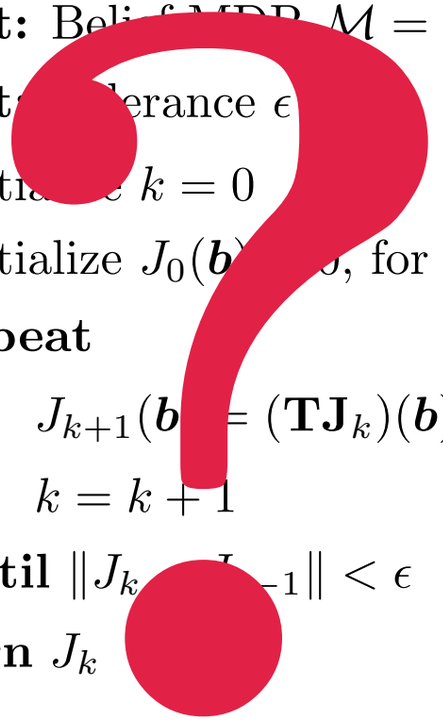
3: **repeat**

4: $J_{k+1}(\boldsymbol{b})$ ... $= (\mathbf{T}\mathbf{J}_k)(\boldsymbol{b})$, for all $\boldsymbol{b}$

5: $k = k + 1$

6: **until** $\|J_k$ ... $J_{k-1}\| < \epsilon$

return J_k



Heuristic solutions

Idea: Use the MDP

- "Under" a POMDP there is an MDP

$$(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\mathbf{P}_a\}, \{\mathbf{O}_a\}, c)$$

  - MDP can be solved efficiently
  - Why not use the MDP solution?

MDP heuristics

MLS heuristic

- Let π_{MDP} be the optimal MDP policy
- At time step t, we have a belief

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}(x_1) & \boldsymbol{b}(x_2) & \dots & \boldsymbol{b}(x_N) \end{bmatrix}$$

Probability that state is x_1

Probability that state is x_2

MLS heuristic

- Let π_{MDP} be the optimal MDP policy

- At time step t, we have a belief

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}(x_1) & \boldsymbol{b}(x_2) & \dots & \boldsymbol{b}(x_N) \end{bmatrix}$$

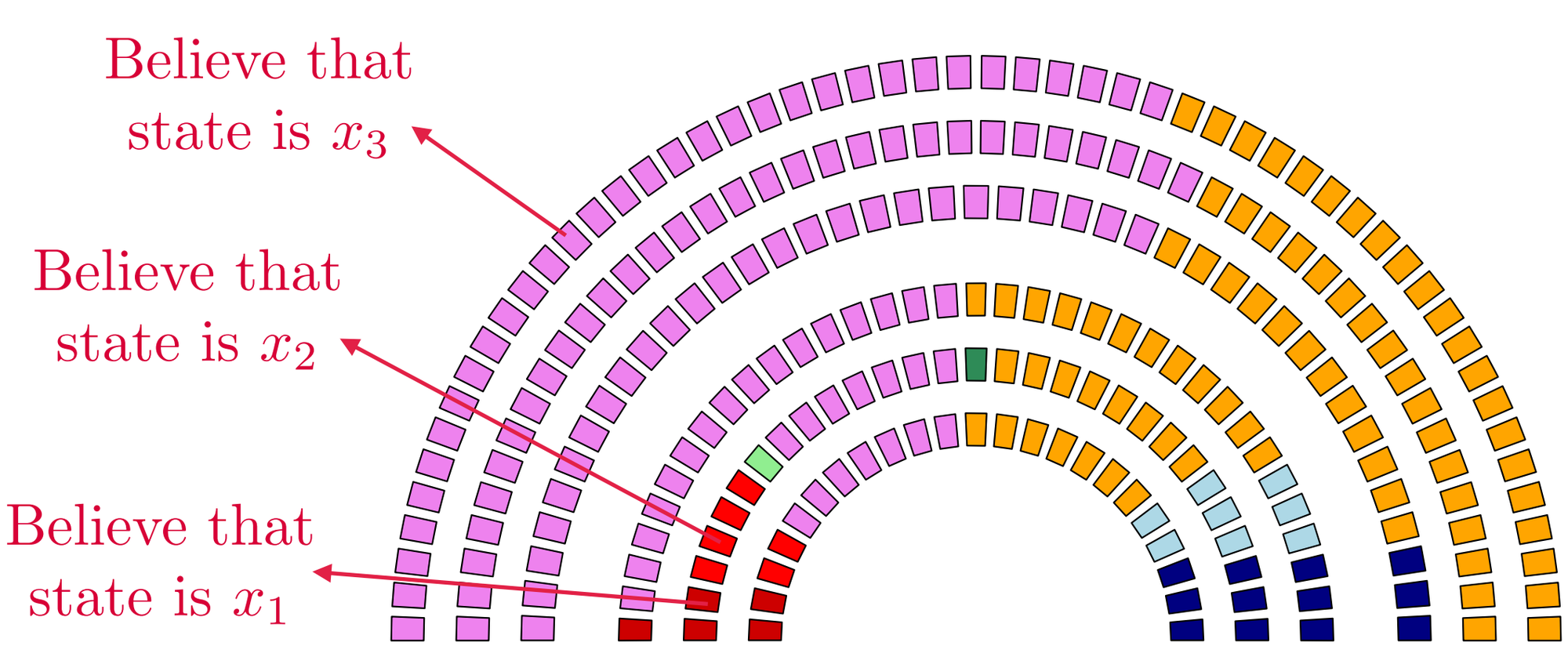
Most likely state

- Select the most likely state (state with largest probability)

- Execute corresponding action — $\pi_{\text{MDP}}(x_2)$

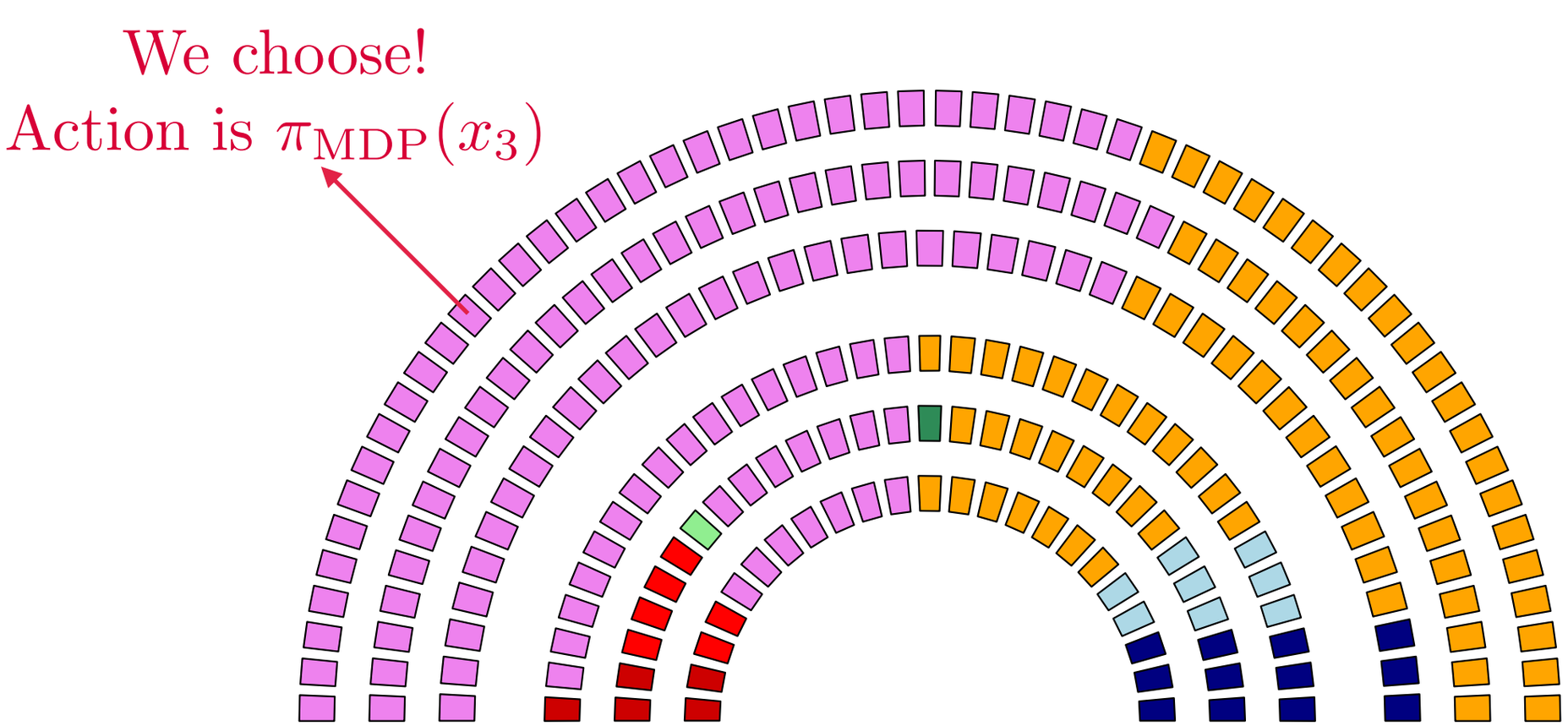
MLS heuristic

- An analogy:
  - The belief as the parliament



MLS heuristic

- An analogy:
  - The belief as the parliament
  - The party with the largest representation chooses the action



MLS heuristic

- The MLS (Most Likely State) heuristic is, then

$$\pi_{\text{MLS}}(\boldsymbol{b}) = \pi_{\text{MDP}}(\operatorname*{argmax}_{x \in \mathcal{X}} \boldsymbol{b}(x))$$

The tiger problem

Tiger POMDP

- $\mathcal{X} = \{L, R\}$
- $\mathcal{A} = \{OL, OR, L\}$
- $\mathcal{Z} = \{L, R\}$
- $\mathbf{P}_{OL} = \mathbf{P}_{OR} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ $\mathbf{P}_{L} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- $\mathbf{O}_{OL} = \mathbf{O}_{OR} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ $\mathbf{O}_{L} = \begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix}$
- $\mathbf{C} = \begin{bmatrix} 1 & 0 & 0.1 \\ 0 & 1 & 0.1 \end{bmatrix}$

What is the belief space?

- The POMDP can be in two possible states:
  - Tiger left
  - Tiger right
- Belief $\boldsymbol{b}$ is a vector

$$\boldsymbol{b} = \begin{bmatrix} \mathbb{P}\left[\text{Tiger left}\right] & \mathbb{P}\left[\text{Tiger right}\right] \end{bmatrix}$$

What is the belief space?

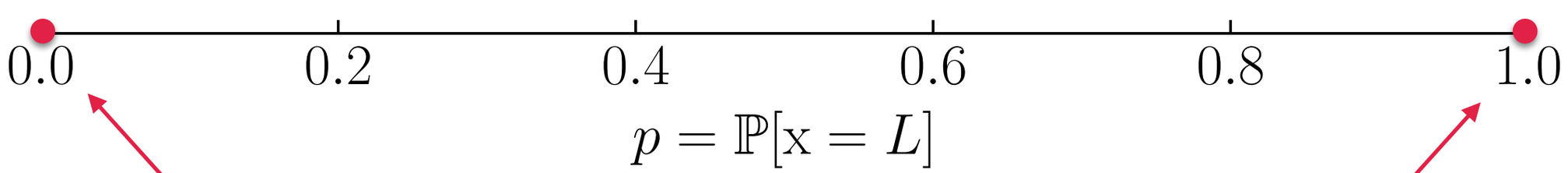
- The POMDP can be in two possible states:
  - Tiger left
  - Tiger right
- Belief $\boldsymbol{b}$ is a vector

$$\boldsymbol{b} = \begin{bmatrix} \mathbb{P}\left[\text{Tiger left}\right] & 1 - \mathbb{P}\left[\text{Tiger left}\right] \end{bmatrix}$$

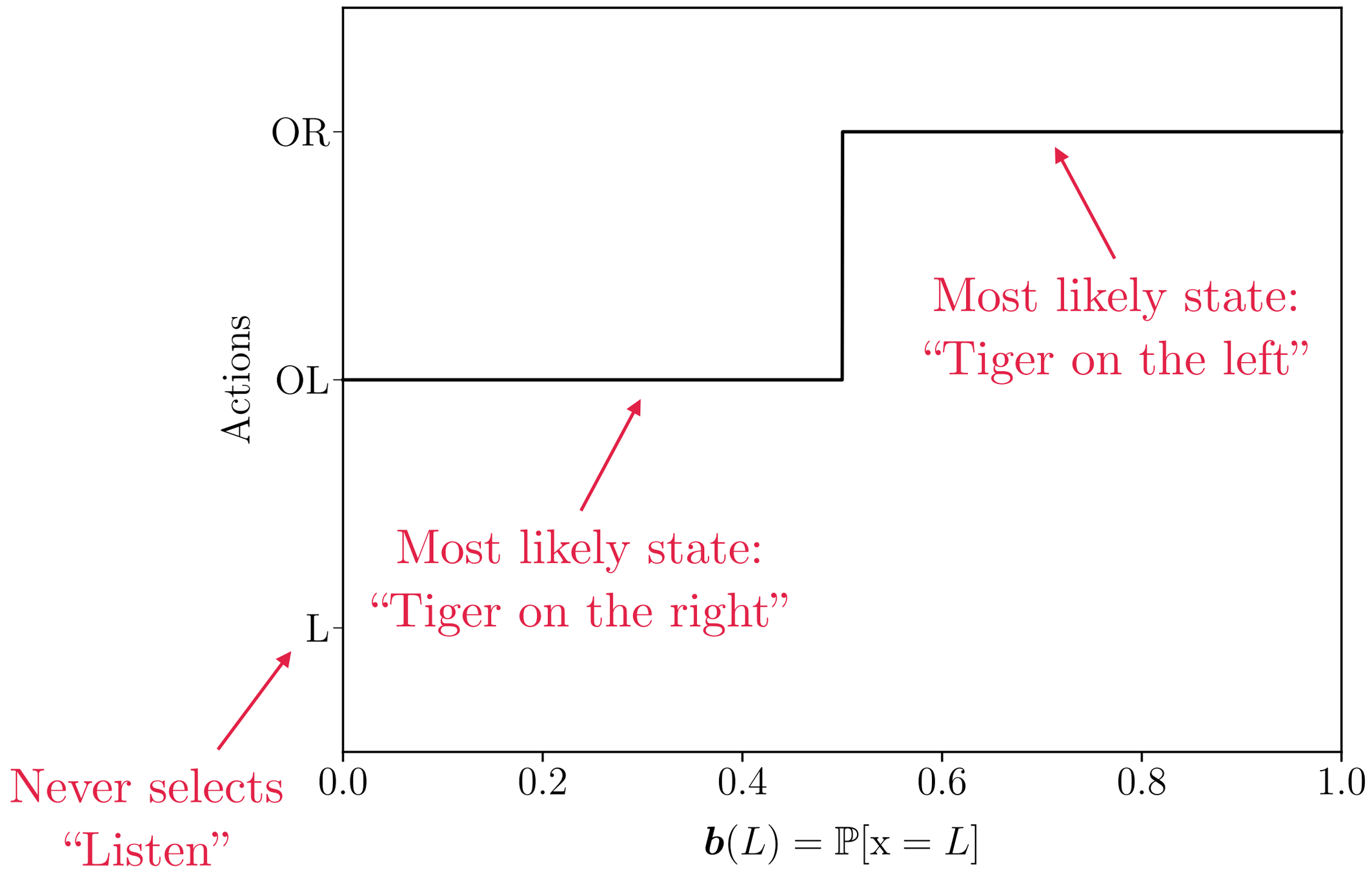
Just a single number

What is the belief space?

- We can represent it as a number in [0, 1]:



MLS heuristic



AV heuristic

- Let π_{MDP} be the optimal MDP policy
- At time step t, we have a belief

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}(x_1) & \boldsymbol{b}(x_2) & \dots & \boldsymbol{b}(x_N) \end{bmatrix}$$

State x_1 votes for $\pi_{\text{MDP}}(x_1)$

State x_2 votes for $\pi_{\text{MDP}}(x_2)$

AV heuristic

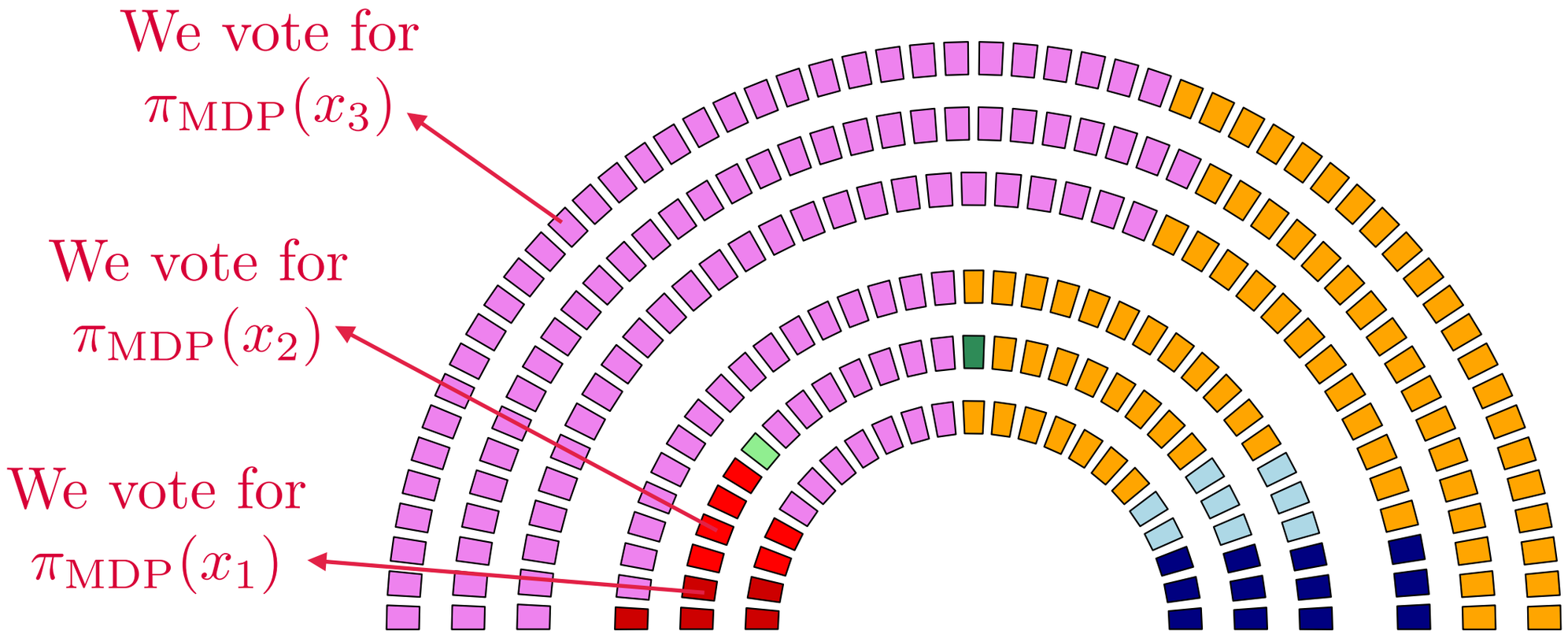
- Let π_{MDP} be the optimal MDP policy
- At time step t, we have a belief

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}(x_1) & \boldsymbol{b}(x_2) & \dots & \boldsymbol{b}(x_N) \end{bmatrix}$$

- States vote proportionally to their probability
- Choose the most voted action

MLS heuristic

- An analogy:
  - The belief as the parliament
  - Each party votes for an action — most voted action wins

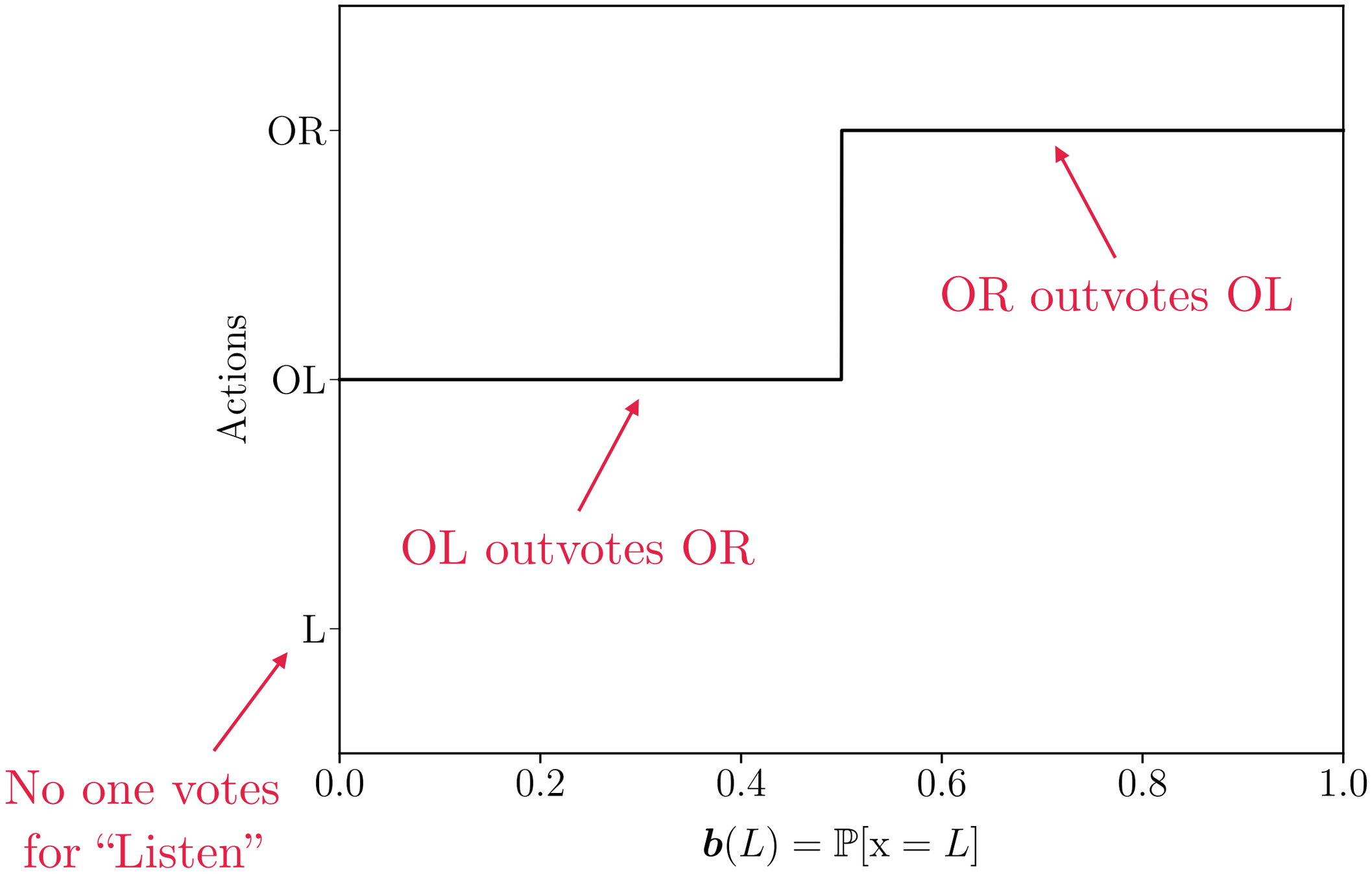


AV heuristic

- The AV (Action Voting) heuristic is, then

$$\pi_{\mathrm{AV}}(\boldsymbol{b}) = \operatorname*{argmax}_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \mathbb{I}(a = \pi_{\mathrm{MDP}}(x))$$

AV heuristic



Let's consider state values...

Q-MDP heuristic

- **Optimistic assumption (at time t):** partial observability is over at time $t + 1$
- What is the cost-to-go?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x, a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*(\mathbf{B}(\boldsymbol{b}, z, a)) \right]$$

No partial observability

Q-MDP heuristic

- **Optimistic assumption (at time t):** partial observability is over at time $t + 1$
- What is the cost-to-go?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x, a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*_{\text{MDP}}(x') \right]$$

Only term that depends on z

Q-MDP heuristic

- **Optimistic assumption (at time t):** partial observability is over at time $t + 1$
- What is the cost-to-go?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x, a) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{P}_a(x' \mid x) J^*_{\text{MDP}}(x') \right]$$

Q-MDP heuristic

- **Optimistic assumption (at time t):** partial observability is over at time $t + 1$
- What is the cost-to-go?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x, a) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{P}_a(x' \mid x) J^*_{\text{MDP}}(x') \right]$$

MDP optimal Q-function

$$Q^*_{\text{MDP}}(x, a)$$

Q-MDP heuristic

- **Optimistic assumption (at time t):** partial observability is over at time $t + 1$
- What is the cost-to-go?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q^*_{\text{MDP}}(x, a)$$

Q-MDP heuristic

- **Optimistic assumption (at time t):** partial observability is over at time $t + 1$
- What is the cost-to-go?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q^*_{\text{MDP}}(x, a)$$

Weighted average of optimal MDP Q-values

Q-MDP heuristic

- The Q-MDP heuristic is, then

$$\pi_{\text{QMDP}}(\boldsymbol{b}) = \arg\min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q^*_{\text{MDP}}(x, a)$$

The tiger problem

- In the tiger problem, MDP agent always knows where tiger is
- Optimal cost-to-go is 0
- Then,

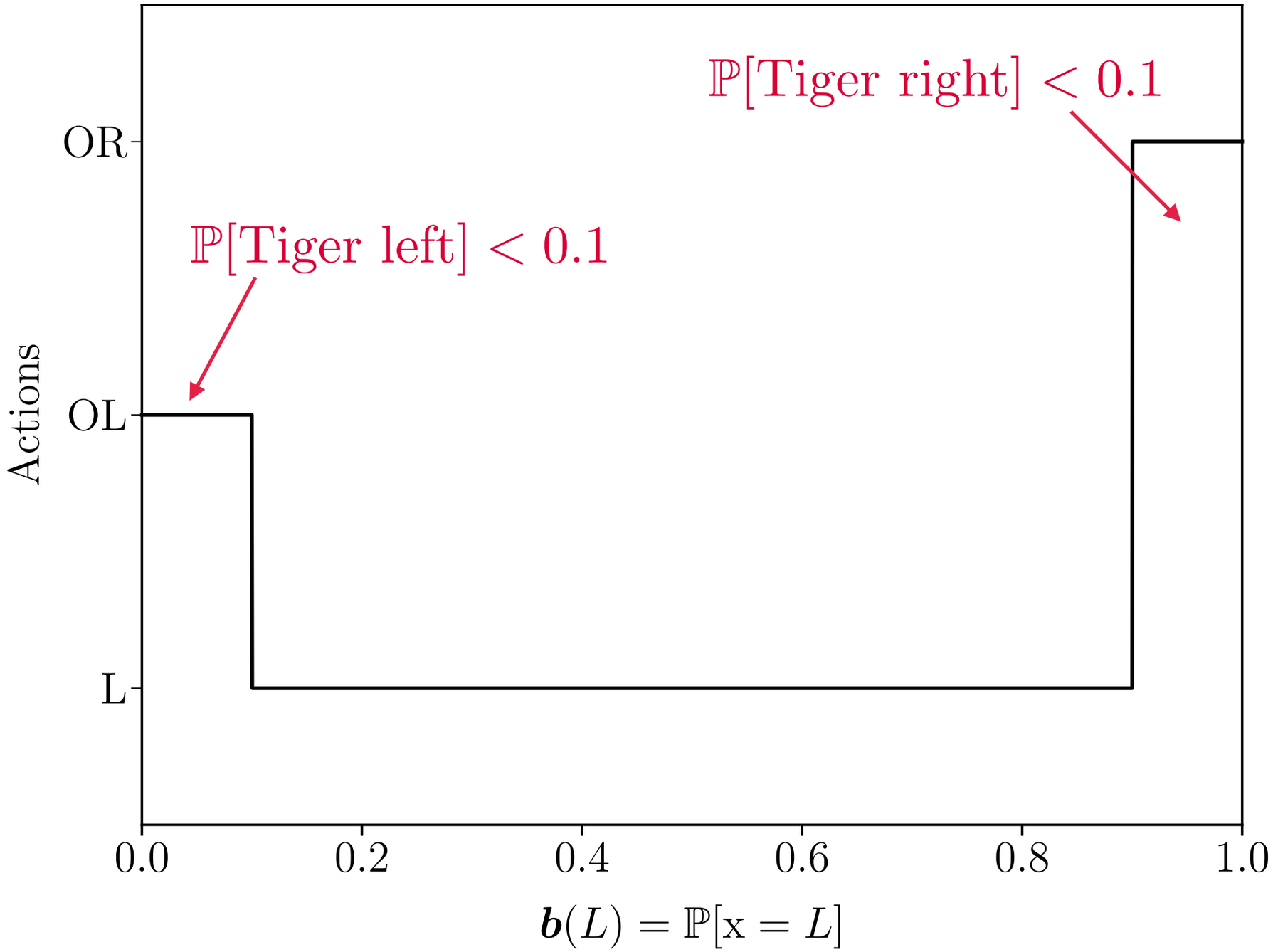
$$\mathbf{Q}^*_{\text{MDP}} = \mathbf{C} = \begin{bmatrix} 1 & 0 & 0.1 \\ 0 & 1 & 0.1 \end{bmatrix}$$

Value for OR is $\mathbb{P}$[Tiger right]

Value for OL is $\mathbb{P}$[Tiger left]

Value for L is 0.1

Q-MDP heuristic



Q-MDP heuristic

- Let's look at Q-MDP in more detail
- Let's start with the optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*(\mathbf{B}(\boldsymbol{b}, z, a)) \right]$$

Q-MDP heuristic

- Let's look at Q-MDP in more detail
- Let's start with the optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x' \in \mathcal{X}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J^*(\mathbf{B}(\boldsymbol{b}, z, a)) \right]$$

Let's split the product

... and replace J^* here

Q-MDP heuristic

- Let's look at Q-MDP in more detail
- Let's start with the optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x, x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') \min_{a' \in \mathcal{A}} \sum_{x'' \in \mathcal{X}} \mathbf{B}(\boldsymbol{b}, z, a)(x'') f(x'', a') \right]$$

Some function of x'' and a'

We now replace $\mathbf{B}(\boldsymbol{b}, z, a)$

Q-MDP heuristic

- Let's look at Q-MDP in more detail
- Let's start with the optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \right.$$

$$\left. \gamma \sum_{z \in \mathcal{Z}} \sum_{x,x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') \min_{a' \in \mathcal{A}} \frac{\sum_{x,x'' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x'' \mid x) \mathbf{O}_a(z \mid x'')}{\sum_{x,x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x')} f(x'', a') \right]$$

These two terms cancel out

Q-MDP heuristic

- Let's look at Q-MDP in more detail
- Let's start with the optimal cost-to-go:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \min_{a' \in \mathcal{A}} \sum_{x, x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') f(x', a') \right]$$

Similar except for the min

- Comparing with the Q-MDP approximation:

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x, x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') \min_{a' \in \mathcal{A}} f(x', a') \right]$$

Q-MDP heuristic

- The position of the minimum ignores partial observability at the next time step

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x,x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') \min_{a' \in \mathcal{A}} f(x', a') \right]$$

Observations out of the min

What if we move the min somewhere else?

Q-MDP heuristic

- We place the min right after $\boldsymbol{b}$

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x,x' \in \mathcal{X}} \boldsymbol{b}(x) \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') \min_{a' \in \mathcal{A}} f(x', a') \right]$$

Q-MDP heuristic

- We place the min right after $\boldsymbol{b}$

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \min_{a' \in \mathcal{A}} \sum_{x' \in \mathcal{X}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') f(x', a') \right]$$

- Factorizing $\boldsymbol{b}$,

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \min_{a' \in \mathcal{A}} \sum_{x' \in \mathcal{X}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') f(x', a') \right]$$

We define a new Q-function
based on this expression

FIB heuristic

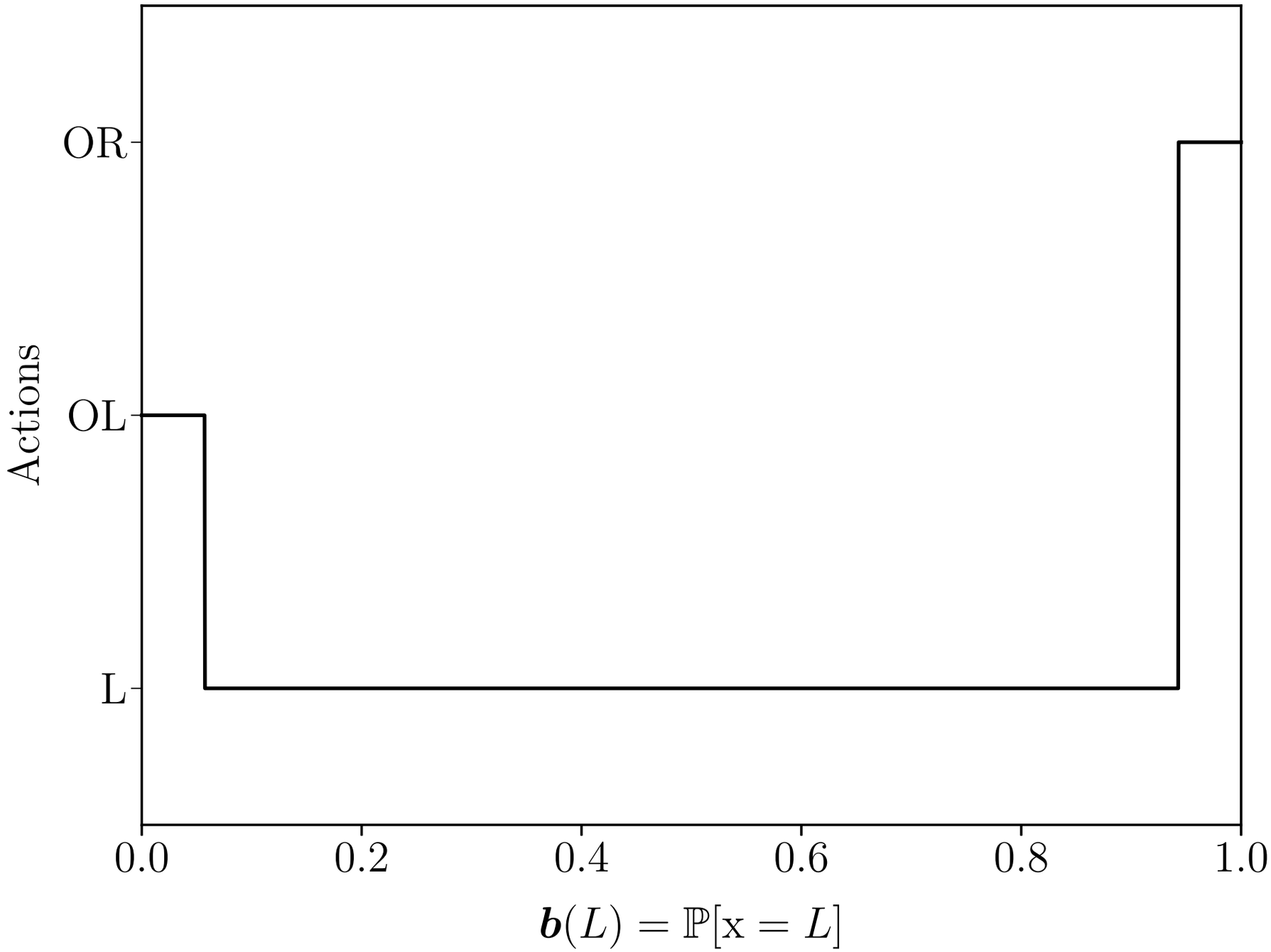
- Equivalent to Q-MDP with modified Q-function:

$$Q_{\text{FIB}}(x,a) = c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \min_{a' \in \mathcal{A}} \sum_{x' \in \mathcal{X}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') Q_{\text{FIB}}(x', a')$$

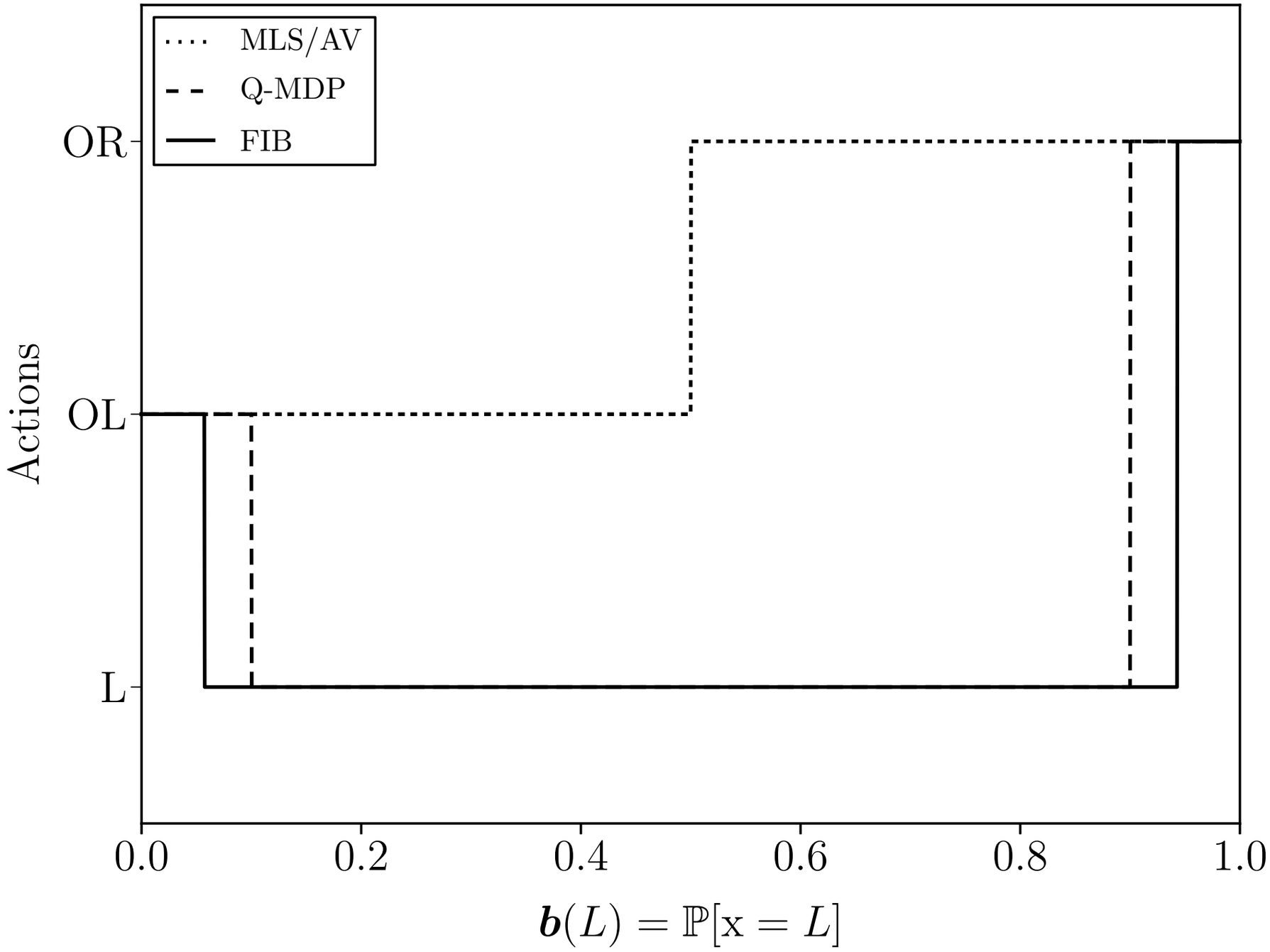
- The FIB heuristic is, then,

$$\pi_{\text{FIB}}(\boldsymbol{b}) = \operatorname*{argmin}_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q_{\text{FIB}}(x,a)$$

FIB heuristic



Comparison


MLS/AV
Q-MDP
FIB
OR
OL
L
Actions
0.0
0.2
0.4
0.6
0.8
1.0
$\boldsymbol{b}(L) = \mathbb{P}[\mathrm{x} = L]$

Heuristic approaches

- MLS heuristic:

$$\pi_{\text{MLS}}(\boldsymbol{b}) = \pi_{\text{MDP}}(\operatorname*{argmax}_{x \in \mathcal{X}} \boldsymbol{b}(x))$$

- AV heuristic:

$$\pi_{\text{AV}}(\boldsymbol{b}) = \operatorname*{argmax}_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \mathbb{I}(a = \pi_{\text{MDP}}(x))$$

- Q-MDP heuristic:

$$\pi_{\text{QMDP}}(\boldsymbol{b}) = \operatorname*{argmin}_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q^*_{\text{MDP}}(x, a)$$

- FIB heuristic:

$$\pi_{\text{FIB}}(\boldsymbol{b}) = \operatorname*{argmin}_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q_{\text{FIB}}(x, a)$$

Can we do better?

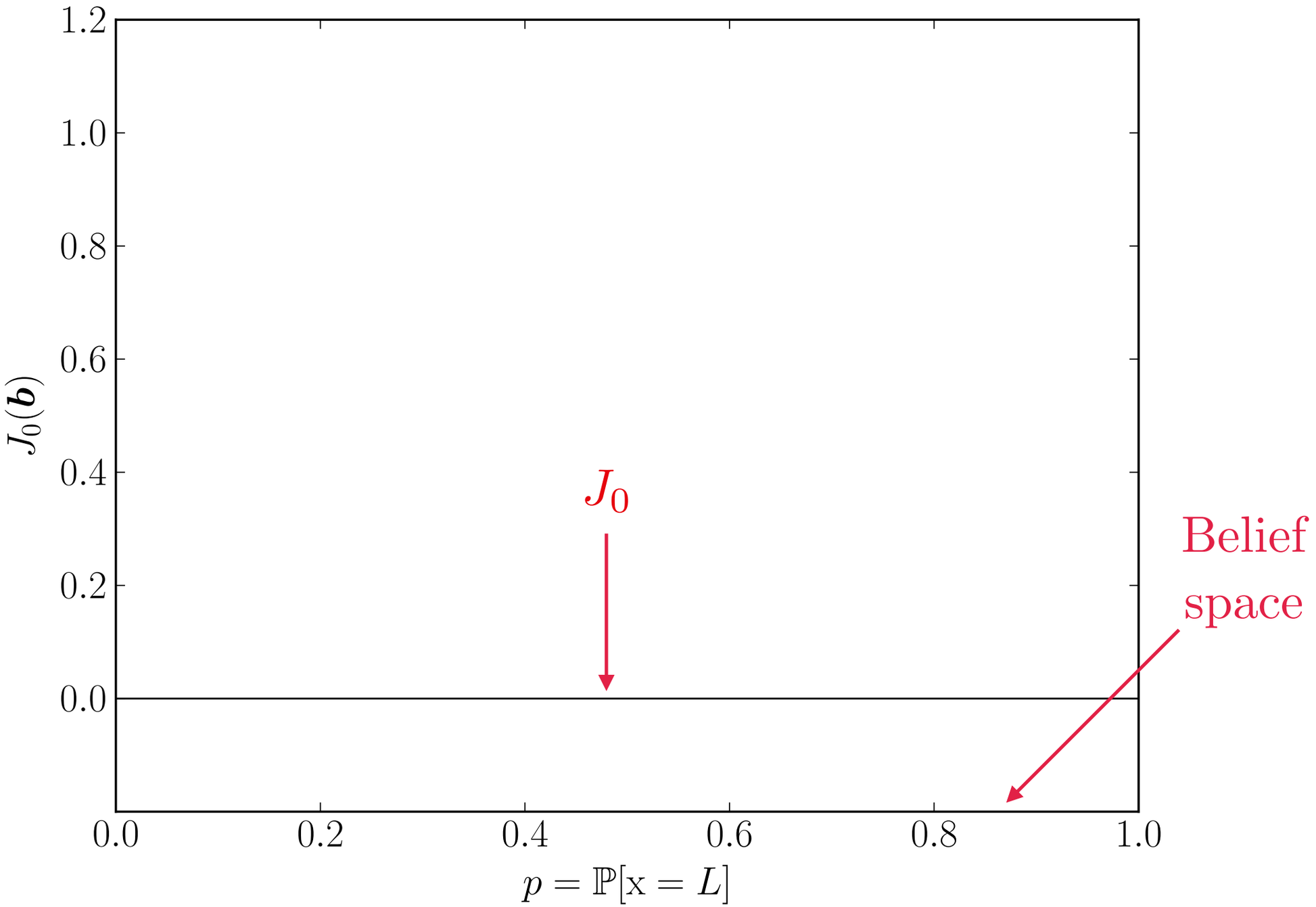
Exact solutions

Value iteration

- Let's try running VI on the tiger problem
  - We start with

$$J_0(\boldsymbol{b}) = 0, \qquad \text{for all } \boldsymbol{b}$$

Value iteration



Value iteration

- Let's try running VI on the tiger problem
  - Iteration 1:

$$J_1(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J_0(\mathbf{B}(\boldsymbol{b}, z, a)) \right]$$

This is zero!

Value iteration

- Let's try running VI on the tiger problem
  - Iteration 1:

$$J_1(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

Value iteration

- Let's try running VI on the tiger problem
  - Iteration 1:

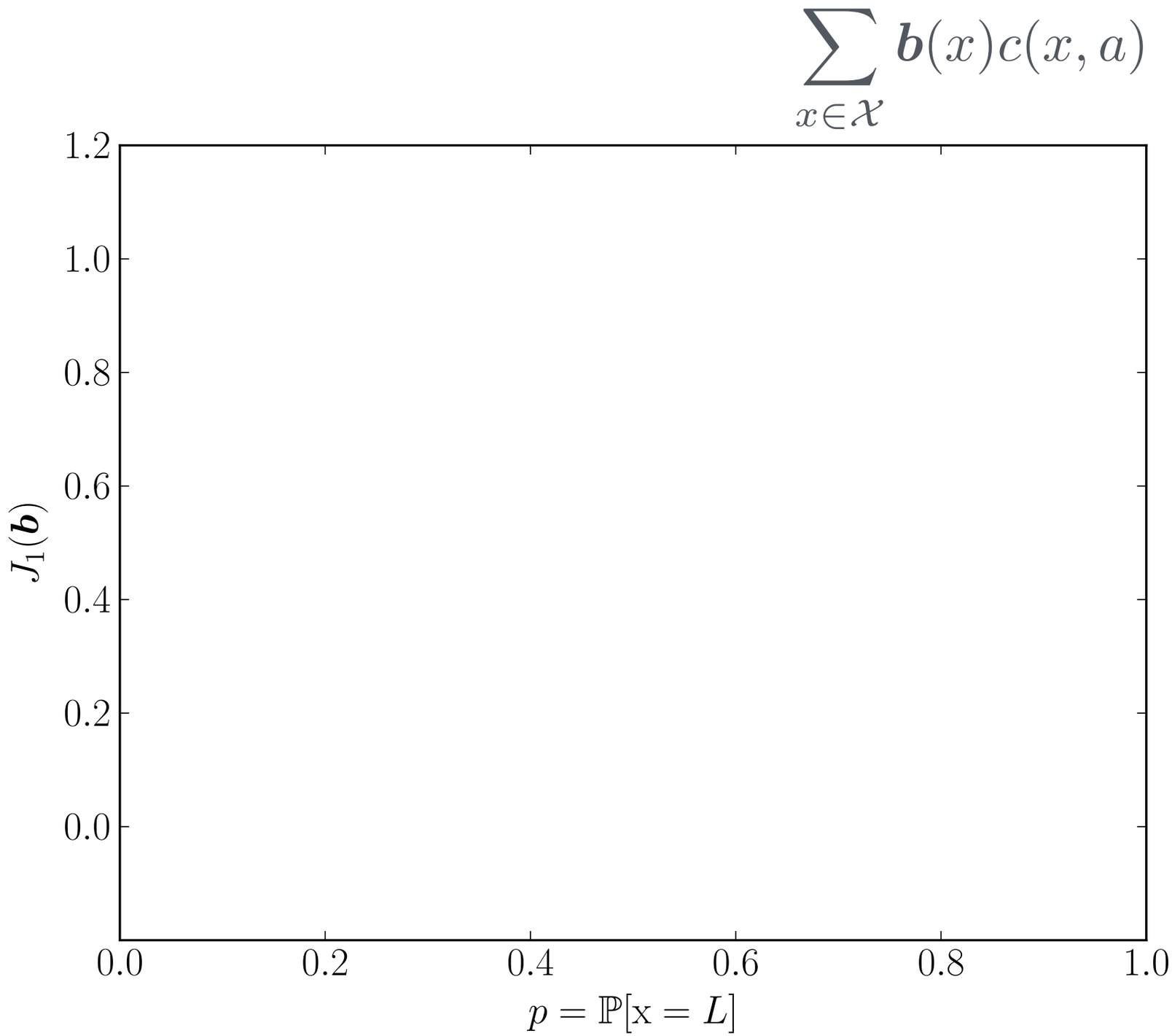
$$J_1(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

We start with this

Value iteration

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

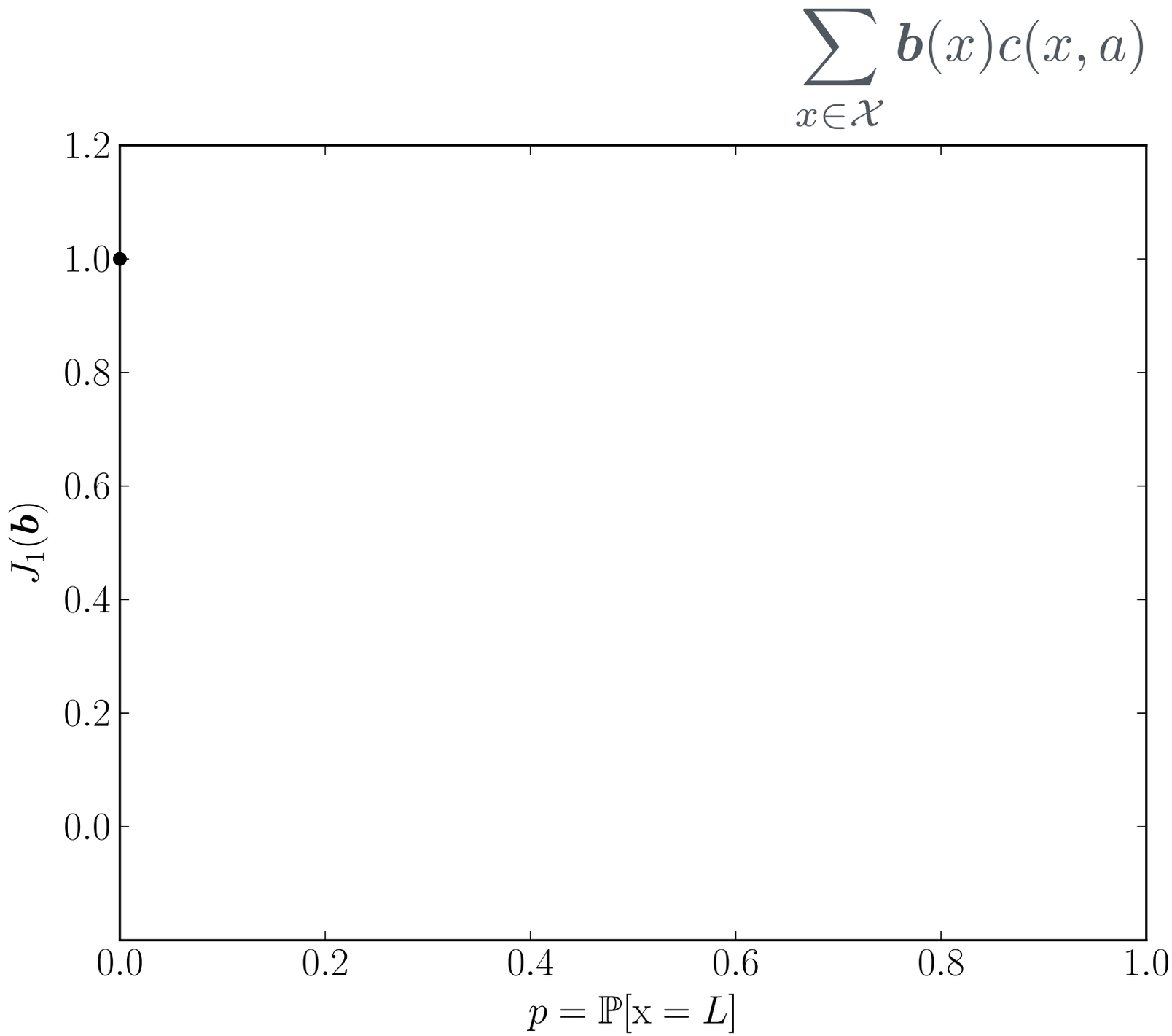
Action OR



Value iteration

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

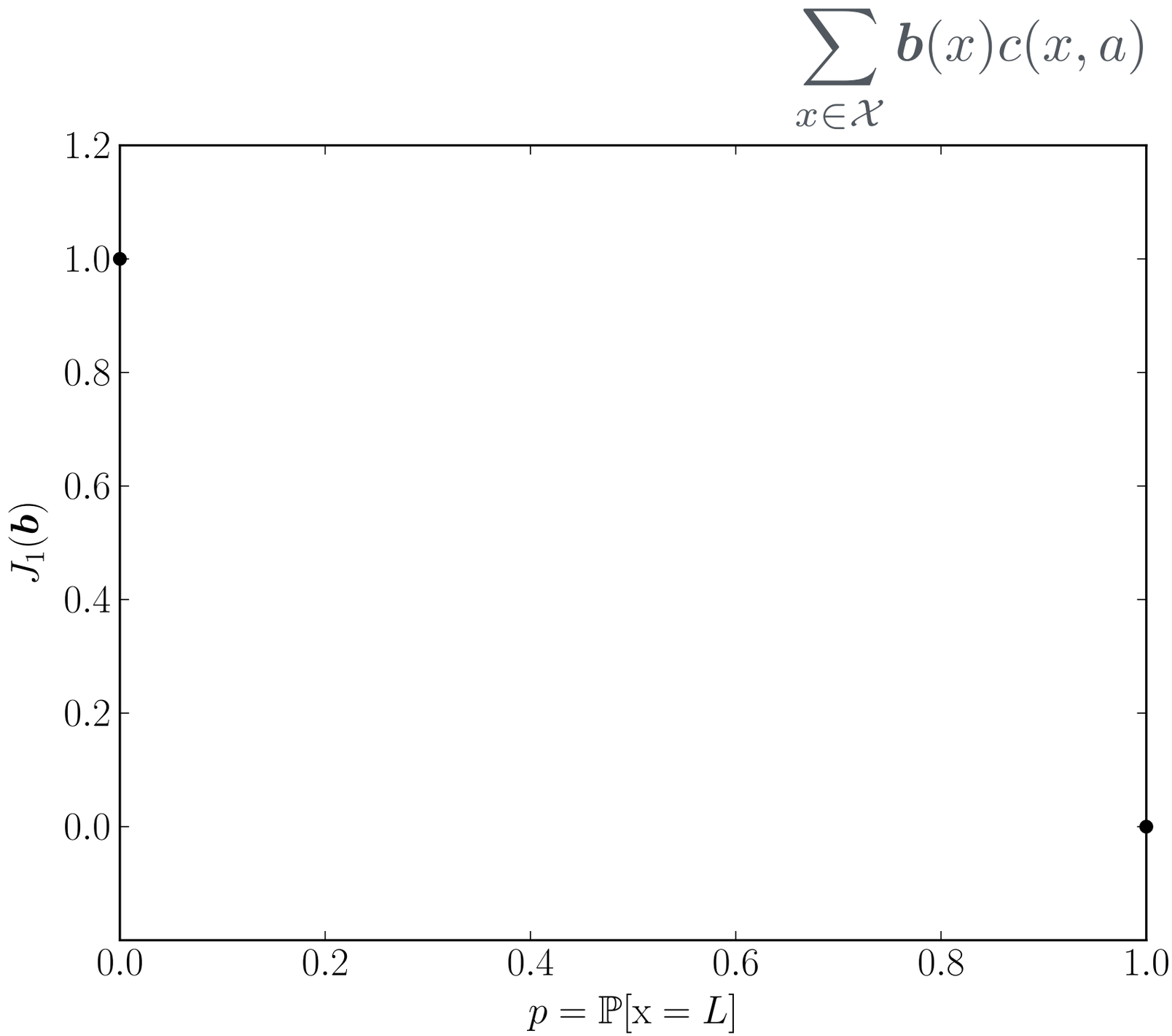
Action OR



Value iteration

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

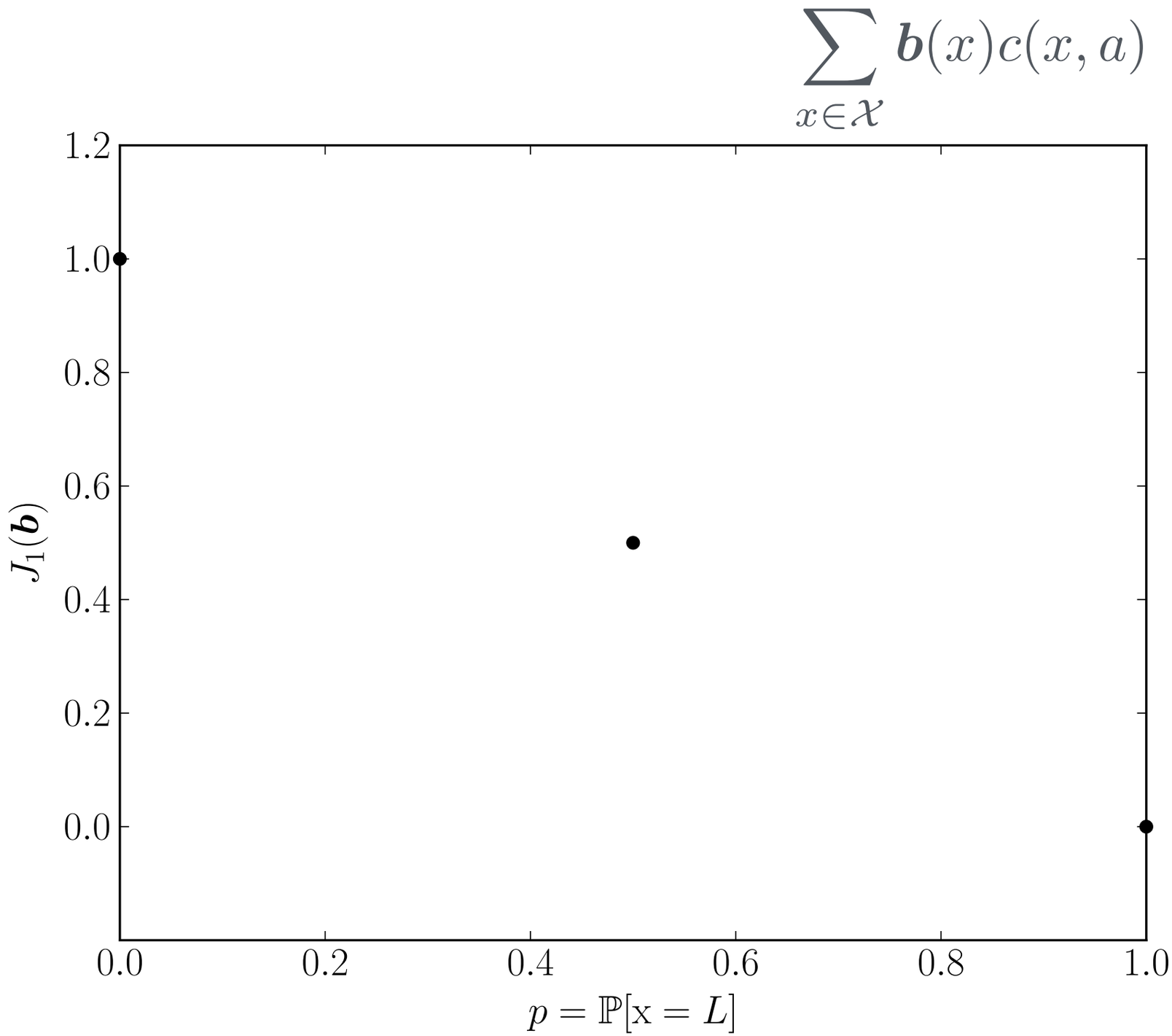
Action OR



Value iteration

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

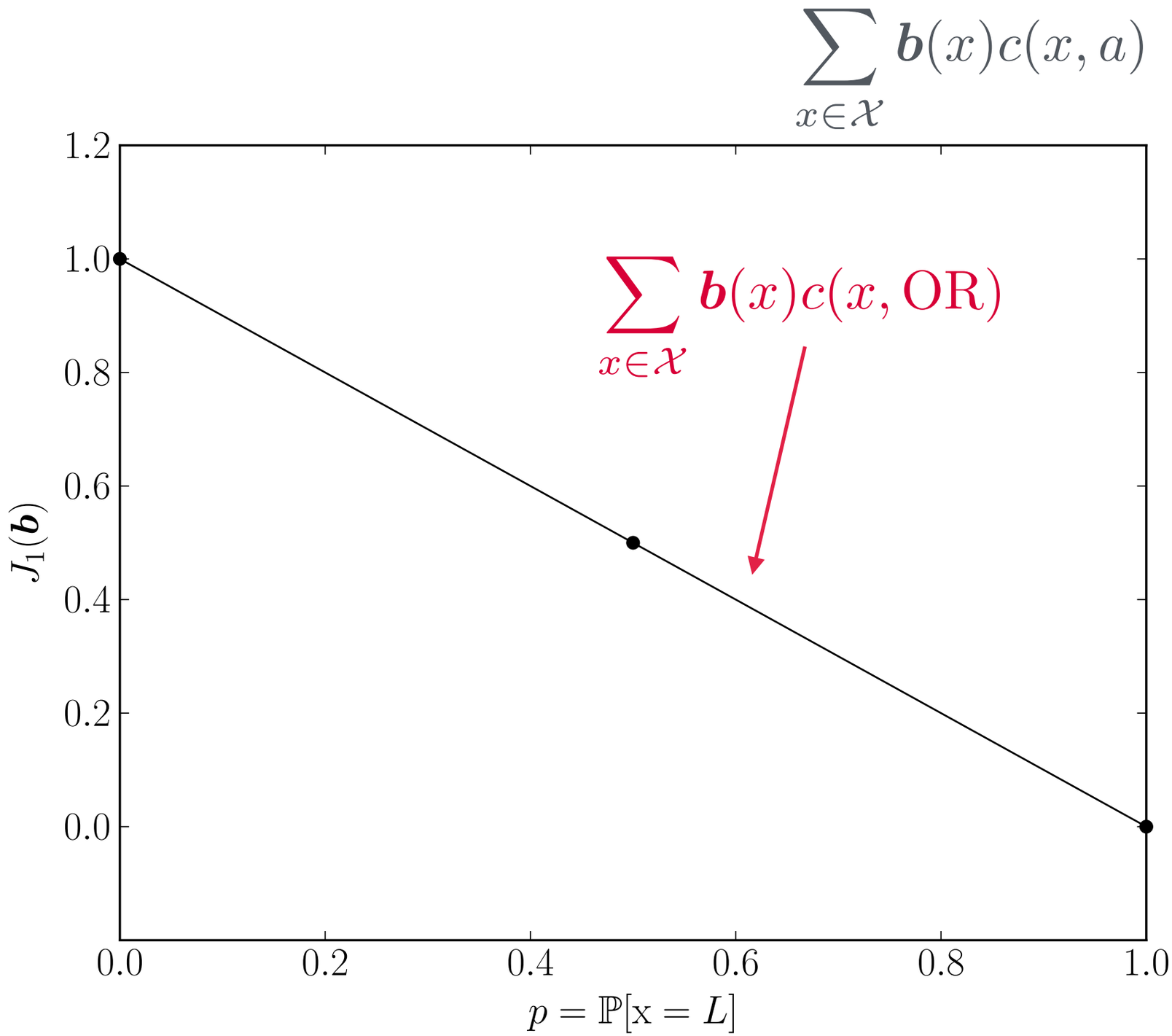
Action OR



Value iteration

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

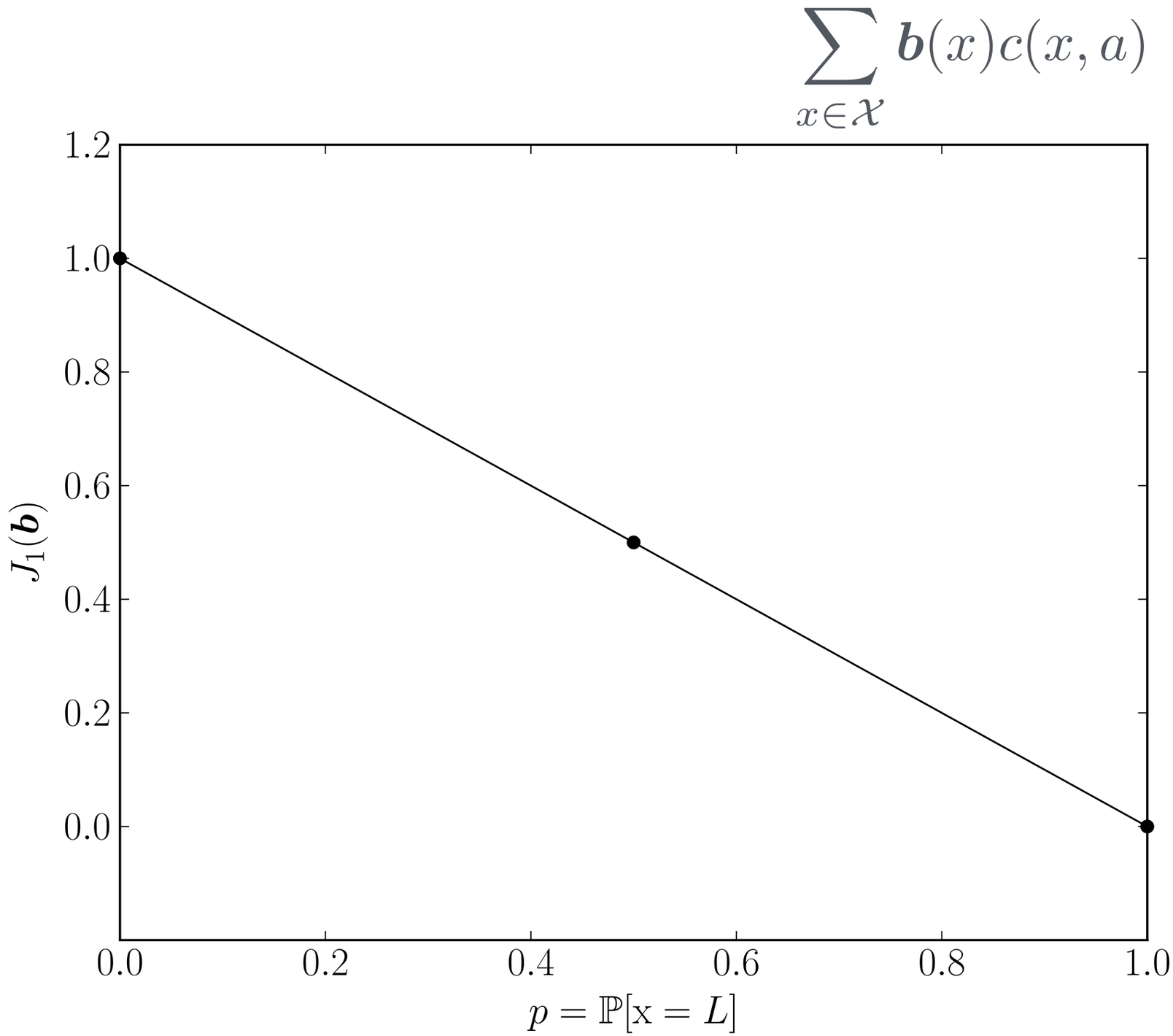
Action OR



Value iteration

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

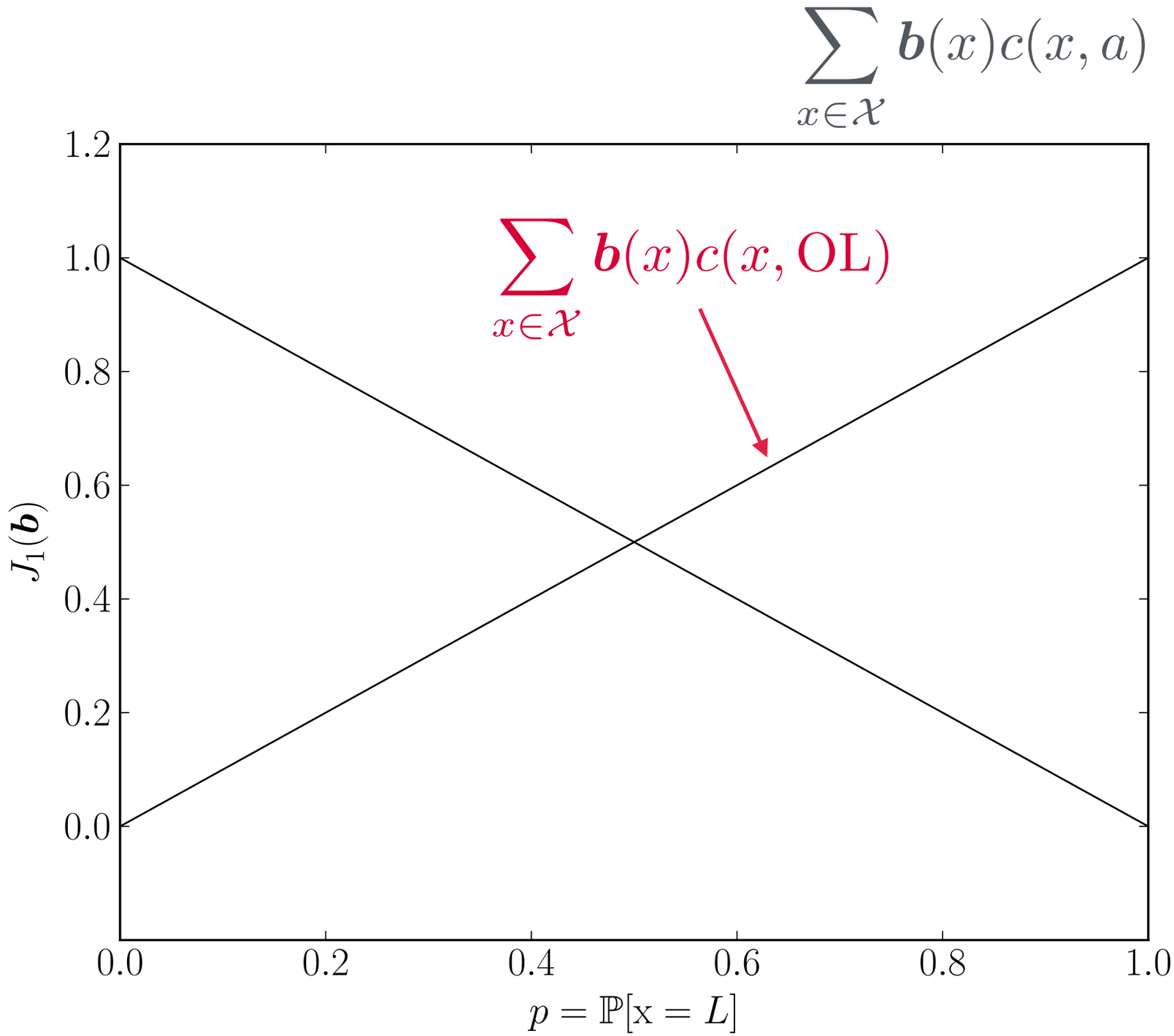
Action OL



Value iteration

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

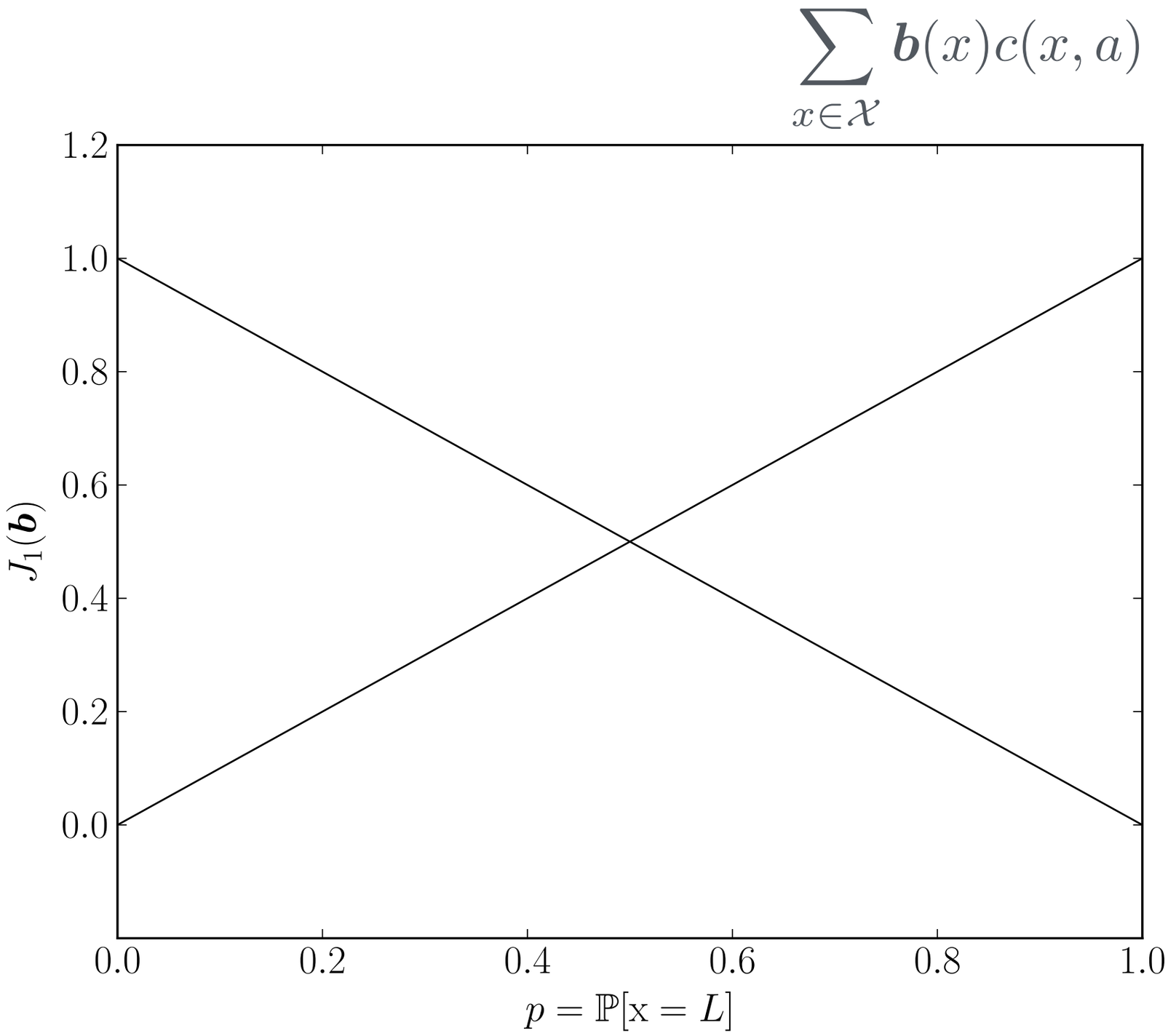
Action OL



Value iteration

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

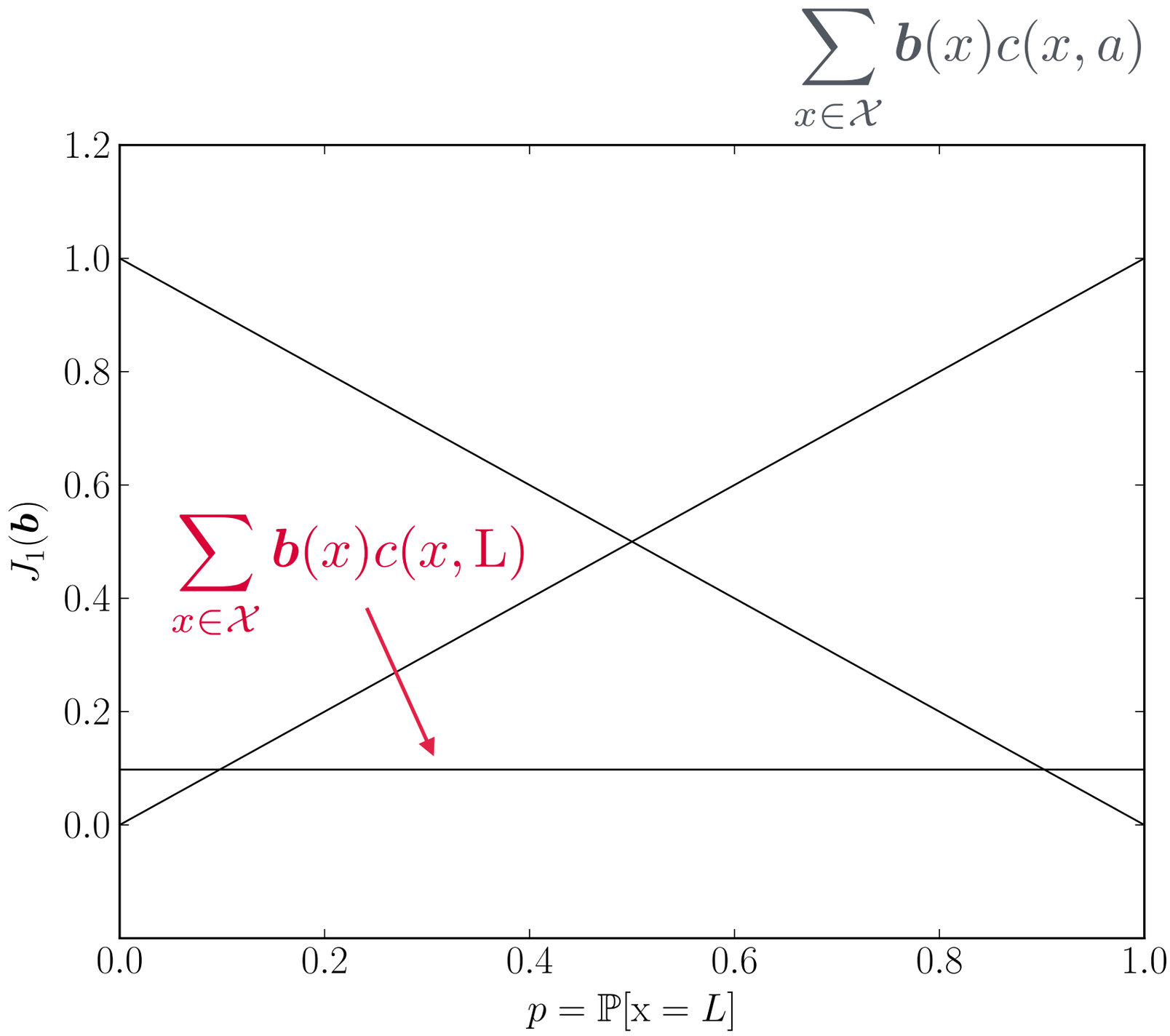
Action L



Value iteration

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

Action L



Value iteration

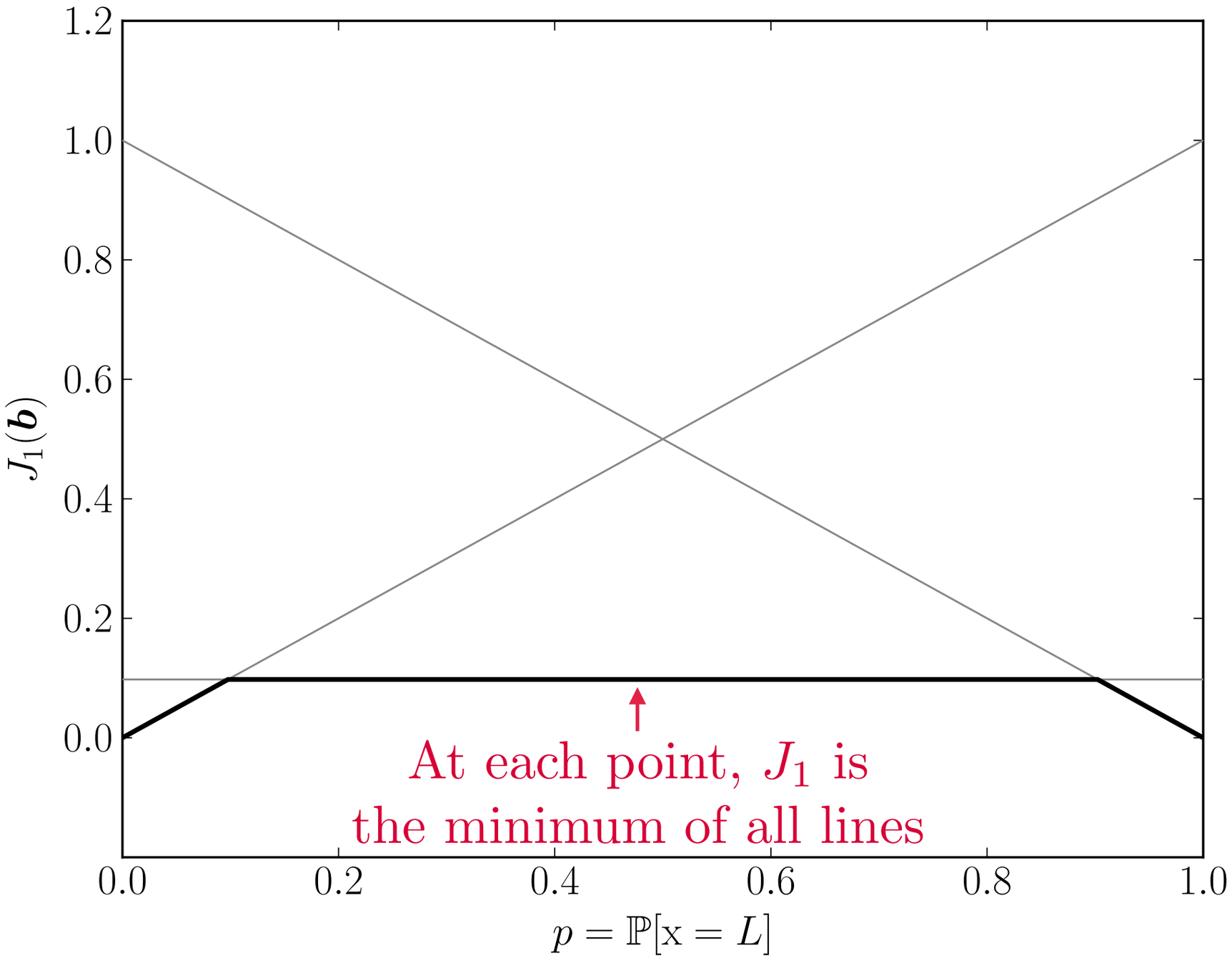
- We just computed, for every belief $\boldsymbol{b}$ and action a,

$$\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

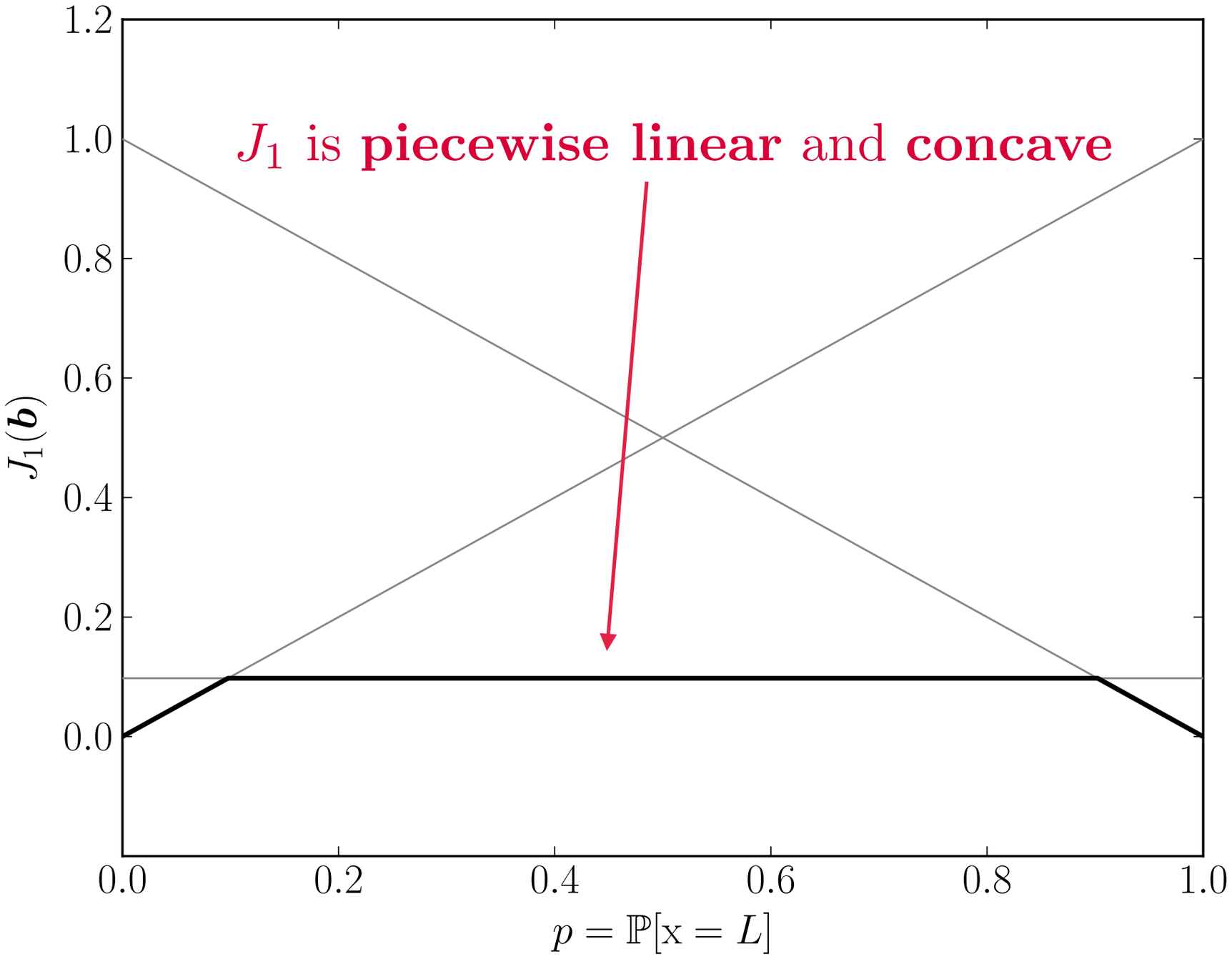
- At iteration 1, we want to compute:

$$J_1(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

Value iteration



Value iteration



Representing $J^{(k)}$

- The cost-to-go at each iteration of VI is always PWLC
  - Can always be written in the form

$$J_k(\boldsymbol{b}) = \min_{\boldsymbol{\alpha} \in \Gamma} \boldsymbol{b} \cdot \boldsymbol{\alpha}$$

Some set of vectors

Representing $J^{(k)}$

- The cost-to-go at each iteration of VI is always PWLC
  - Can always be written in the form

$$J_k(\boldsymbol{b}) = \min_{\boldsymbol{\alpha} \in \Gamma} \boldsymbol{b} \cdot \boldsymbol{\alpha}$$

$$= \min_{\boldsymbol{\alpha} \in \Gamma} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x)\alpha(x)$$

We refer to these as the α-vectors

How do we compute the α-vectors?

Value iteration

- How do we compute the α-vectors?

$$J_{k+1}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') J_k(\mathbf{B}(\boldsymbol{b}, z, a)) \right]$$

Replace PWLC representation

Value iteration

- How do we compute the α-vectors?

$$J_{k+1}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{x' \in \mathcal{X}} \sum_{z \in \mathcal{Z}} \mathbf{P}_a(x' \mid x) \mathbf{O}_a(z \mid x') \min_{\boldsymbol{\alpha} \in \Gamma_k} \mathbf{B}(\boldsymbol{b}, z, a) \cdot \boldsymbol{\alpha} \right]$$

Minimizing α-vector depends on $\boldsymbol{b}$

↓

We compute all alternatives

Value iteration

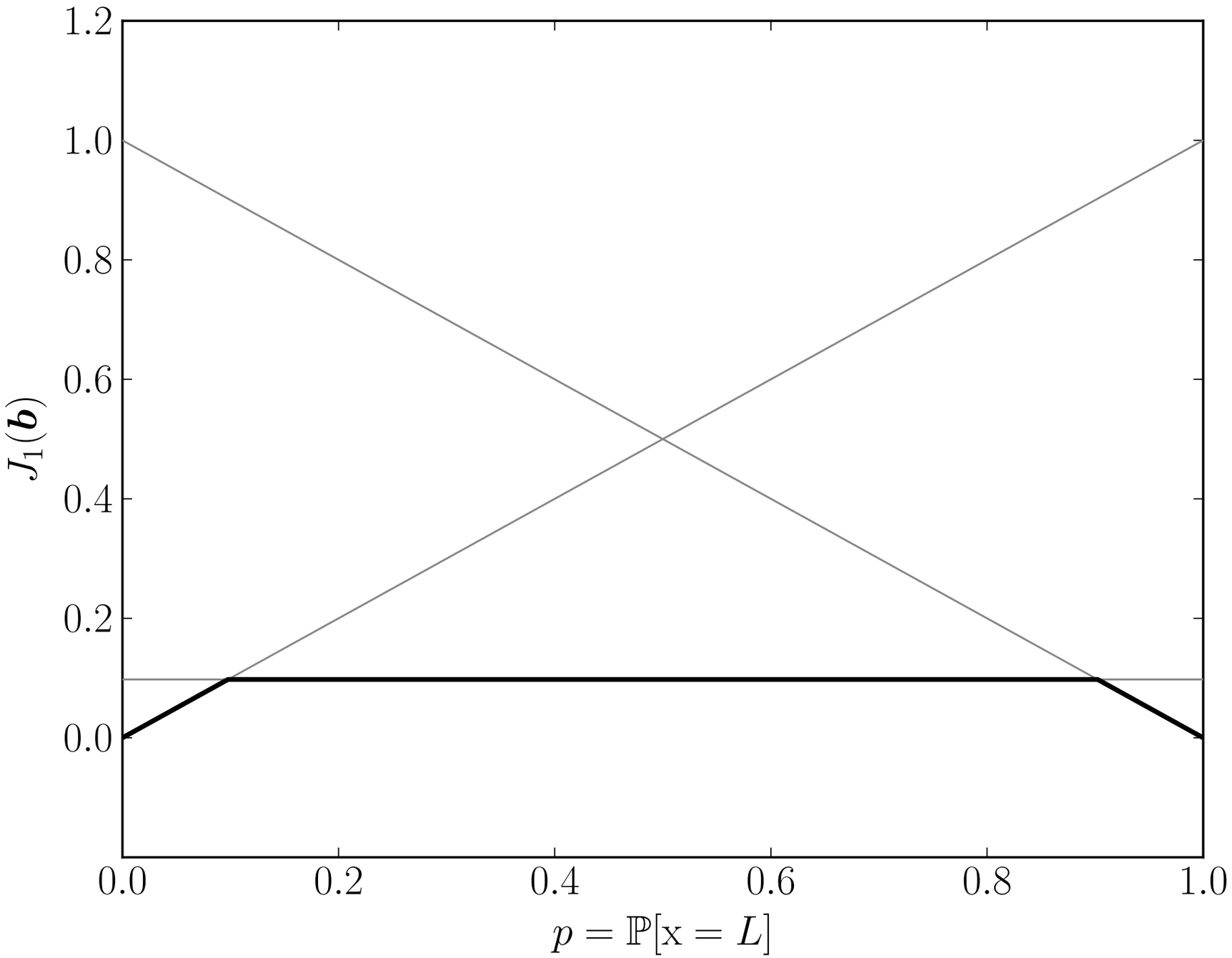
- Compute for each iteration $k+1$, the set Γ_{k+1} from Γ_k
  - For each $\boldsymbol{\alpha} \in \Gamma_k$, each $a \in \mathcal{A}$ and each $z \in \mathcal{Z}$, compute:

$$\boldsymbol{\alpha}_{a,z}^{\text{new}} = \frac{1}{|\mathcal{Z}|}\mathbf{c}_a + \gamma \mathbf{P}_a \text{diag}(\mathbf{O}_a(z \mid \cdot))\boldsymbol{\alpha}$$

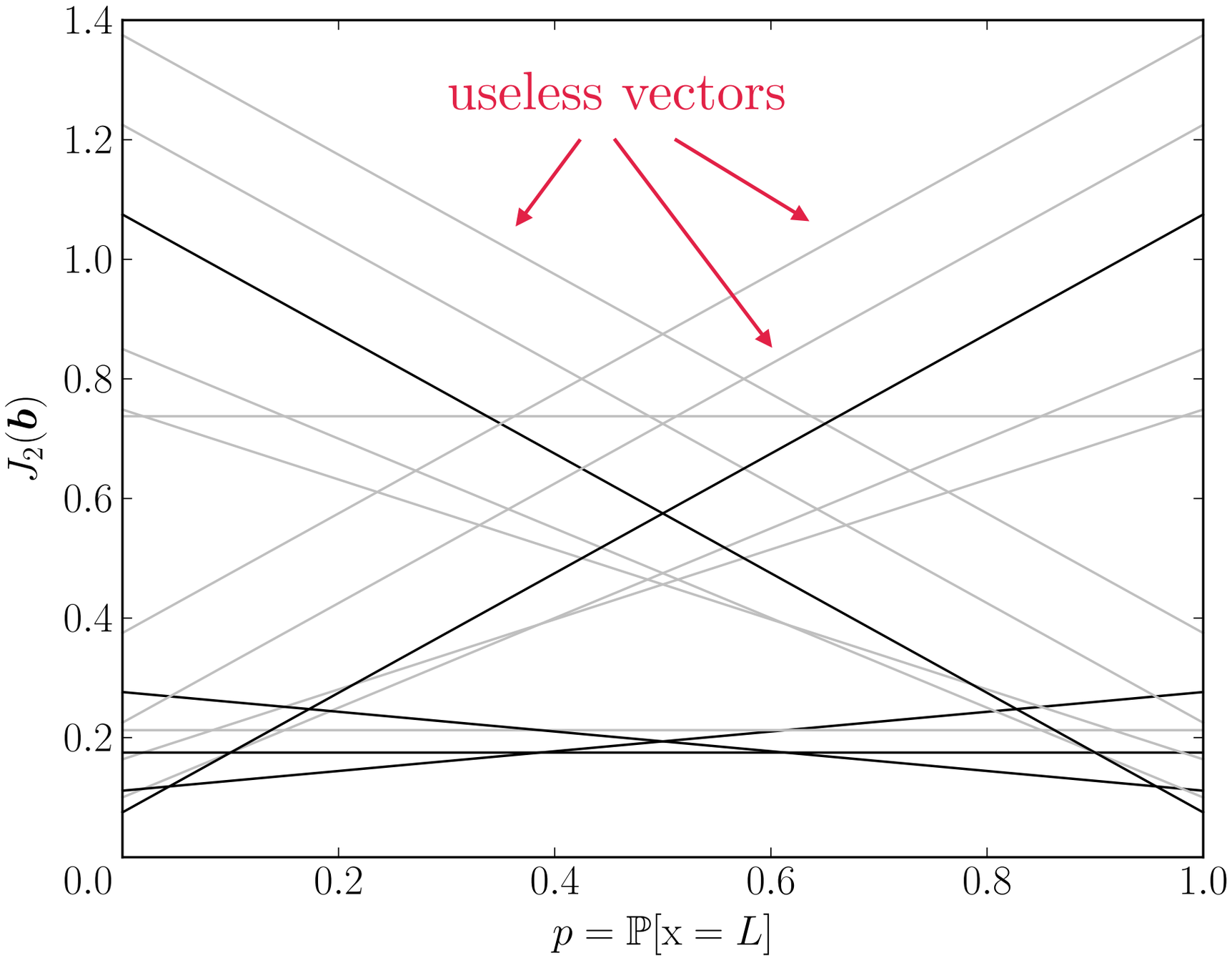
  - Compute all possible combinations of $|\mathcal{Z}|$ vectors $\boldsymbol{\alpha}_{a,z}^{\text{new}}$, one for each $z \in \mathcal{Z}$
  - For each such combination, let

$$\boldsymbol{\alpha}_{a}^{\text{new}} = \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}_{a,z}^{\text{new}}$$

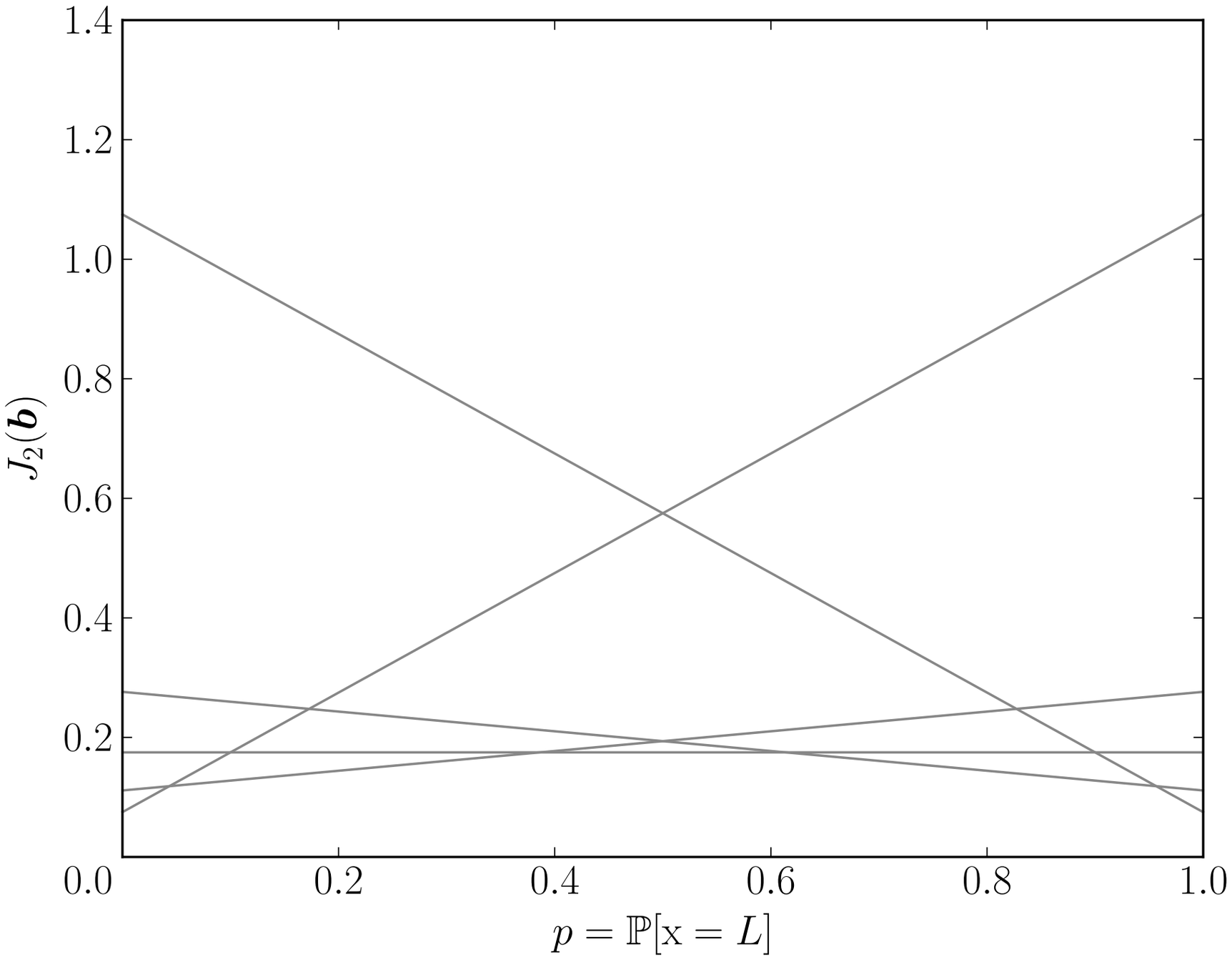
Value iteration



Value iteration



Value iteration



Value iteration

- Approaches to build Γ_{k+1} from Γ_k:
  - **Region based methods:**

    Start with empty Γ_{k+1} and only add vectors that are necessary

    (e.g.: Witness algorithm)

A vector is necessary if it represents J in a non-empty region (witness region)

Value iteration

- Approaches to build Γ_{k+1} from Γ_k:

  - **Region-based methods:**

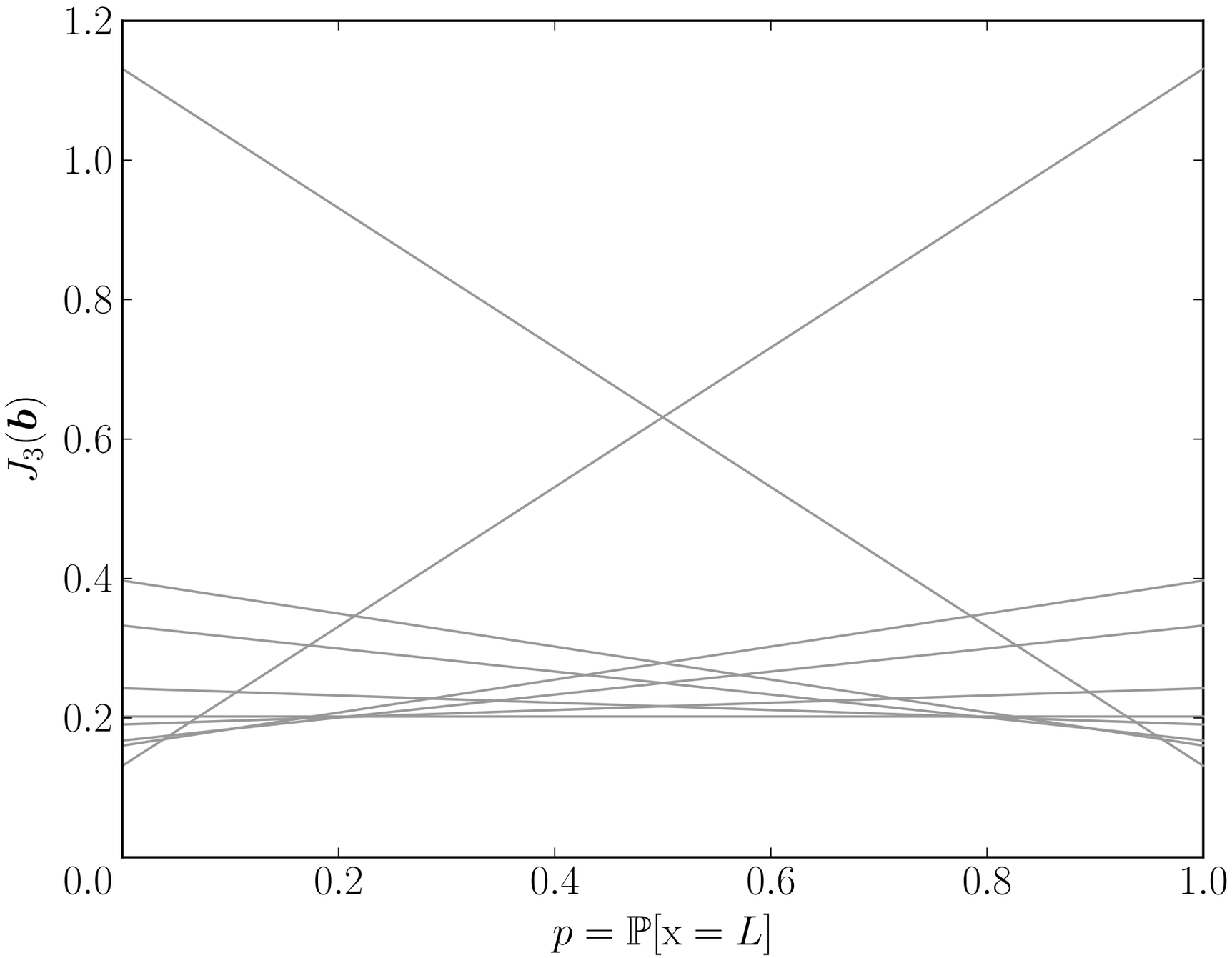
    Start with empty Γ_{k+1} and only add vectors that are necessary (e.g.: Witness algorithm)

  - **Prunning-based methods:**

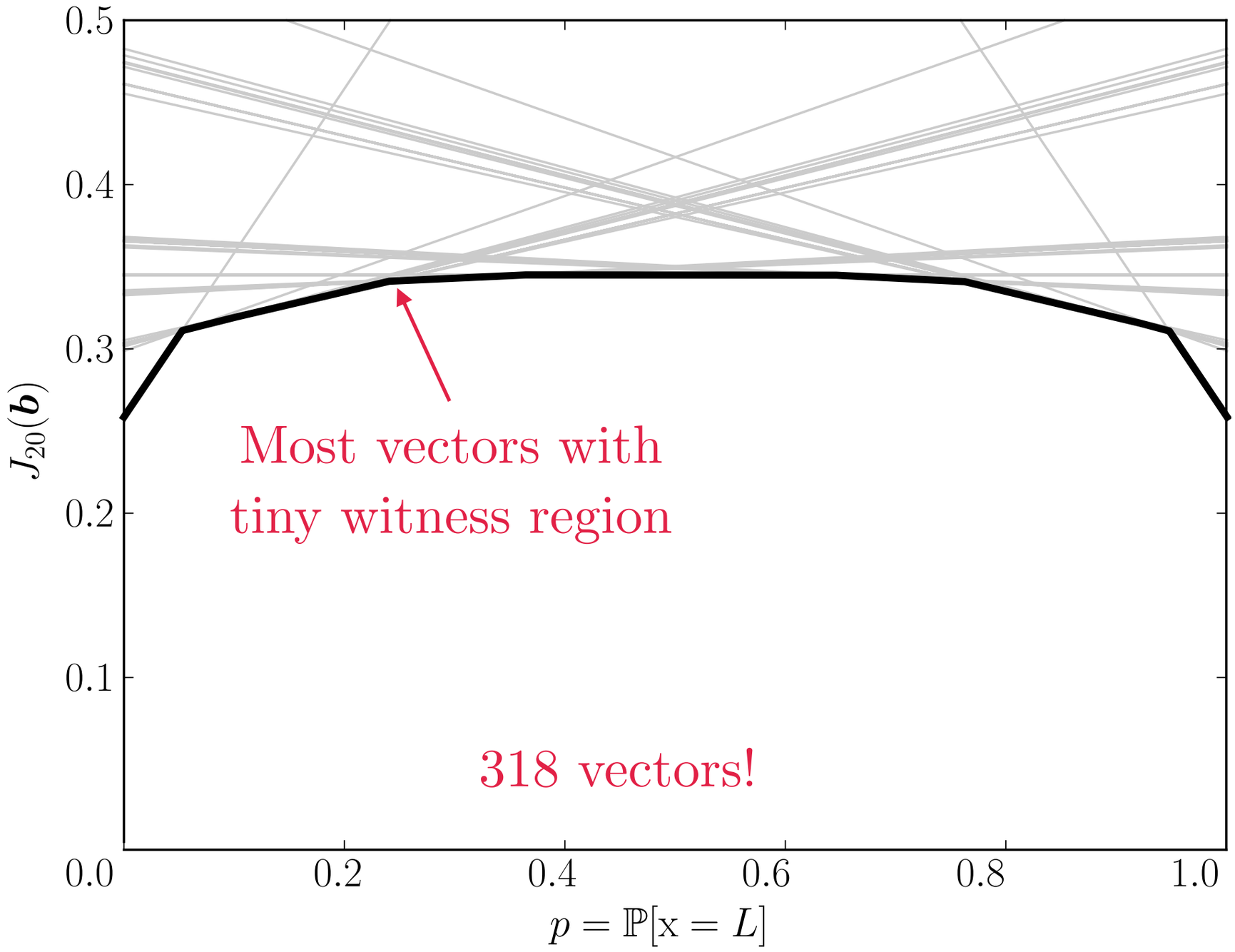
    Start with full Γ_{k+1} and remove vectors that are unnecessary

    (e.g., Incremental pruning algorithm)

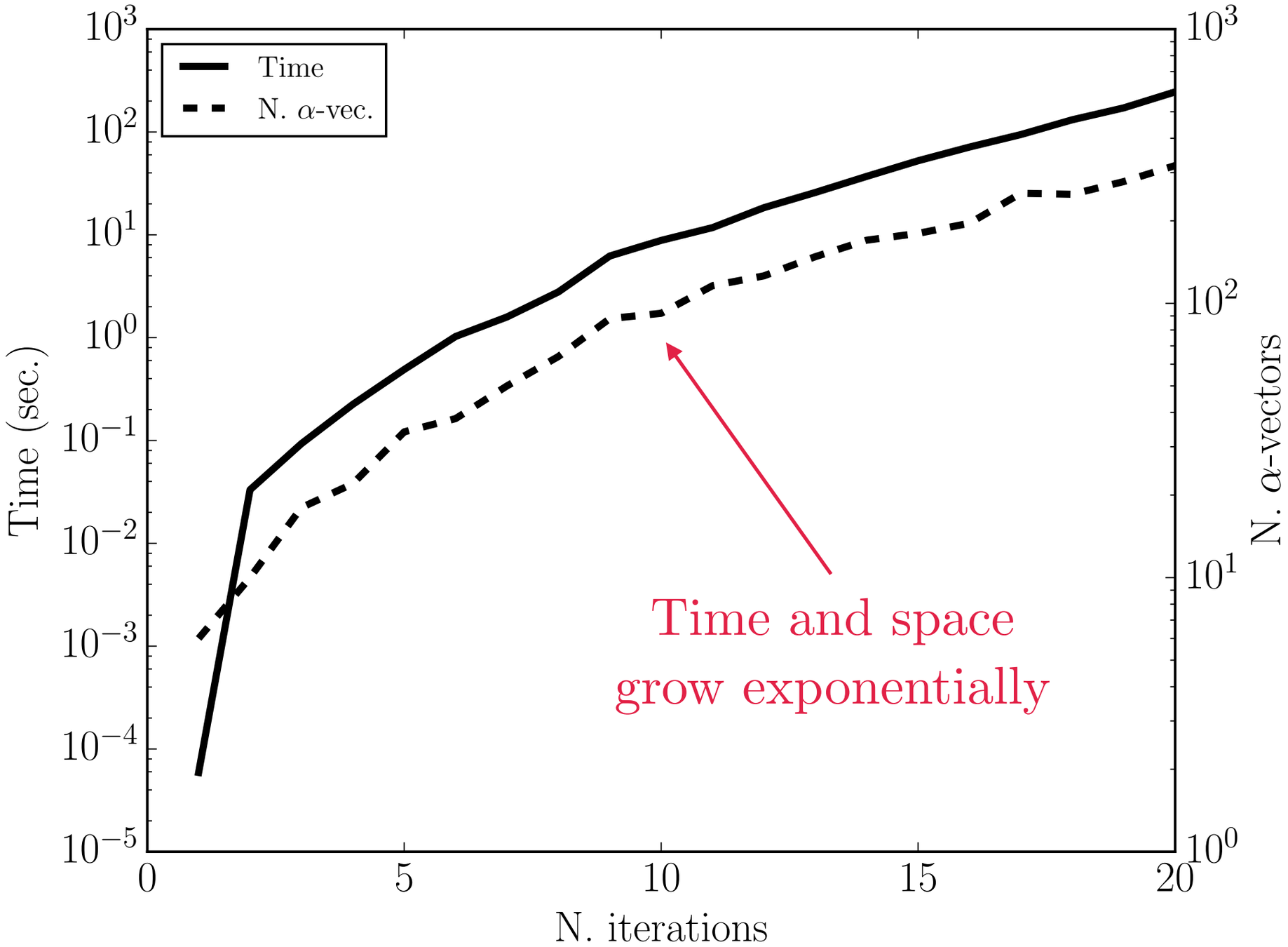
Value iteration



Value iteration

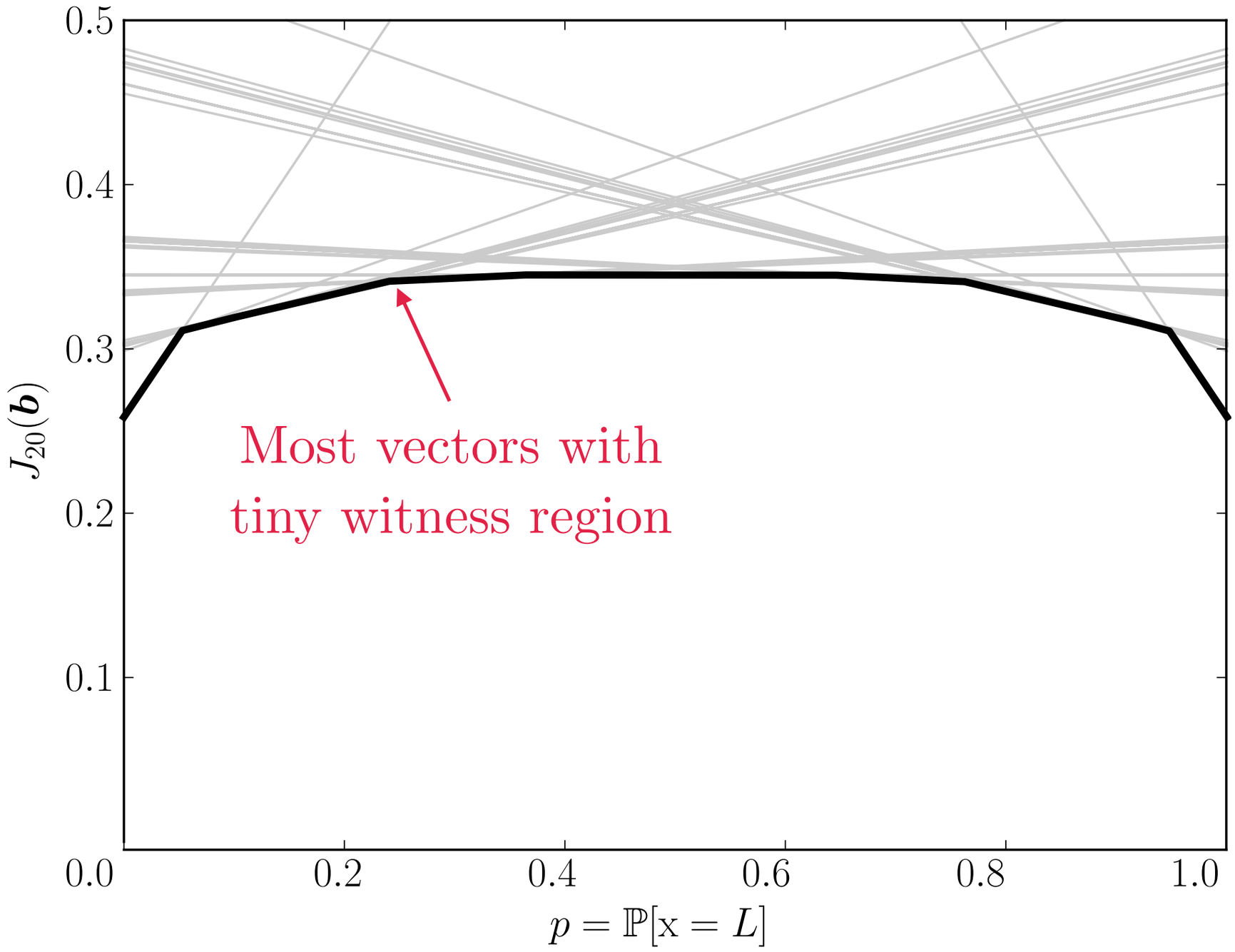


Computation time



Point-based methods

Value iteration

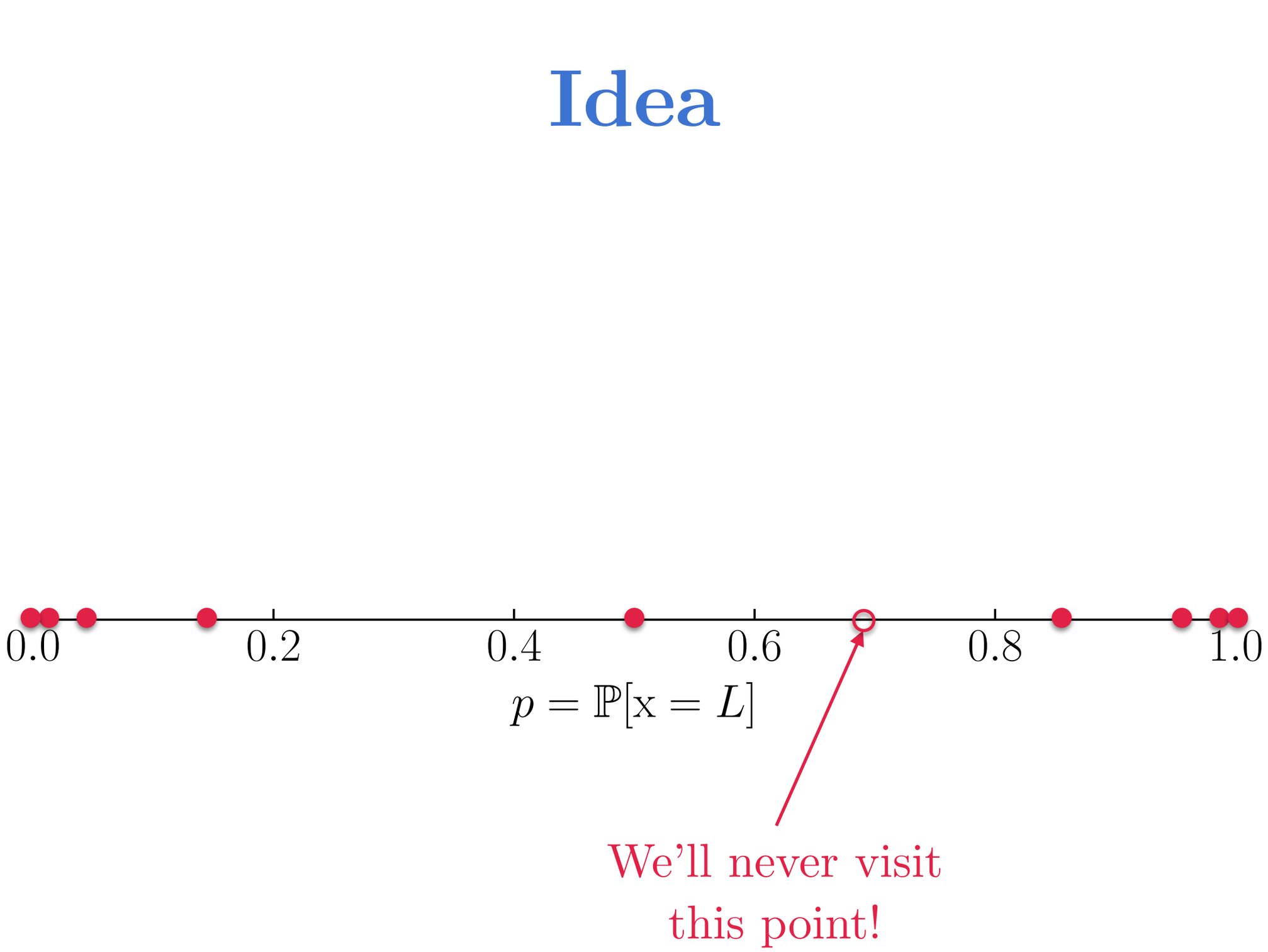


Value iteration

- Most α-vectors play little role in representing J
- What if we only compute the vectors that "truly matter"?

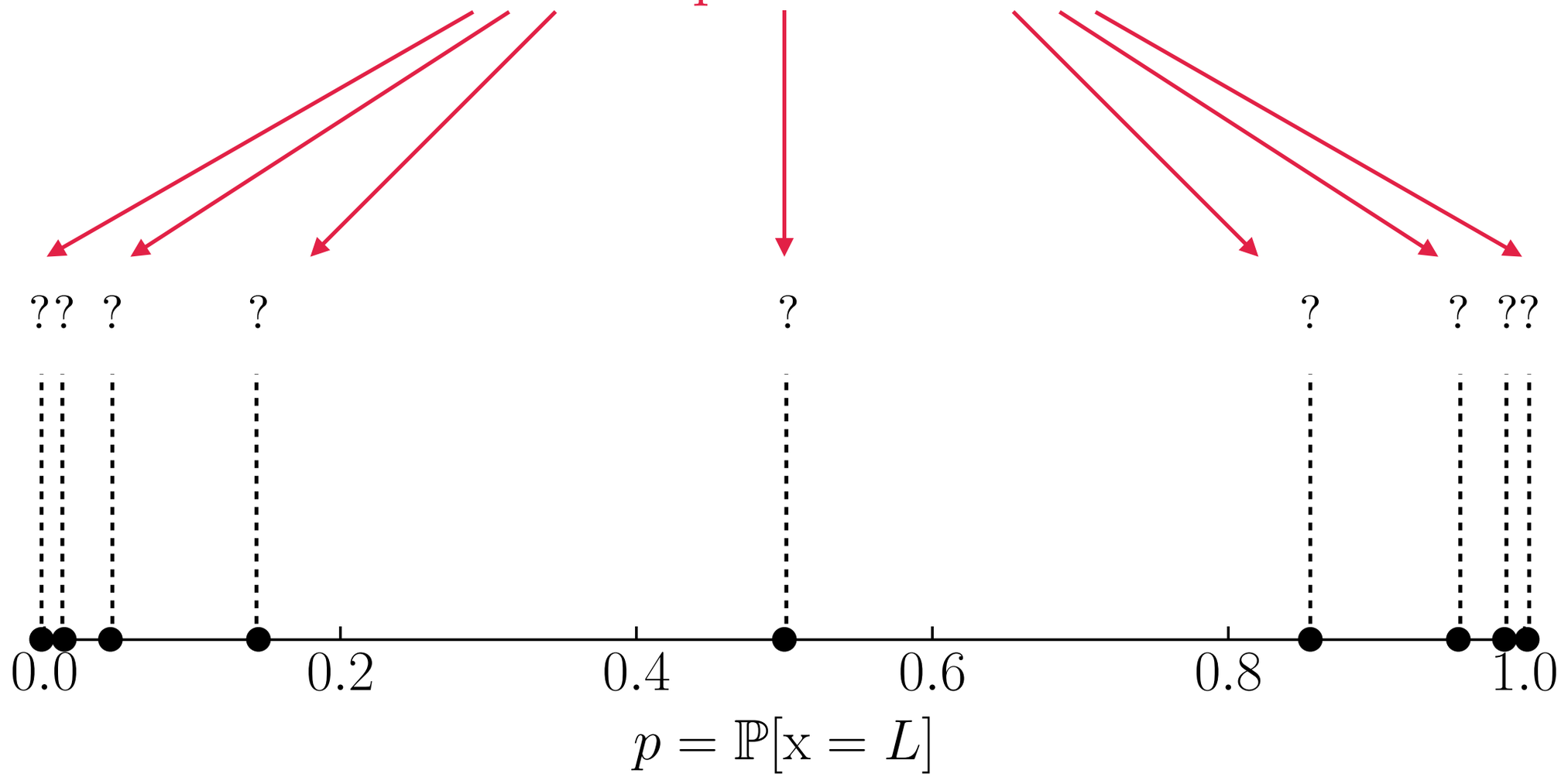
… which ones?

Idea


0.0
0.2
0.4
0.6
0.8
1.0
$p = \mathbb{P}[\mathrm{x} = L]$
We'll never visit
this point!

Idea

What if we only worry about the α-vectors in these points?



Point-based methods

- Select a finite set $\mathcal{B}_{\text{sample}}$ of beliefs to perform updates
- For each belief, compute the corresponding α-vectors

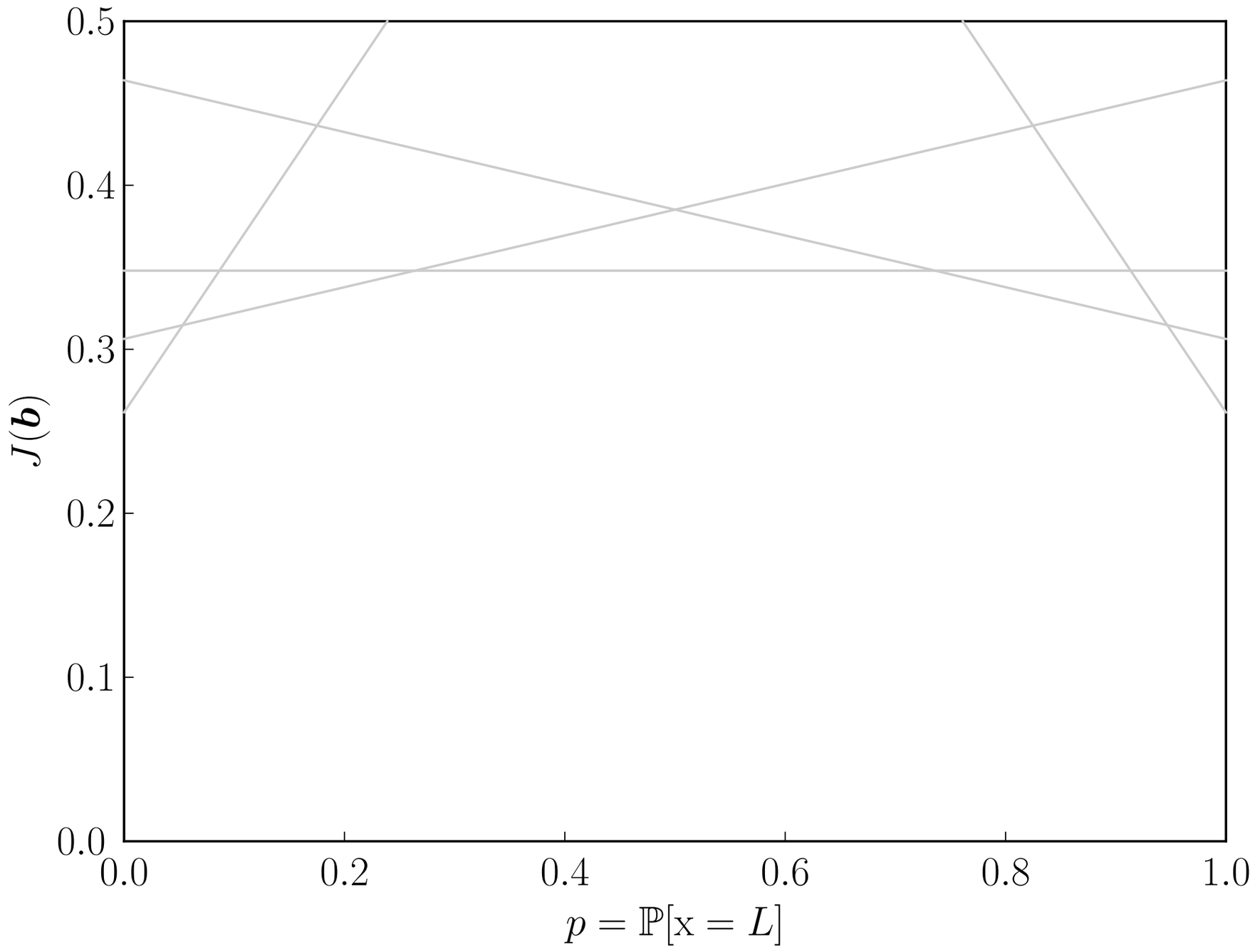
$$\boldsymbol{\alpha}^{\text{new}}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\mathbf{c}_a + \gamma \sum_{z \in \mathcal{Z}} \mathbf{P}_a \text{diag}(\mathbf{O}_a(z \mid \cdot)) \min_{\boldsymbol{\alpha} \in \Gamma} \mathbf{B}(\boldsymbol{b}, z, a) \cdot \boldsymbol{\alpha} \right]$$

- If needed, rebuild $\mathcal{B}_{\text{sample}}$

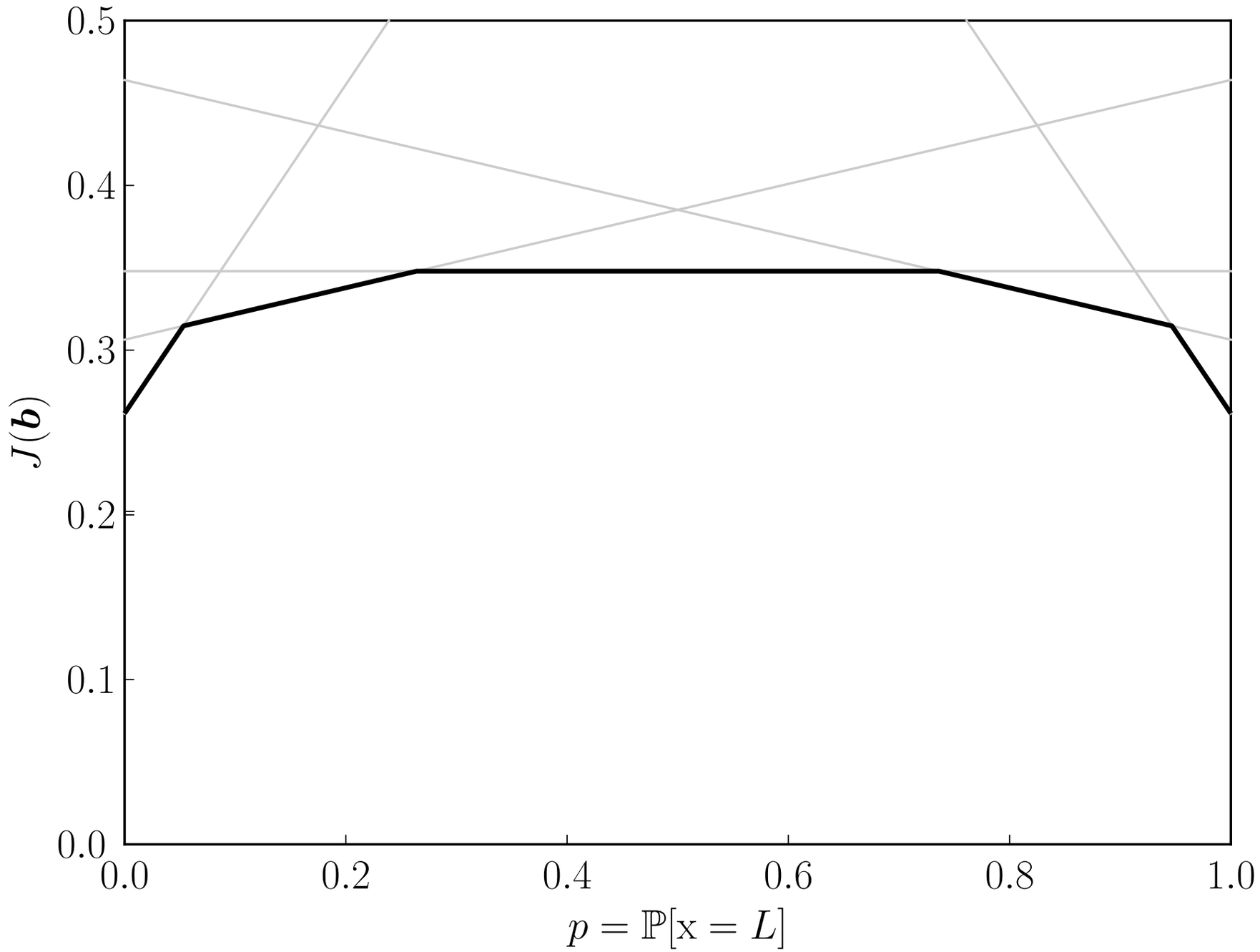
Point-based methods

- Many point-based methods:
  - PBVI (Pineau et al., 2003)
  - Perseus (Spaan & Vlassis, 2005)
  - HSVI (Smith & Simmons, 2005)
  - FSVI (Shani et al., 2007)
  - SARSOP (Kurniawati et al., 2008)
  - GapMin (Poupart et al., 2011)
- Much code available

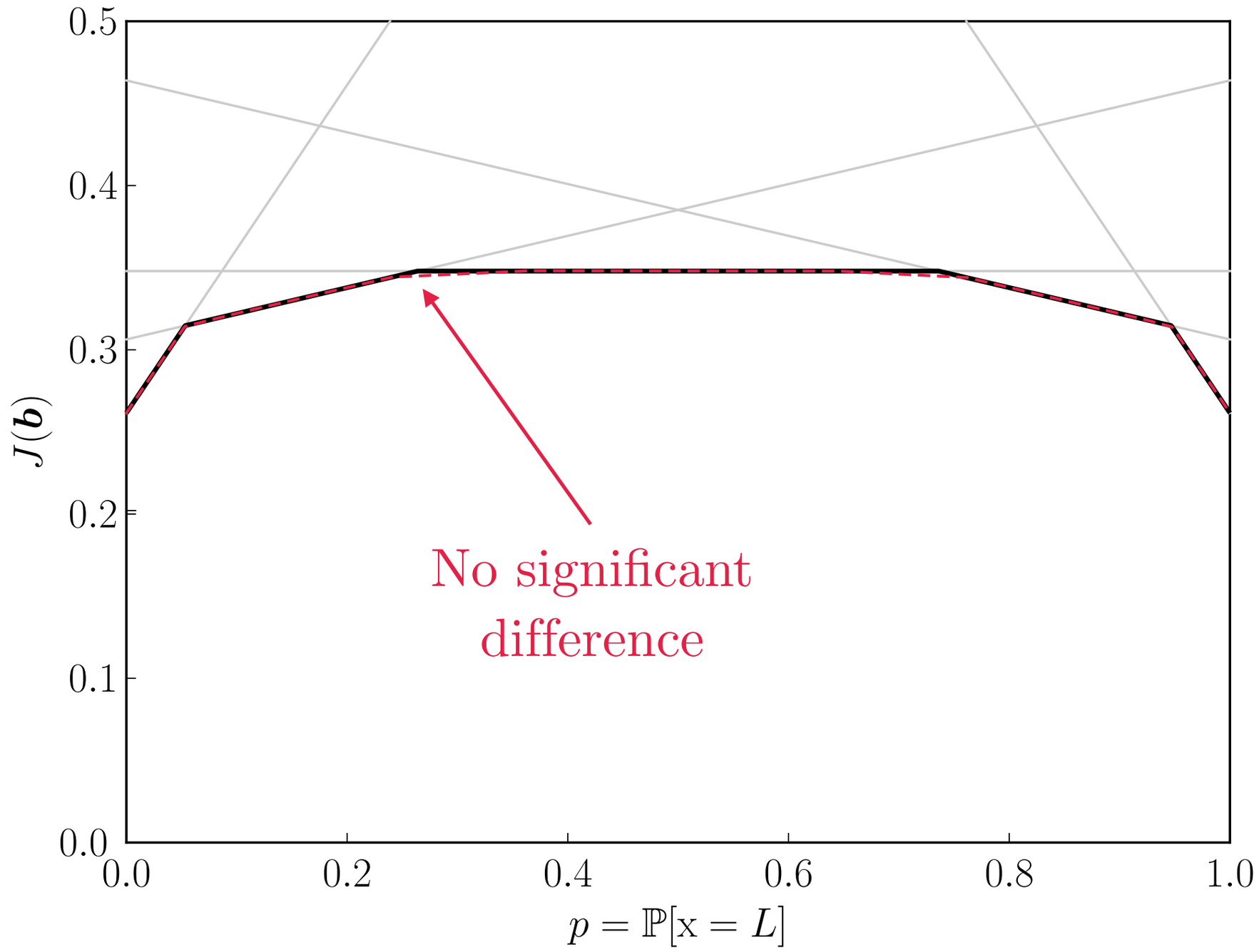
Point-based methods



Point-based methods



Point-based methods



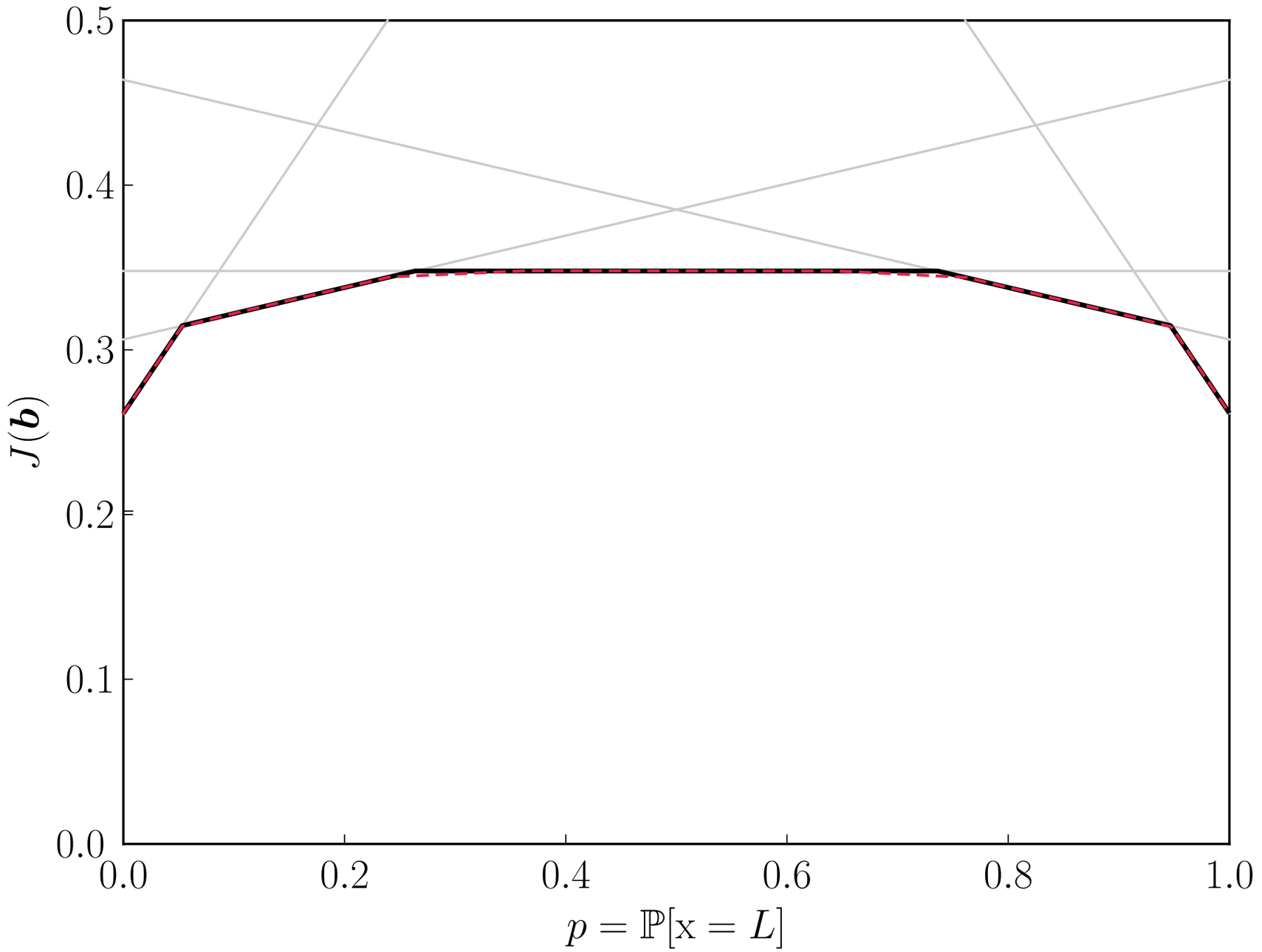
Point-based methods

VI:

- 318 vectors
- ~4 minutes

PERSEUS:

- 5 vectors
- 226 ms



What about policy iteration?

Policy iteration?

- Value iteration for POMDPs
  - How do we represent a cost-to-go function?
  - **At each iteration of VI, the cost-to-go is PWLC**
- Policy iteration for POMDPs
  - How do we represent a policy?

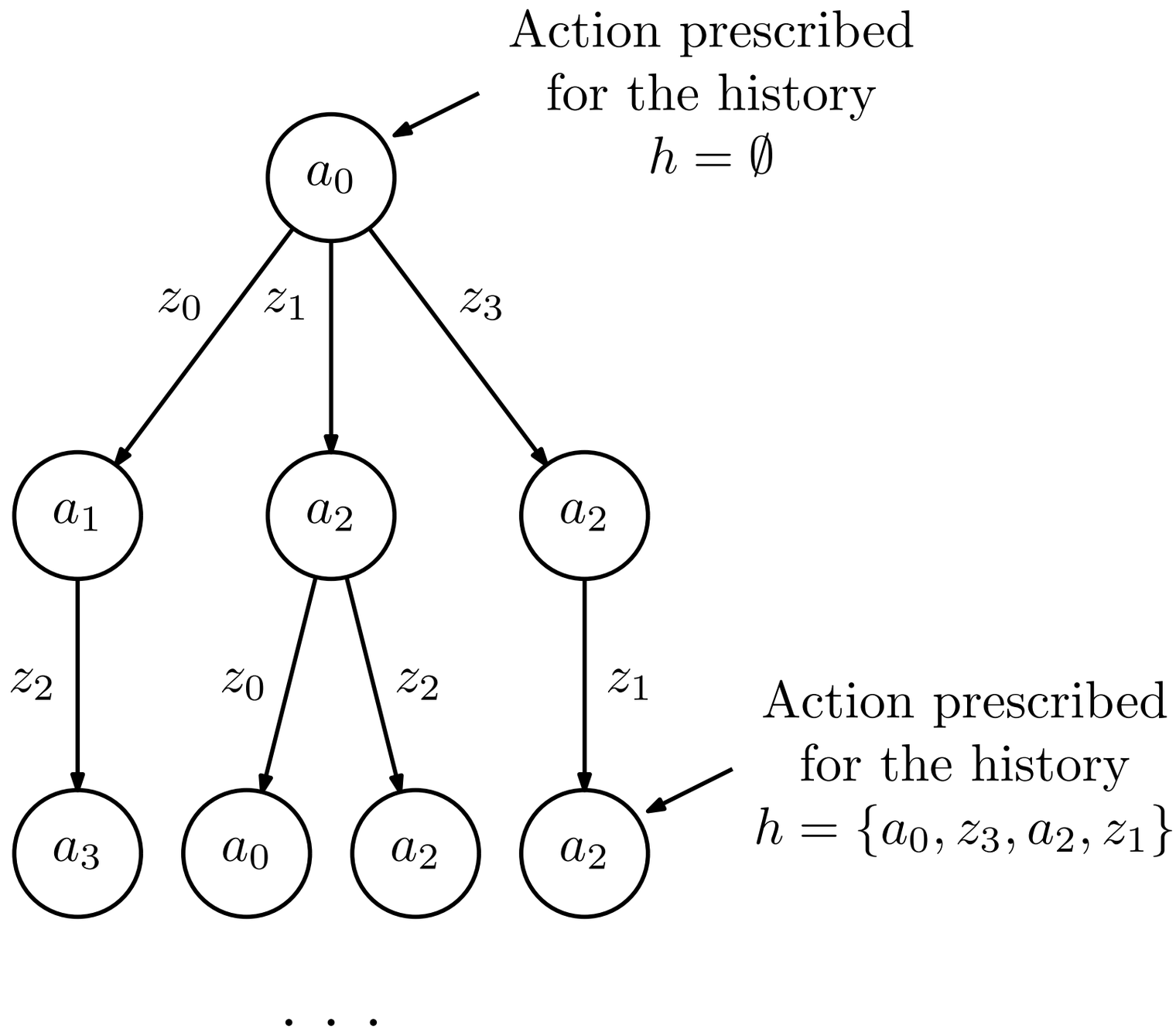
Policy graphs

- How can we represent a POMDP policy?
  - Compute it in runtime from J
  - Alternatively, we can consider policies as mapping **histories** to actions

Policy graphs

- We can represent the possible histories in a **policy tree**
  - Each node contains the **action** for that history
  - Branches correspond to **observations** from the node

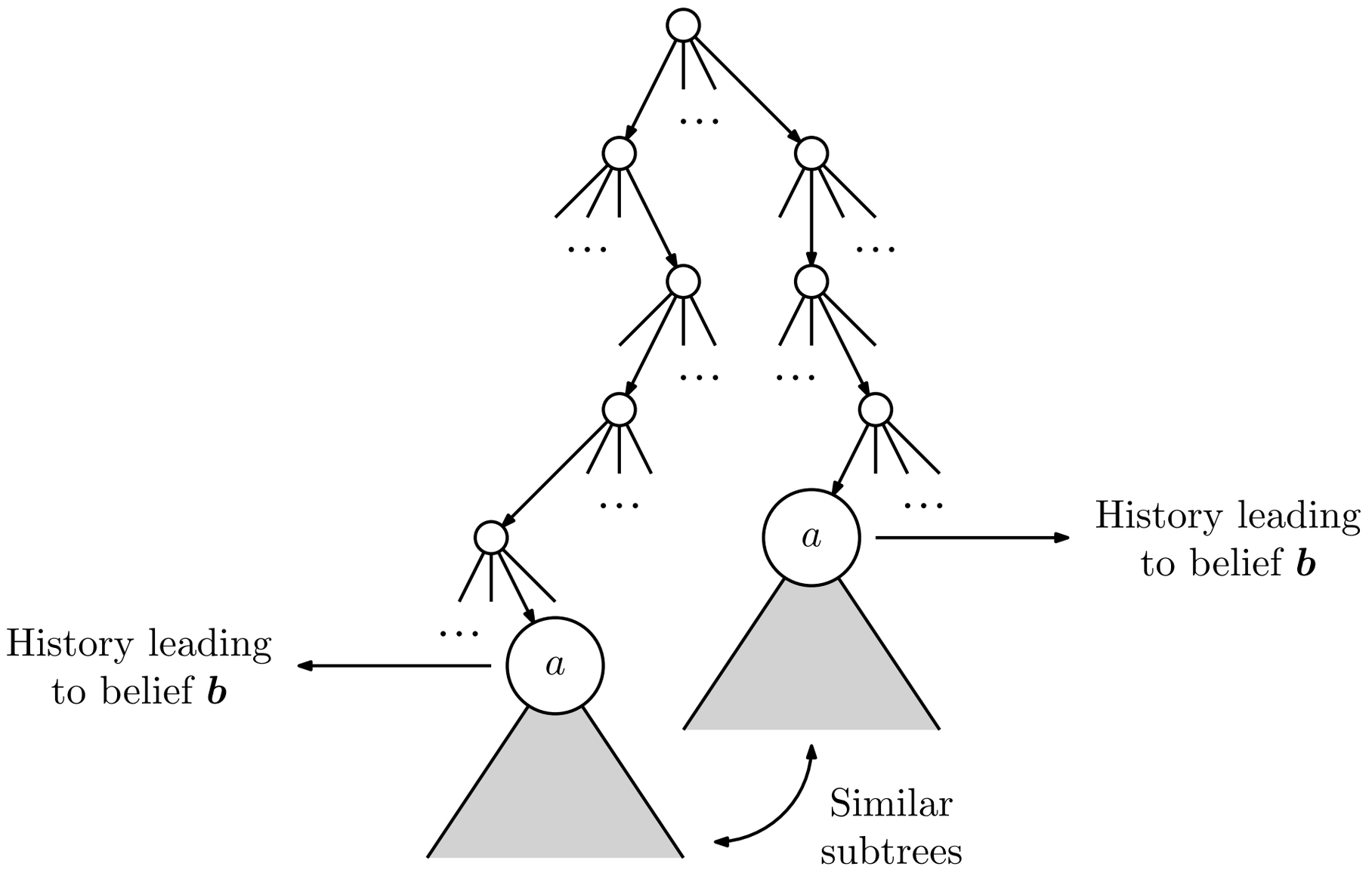
Policy graphs



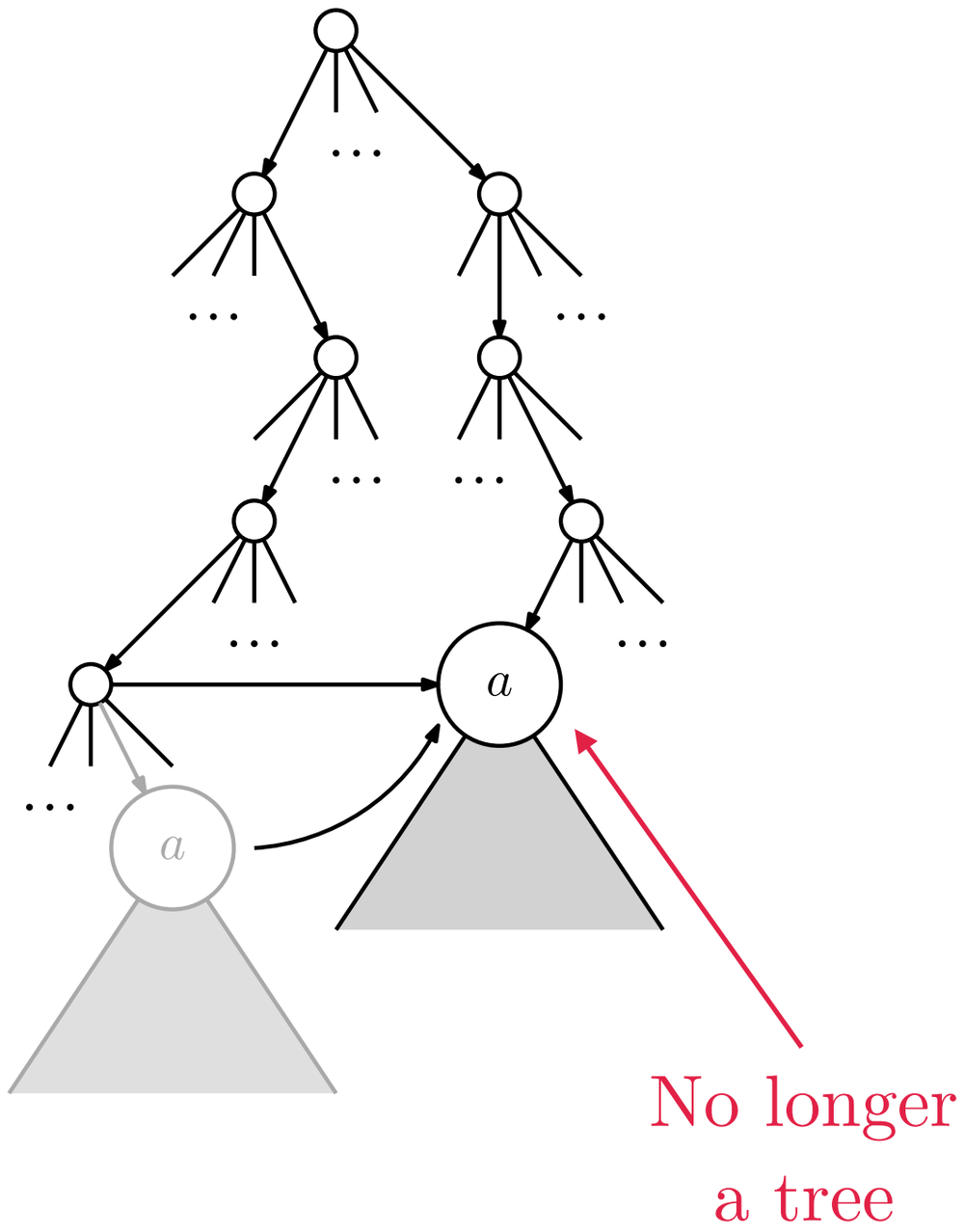
Policy graphs

- There is a lot of redundancy in policy trees
  - Histories leading to the same belief will have equivalent subtrees

Policy graphs



Policy graphs

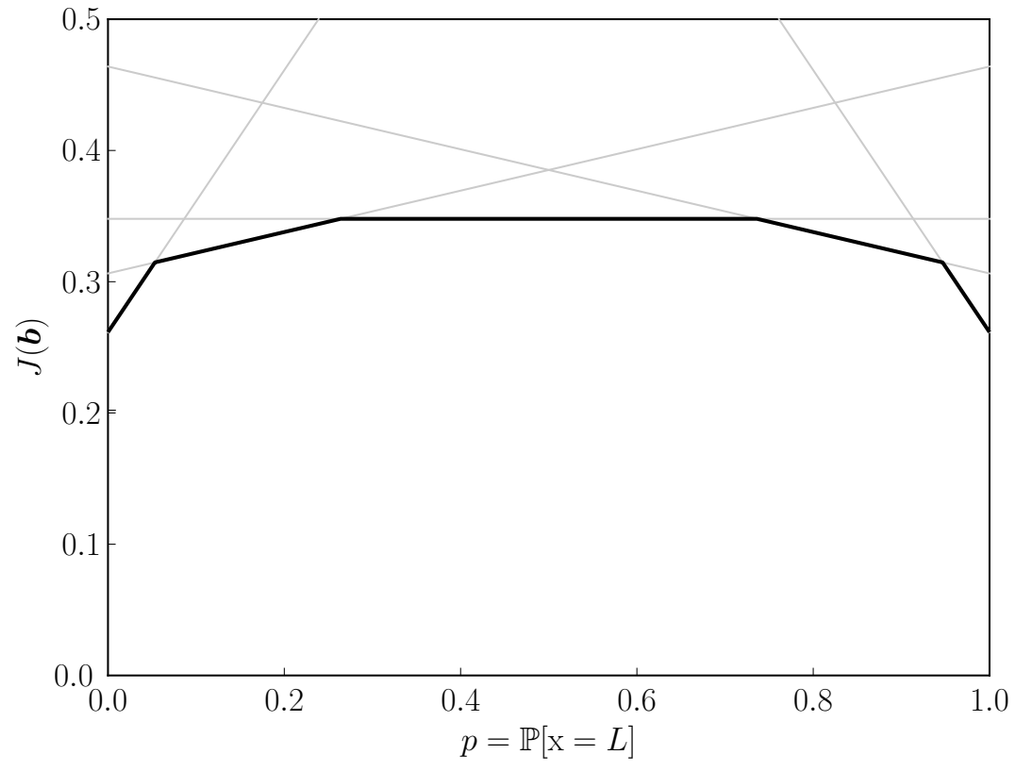


Policy graphs

- Policy graphs provide convenient representations for POMDP policies
- Close relation between policy graphs and α-vectors
  - Each node in the graph corresponds to an α-vector and vice-versa

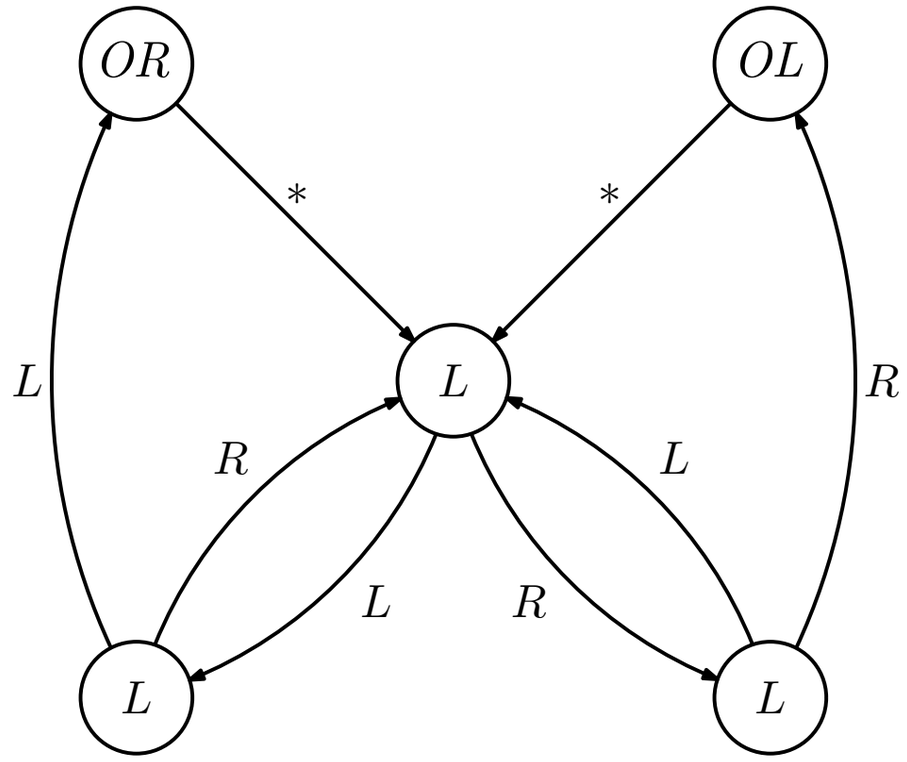
Policy graphs

- Example: Tiger problem



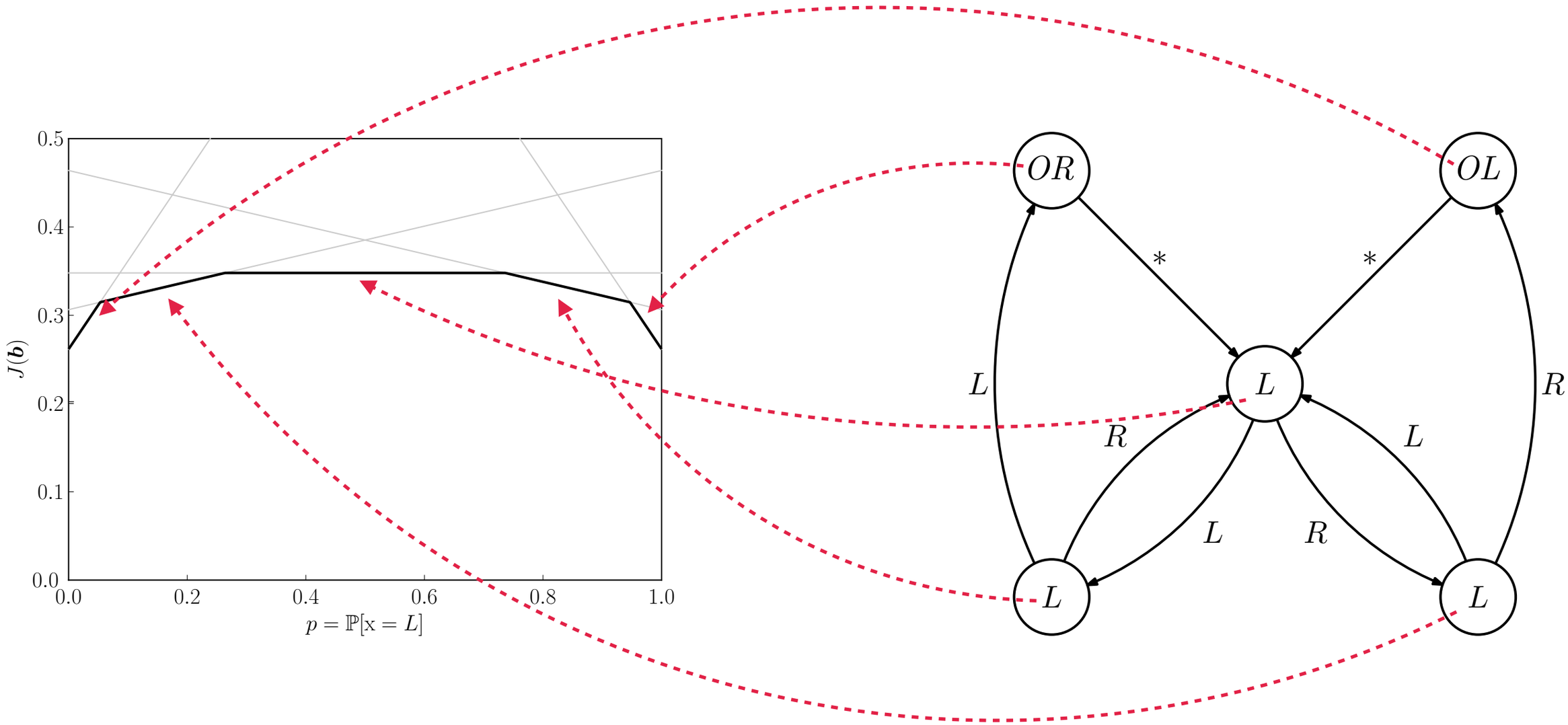
5 vectors

5 nodes



Policy graphs

- Example: Tiger problem



Key points about POMDPs

- Very general model for decision making under uncertainty
- Very hard to solve
- Beliefs provide a **summary** of the history
- POMDP $\leftrightarrow$ Belief MDP
  - We can use VI $\rightarrow$ Cost-to-go is PWLC (finite representation)
  - We can use PI $\rightarrow$ Policy graphs w/ finite number of nodes
- Approximate methods:
  - MDP heuristics
  - Point-based methods

Comments on complexity

How hard is an MDP?

- MDPs can be solved by a linear program
  - LP is known to be **polynomial-time (P)**
  - MDPs are solvable in **polynomial-time (P)**

How hard is a POMDP?

- Infinite horizon POMDPs are **undecidable**

- Finite-horizon POMDPs are **PSPACE-complete**
  - ... little hope for exact solution methods

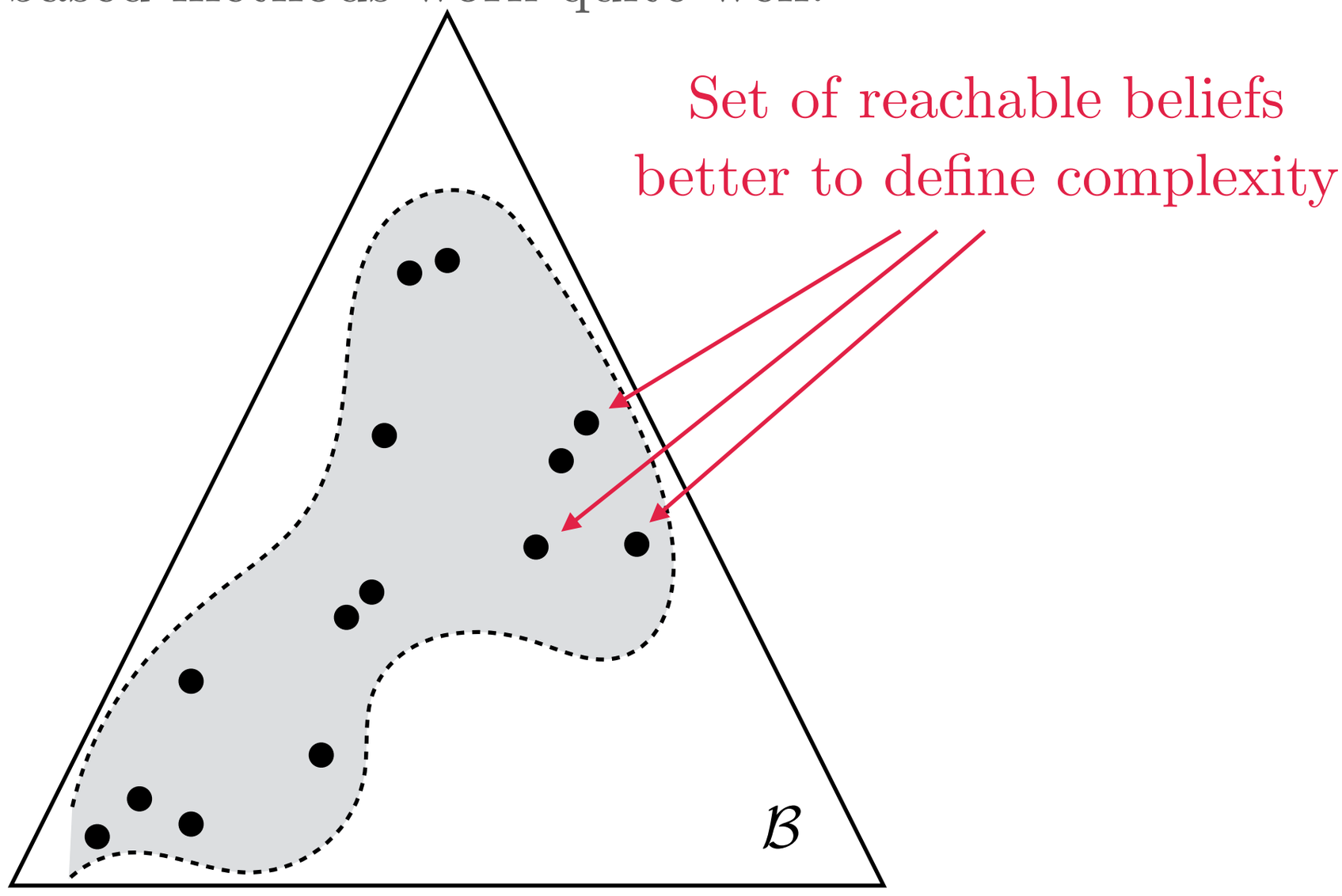
- POMDPs are **non-approximable**
  - ... in the worst case, you can't even guarantee a good approximation!

How hard is a POMDP?

- However, point-based methods work quite well!

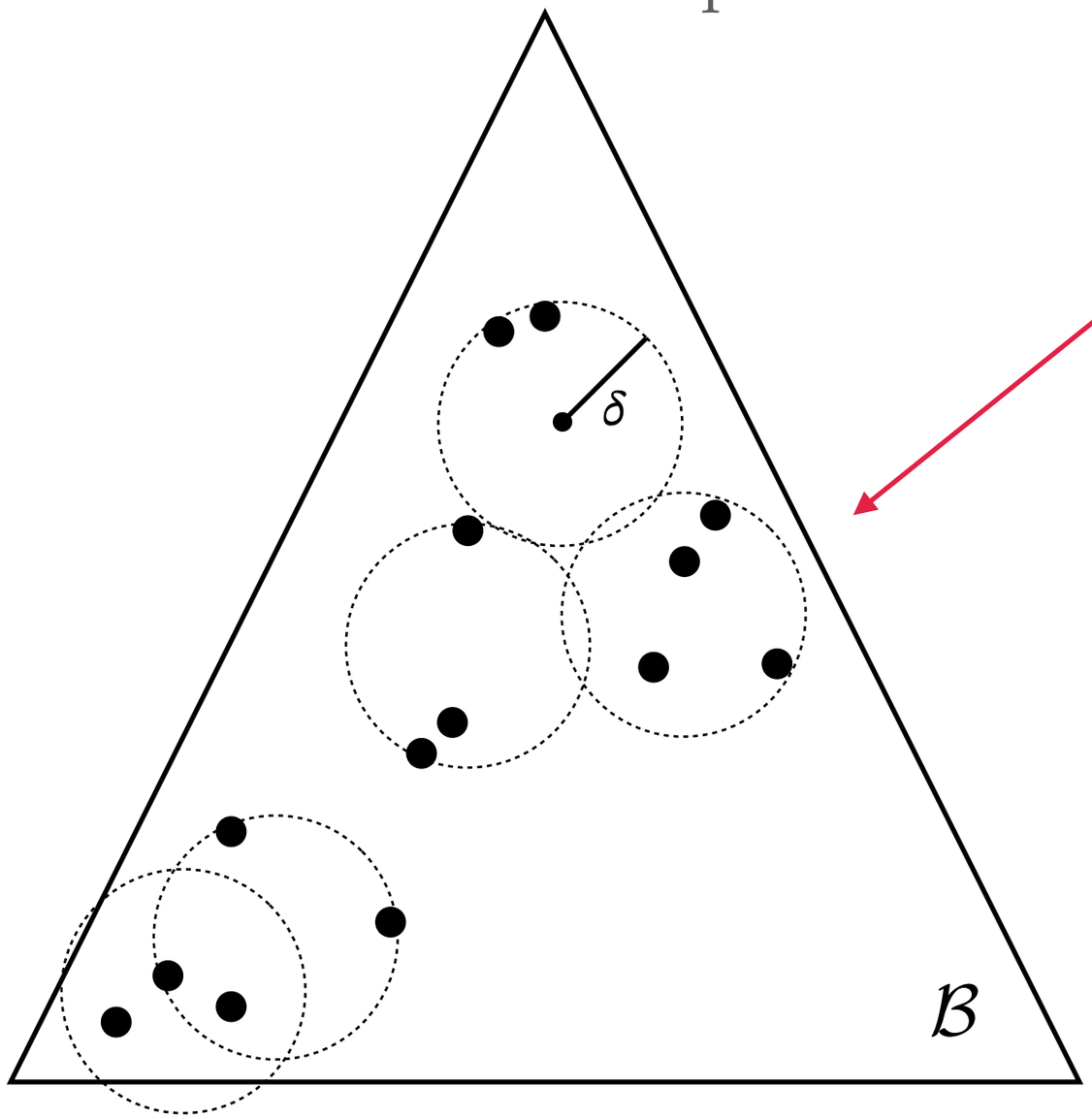
How hard is a POMDP?

- However, point-based methods work quite well!



How hard is a POMDP?

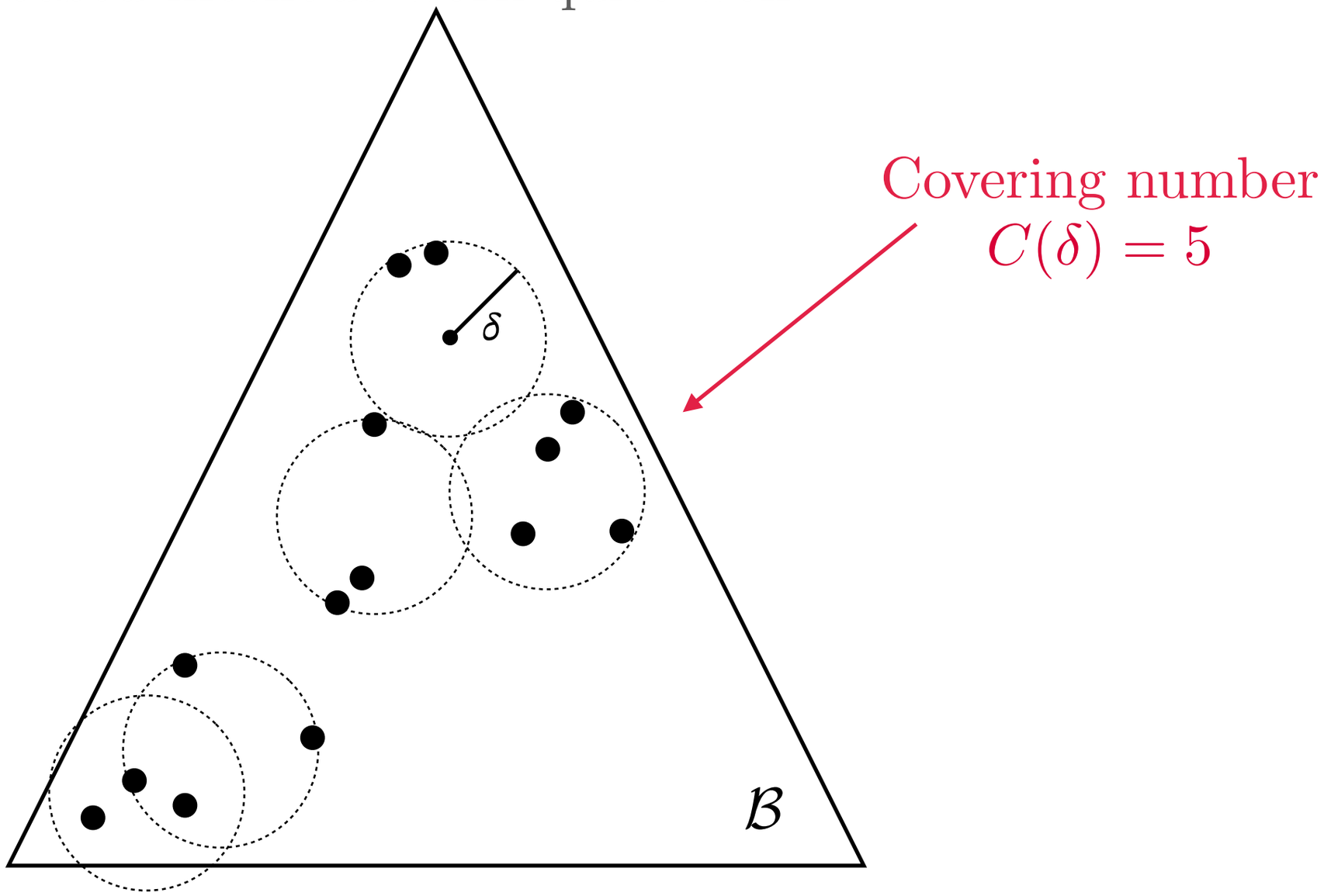
- However, point-based methods work quite well!



Need 5 balls or radius δ to "cover" reachable points

How hard is a POMDP?

- However, point-based methods work quite well!



How hard is a POMDP?

- Complexity of POMDP planning better captured by the covering number of the reachable belief space
- Some point-based methods are built on such argument (they sample beliefs to cover reachable space)
  - Ex: SARSOP (Kurniawati et al., 2008)