TÉCNICO LISBOA

Planning, Learning and Decision Making

Lecture 10. POMDPs (cont.)

Partially observable MDPs

- **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 MDPs

Key Property: Markov property

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

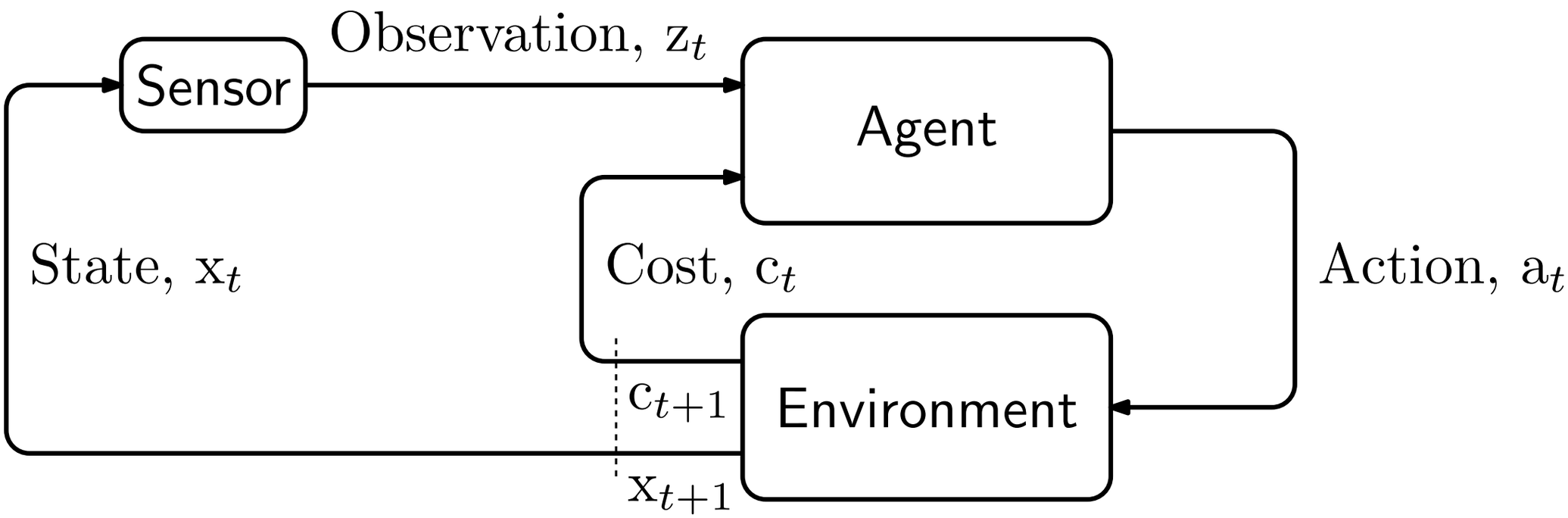
$$\mathbb{P}\left[\mathrm{x}_{t+1} = y \mid \mathbf{x}_{0:t} = \boldsymbol{x}_{0:t}, \mathbf{a}_{0:t} = \boldsymbol{a}_{0:t}\right] = \mathbb{P}\left[\mathrm{x}_{t+1} = y \mid \mathrm{x}_t = x_t, \mathrm{a}_t = a_t\right]$$

State-dependent observations

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

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

Partially observable MDPs



Challenges

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

A brief recap

- Remember the tiger problem?
  - When escaping a dungeon, you face two doors
  - Behind one of the doors lies your freedom
  - Behind the other door lies a fearsome tiger

A brief recap

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

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?

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

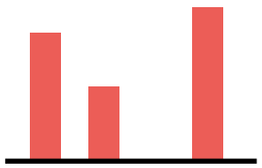
$$\mathbf{P}_{L} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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

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

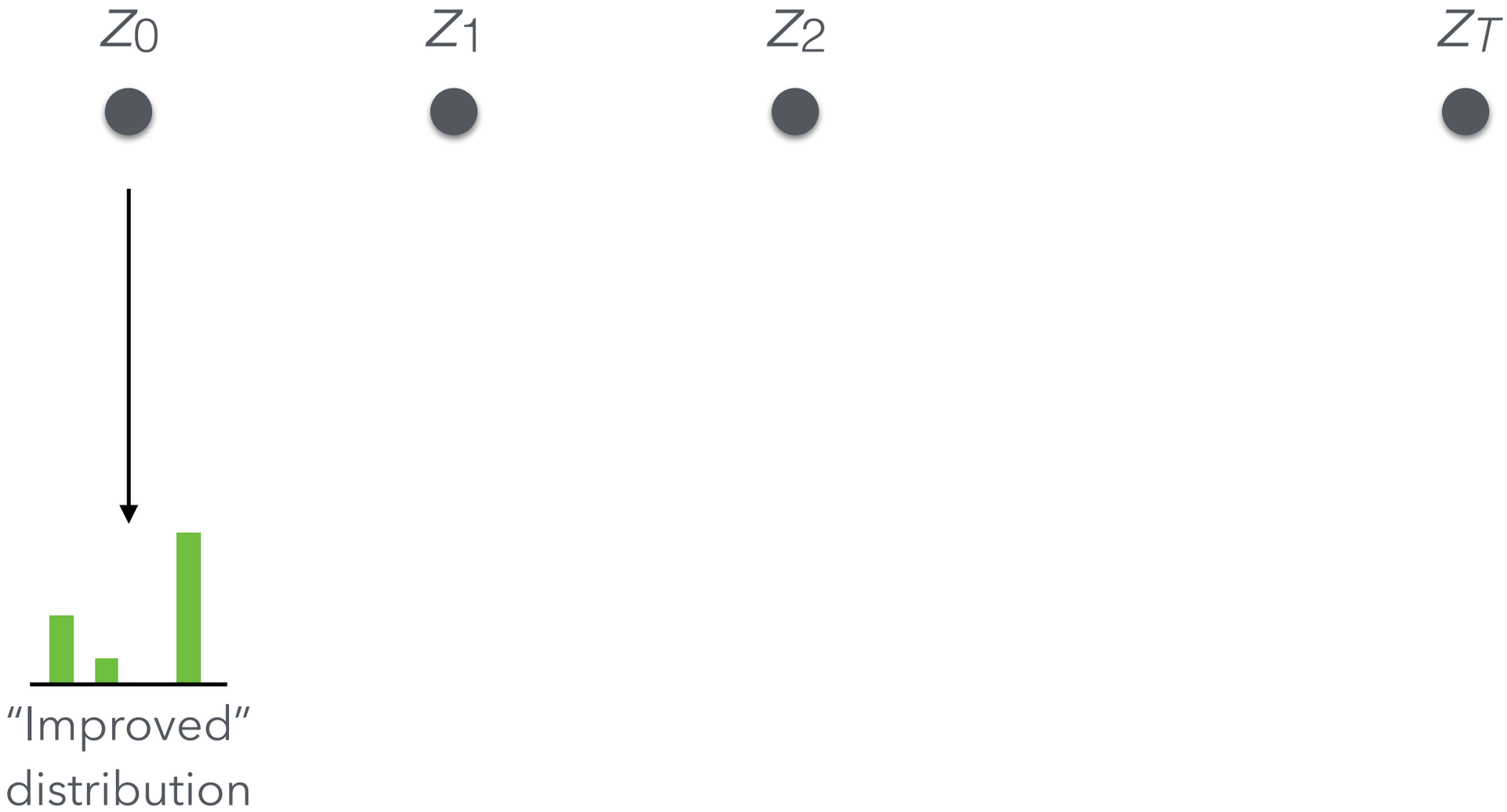
Remember?

Z_0 Z_1 Z_2 Z_T



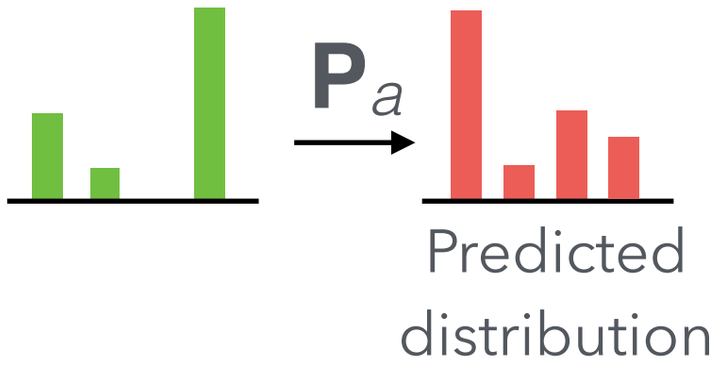
Initial distribution

Remember?

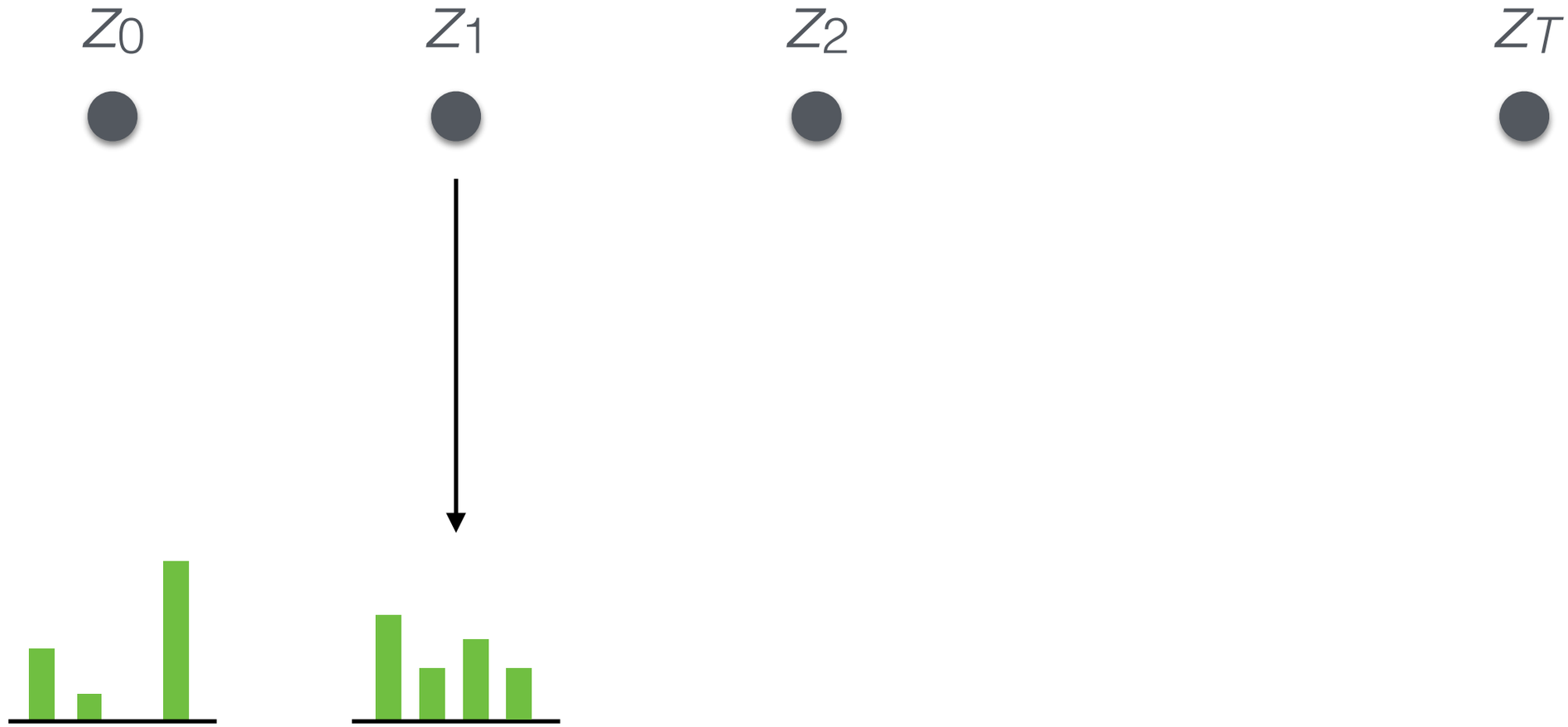


Remember?

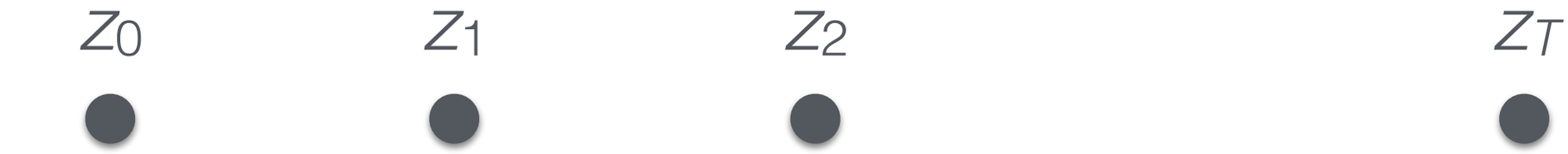
z_0 z_1 z_2 z_T



Remember?

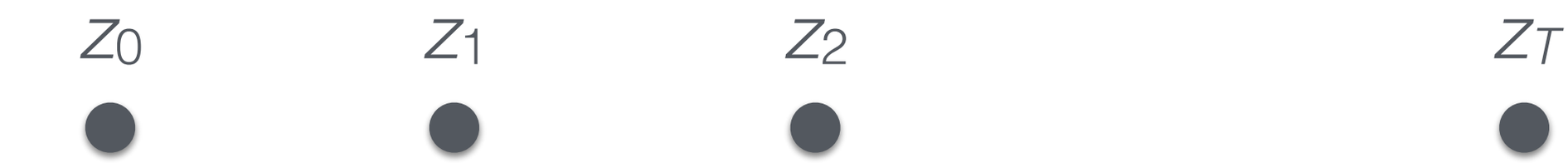


Remember?



Remember?

Remember?



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":

$$\hat{\alpha}_1 = \alpha_0 \mathsf{P}_L$$

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

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

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

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

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

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

Let's do this:

- Consider observation "R":

$$\boldsymbol{\alpha}_1 = \hat{\boldsymbol{\alpha}}_1 \text{diag}(\mathbf{O}_{L,R})$$

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

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

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

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

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

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

Let's do this:

- Again, action "Listen":

$$\hat{\alpha}_2 = \alpha_1 \mathsf{P}_L$$

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

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

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

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

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

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

Let's do this:

- … and observation "R":

$$\boldsymbol{\alpha}_2 = \hat{\boldsymbol{\alpha}}_2 \text{diag}(\mathbf{O}_{L,R})$$

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

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

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

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

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

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

Let's do this:

- Finally, normalize:

$$\begin{aligned}\boldsymbol{\mu}_{2|0:2} &= \frac{\boldsymbol{\alpha}_2}{\|\boldsymbol{\alpha}_2\|_1} \\ &= \frac{1}{0.373}\begin{bmatrix}0.011 & 0.361\end{bmatrix} \\ &= \begin{bmatrix}0.03 & 0.97\end{bmatrix}\end{aligned}$$

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

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

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

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

One extra observation?

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

Step ↓ Obs. ↓ Norm.

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

$$\rightarrow \boldsymbol{\mu}_{2|0:2}\mathsf{P}_L \mathrm{diag}(\mathsf{O}_{L,L})$$

$$\rightarrow \frac{\boldsymbol{\mu}_{2|0:2}\mathsf{P}_L \mathrm{diag}(\mathsf{O}_{L,L})}{\left\| \boldsymbol{\mu}_{2|0:2}\mathsf{P}_L \mathrm{diag}(\mathsf{O}_{L,L}) \right\|_1}$$

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

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

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

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

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

The belief

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

$$\boldsymbol{b}_{t+1} = \frac{\boldsymbol{b}_t \mathsf{P}_a \mathrm{diag}(\mathsf{O}_{a,z})}{\|\boldsymbol{b}_t \mathsf{P}_a \mathrm{diag}(\mathsf{O}_{a,z})\|_1}$$

or, component-wise, …

$$\boldsymbol{b}_{t+1}(y) = \frac{\sum_{x \in \mathcal{X}} b_t(x) \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a)}{\sum_{x, x' \in \mathcal{X}} b_t(x) \mathsf{P}(x' \mid x, a) \mathsf{O}(z \mid x', a)}$$

Belief update

• • •

The belief

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

$$\boldsymbol{b}_{t+1}(y) = \frac{\sum_{x \in \mathcal{X}} b_t(x) \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a)}{\sum_{x, x' \in \mathcal{X}} b_t(x) \mathsf{P}(x' \mid x, a) \mathsf{O}(z \mid x', a)}$$

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:

$$\boldsymbol{b}_{t+1}(y) = \frac{\sum_{x \in \mathcal{X}} b_t(x) \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a)}{\sum_{x, x' \in \mathcal{X}} b_t(x) \mathsf{P}(x' \mid x, a) \mathsf{O}(z \mid x', a)}$$

  - … the action at time t
  - … the observation at time t
  - … the belief at time $t + 1$
- The belief at time t **summarizes the history** up to time t!

The belief is Markov!

Partially observable MDPs

- **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, $\mathbf{c}$

Belief MDPs

- **Model** for sequential decision processes
- Described by:
  - State space, (the space of beliefs) $\mathscr{B}$
  - Action space, $\mathscr{A}$
  - Transition probabilities (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:

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

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \underbrace{\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x, a) \right]}_{c_B(\boldsymbol{b}, a)} + \gamma \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a) J^*(\boldsymbol{b}'_{z,a}) \Bigg]$$

Cost-to-go function

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

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

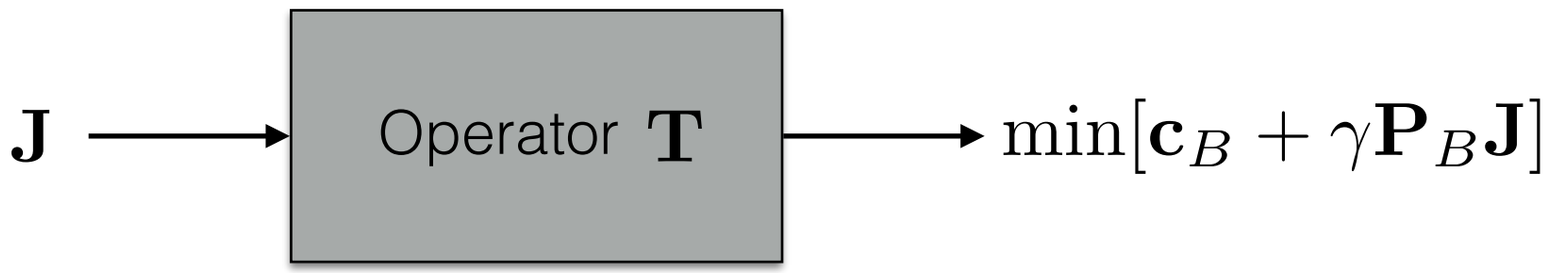
$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x, a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a) J^*(\boldsymbol{b}'_{z,a}) \right]$$

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

Value iteration for POMDPs

- We can use the recursive expression:

$$J^*(\boldsymbol{b}) = \mathbf{T}\left[\min_{a \in \mathcal{A}} \left[c_B(\boldsymbol{b}, a) + \gamma \sum_{\boldsymbol{b}'} \mathsf{P}_B(\boldsymbol{b}' \mid \boldsymbol{b}, a) J^*(\boldsymbol{b}') \right]\right]$$



… however…

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

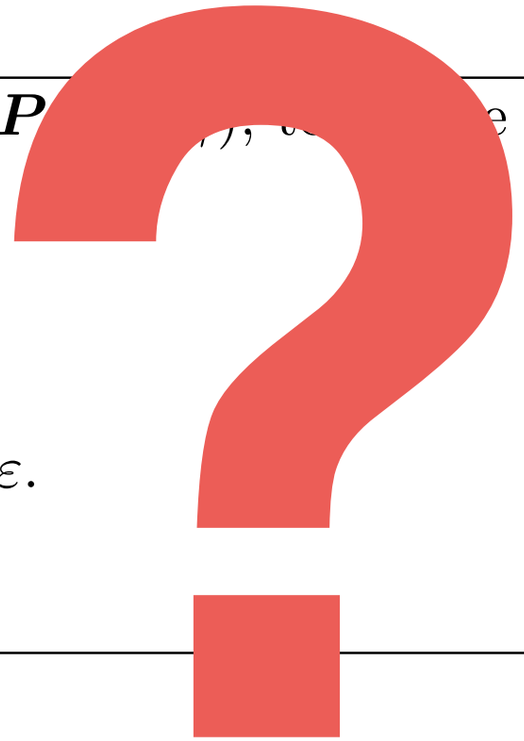
How do we represent this?

Representing J*

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

Require: MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \{\boldsymbol{P}$ $\varepsilon > 0$;

1: Initialize $k = 0$, $J^{(0)} \equiv 0$
2: **repeat**
3: $\quad J^{(k+1)} \leftarrow \mathsf{T} J^{(k)}$
4: $\quad k \leftarrow k + 1$
5: **until** $\left\| J^{(k-1)} - J^{(k)} \right\|_\infty < \varepsilon$.
6: Compute π^* using (2.21)
7: **return** π^*



The tiger problem

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?

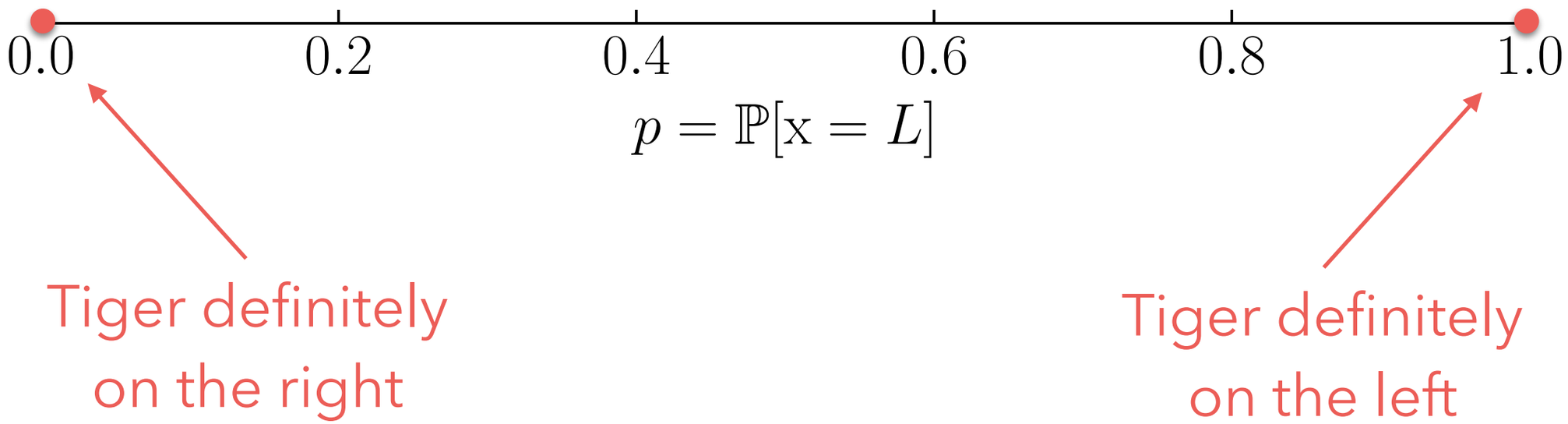
- The POMDP can be in two possible states:
  - Tiger left
  - Tiger right
- Belief $\boldsymbol{b}$ is a vector

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

Just a single number

What is the belief space?

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

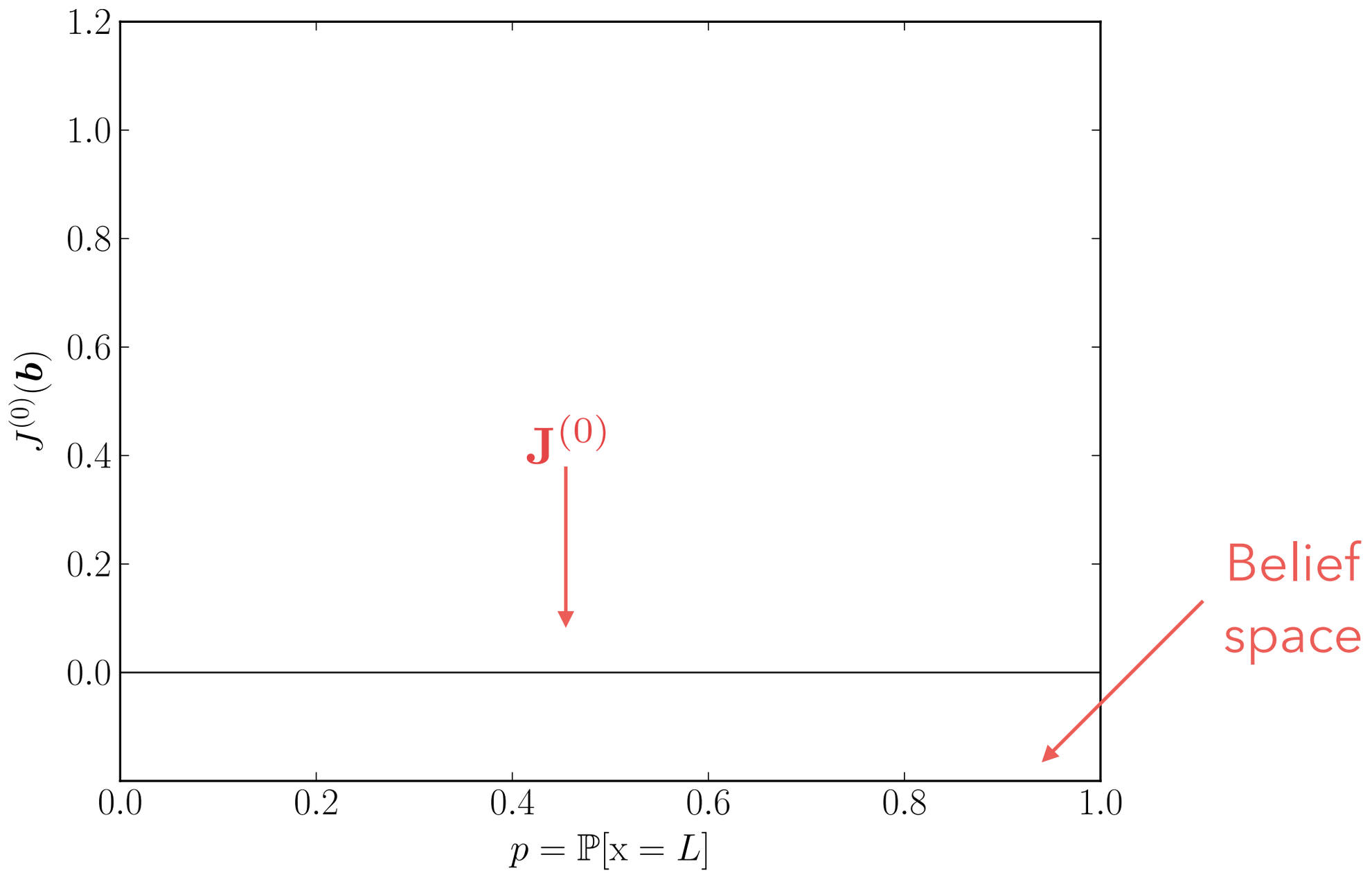


Value iteration

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

$$J^{(0)} \equiv \mathbf{0}$$

Value iteration



Value iteration

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

$$J^{(1)}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x, a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a) J^{(0)}(\boldsymbol{b}'_{z,a}) \right]$$

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

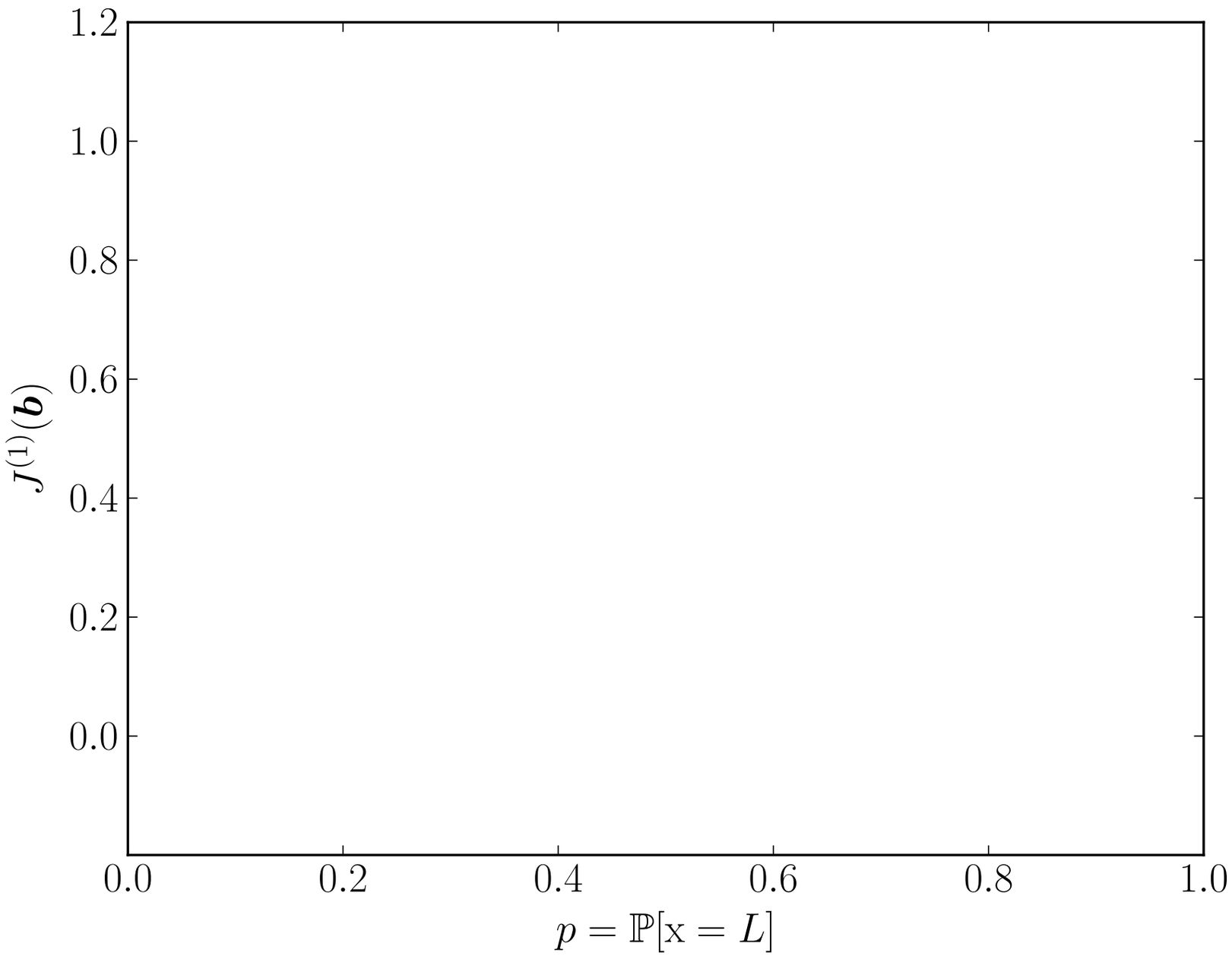
- Let’s try running VI on the tiger problem
  - Iteration 1:

$$J^{(1)}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

Value iteration

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

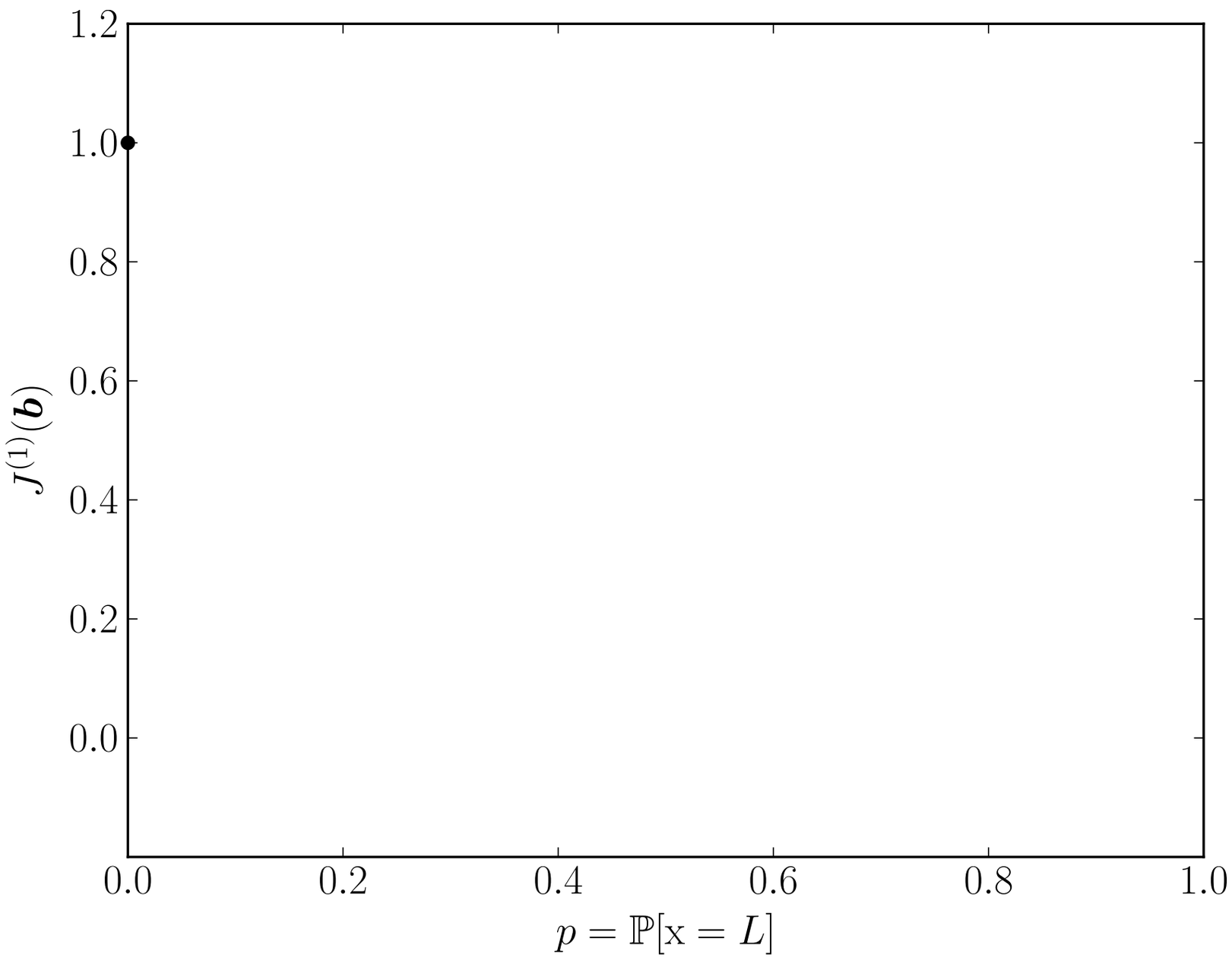
Action OR



Value iteration

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

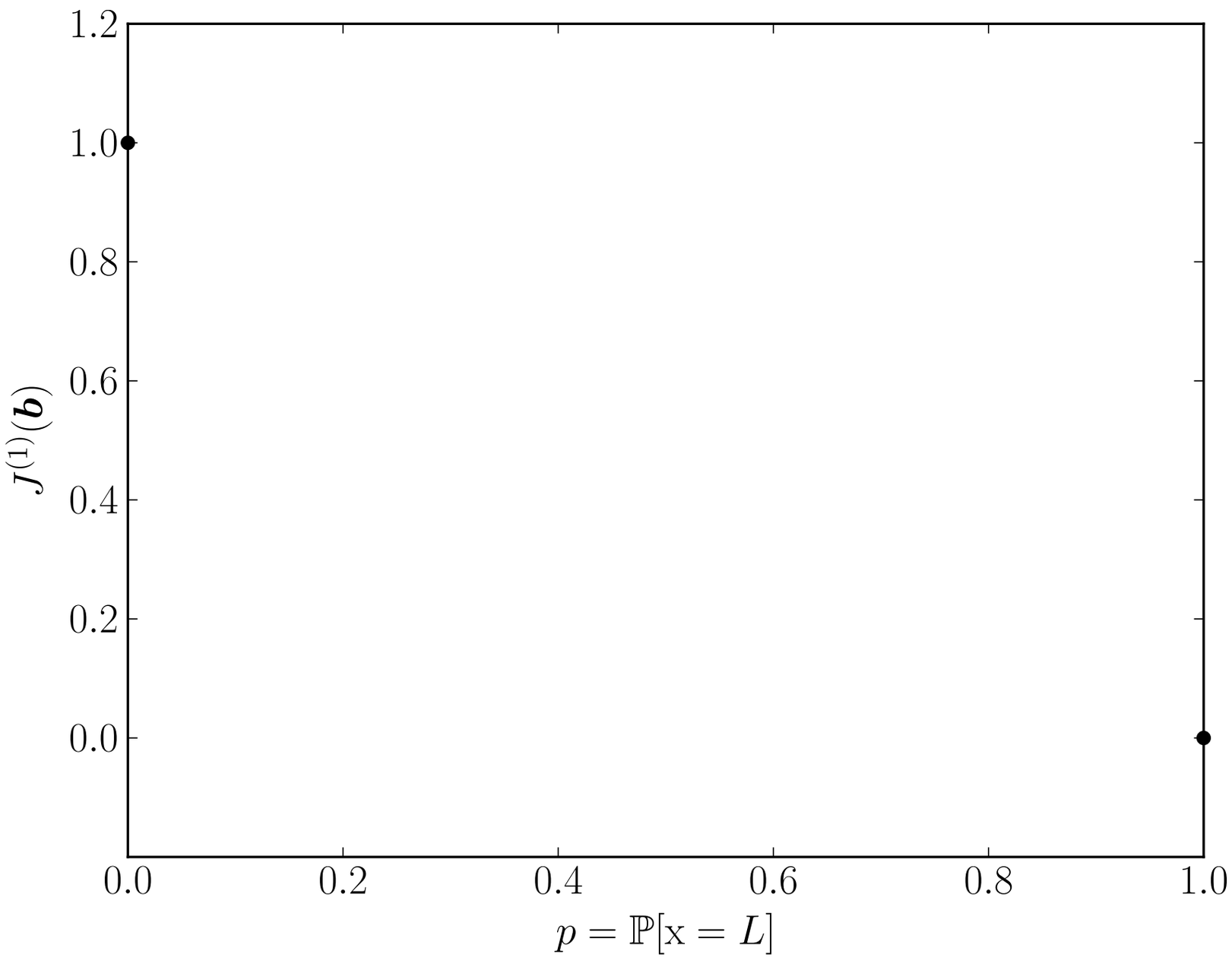
Action OR



Value iteration

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

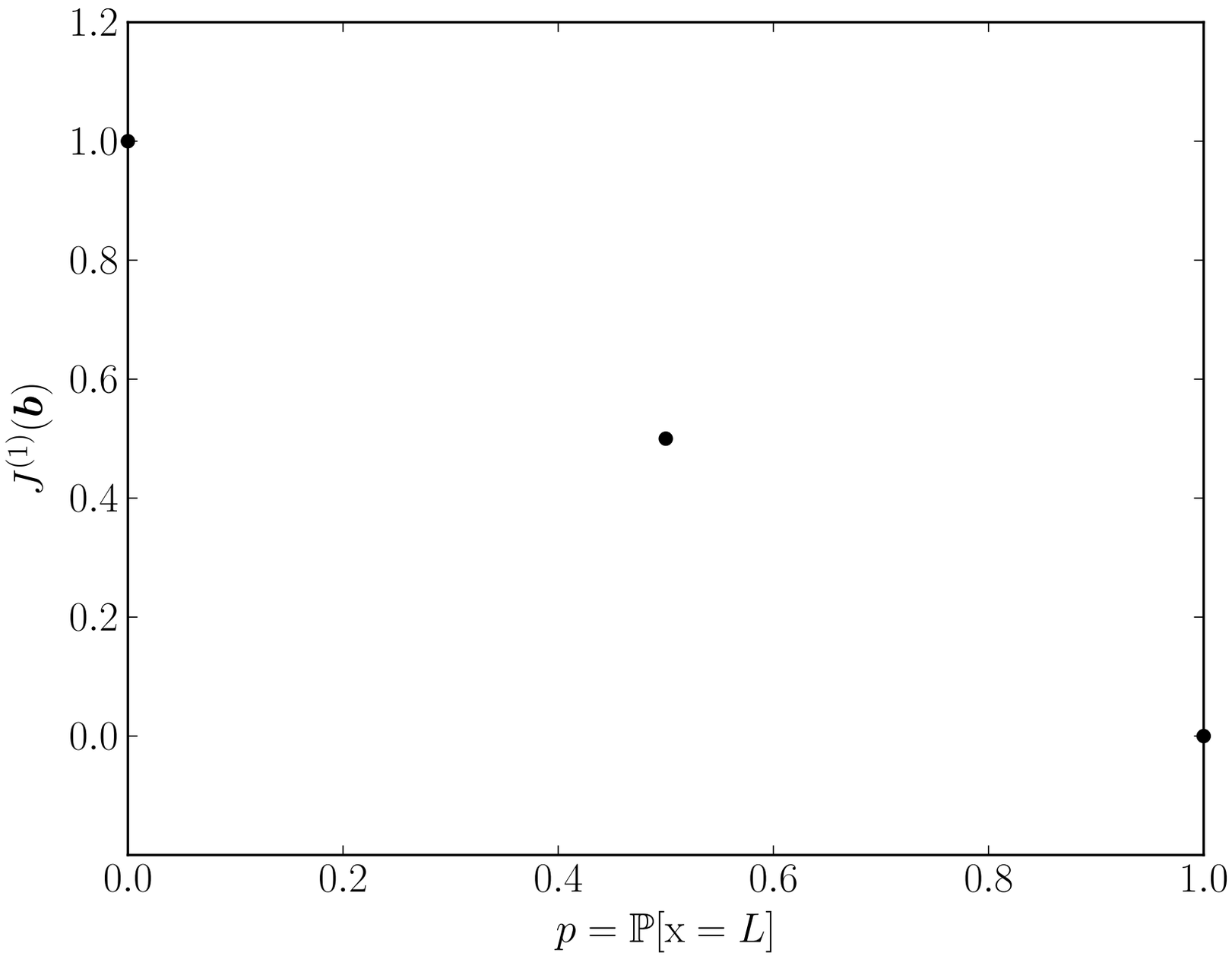
Action OR



Value iteration

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

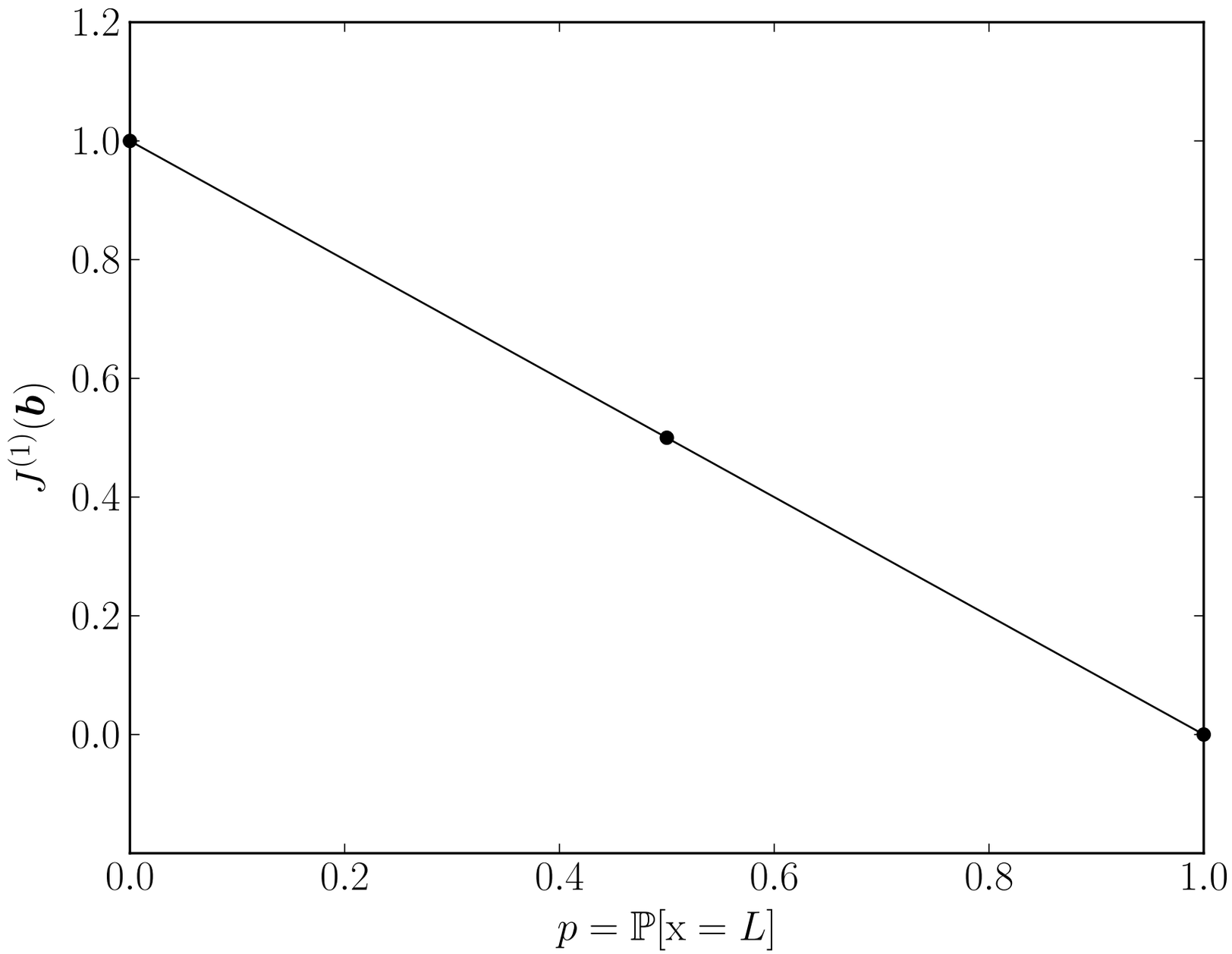
Action OR



Value iteration

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

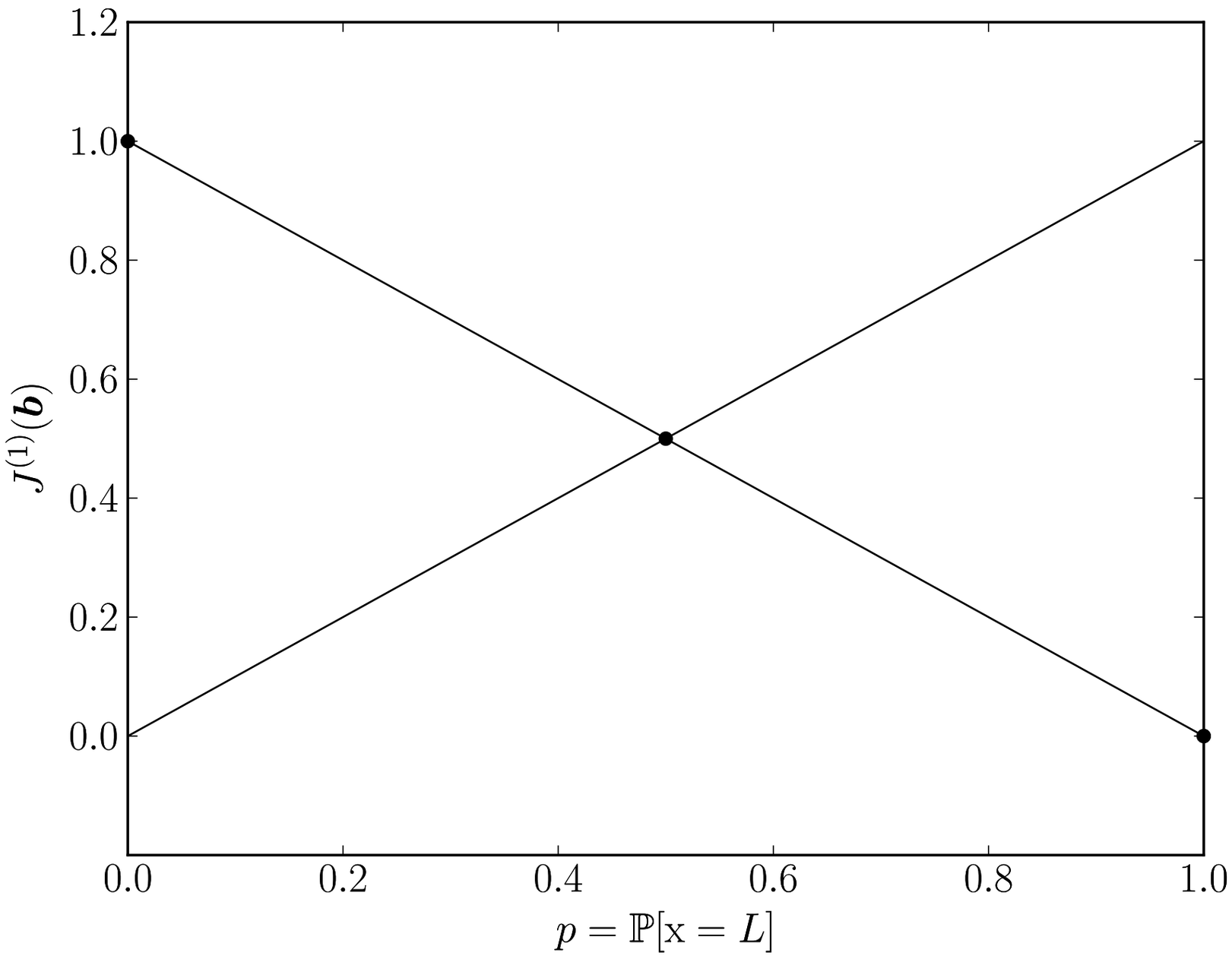
Action OR



Value iteration

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

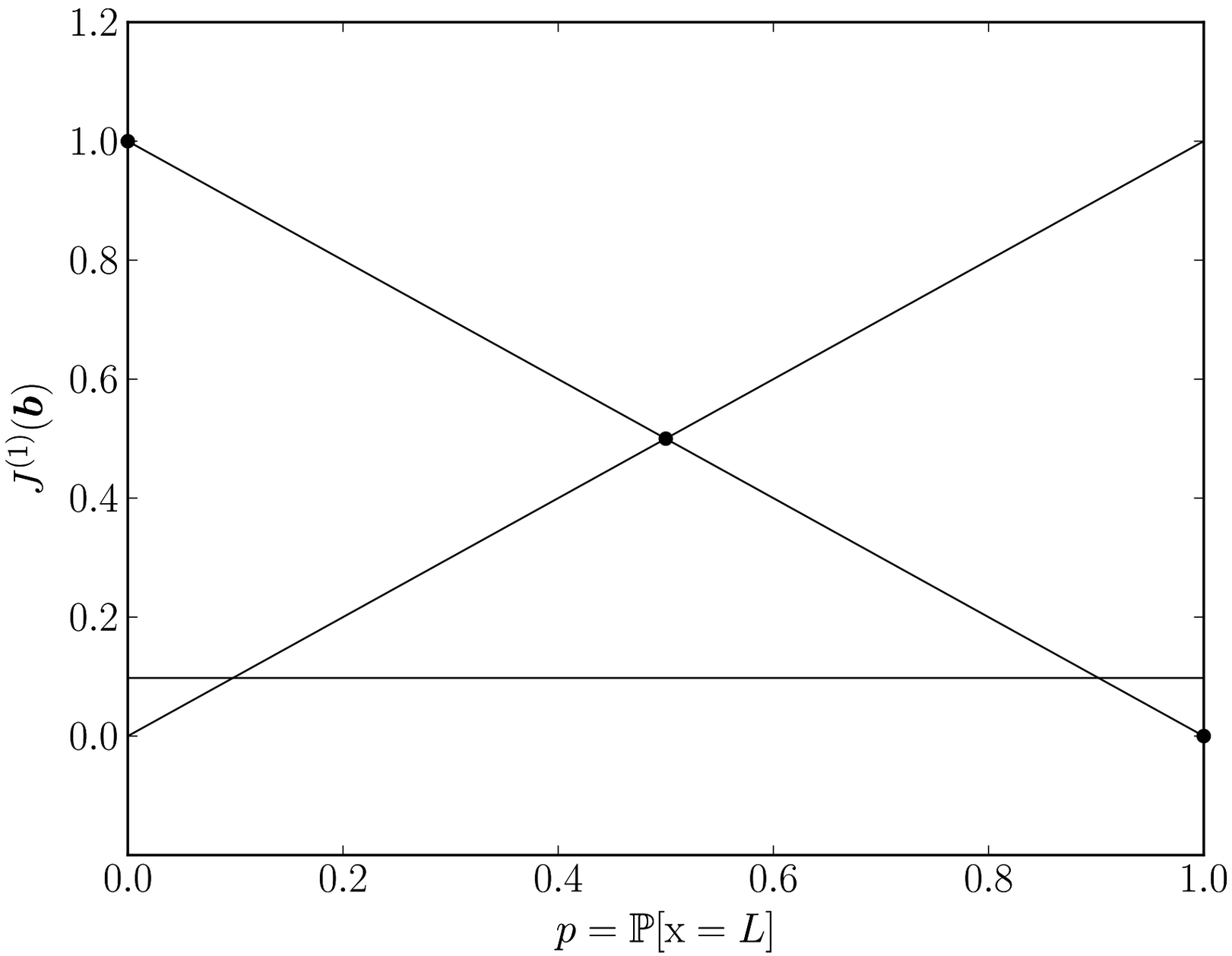
Action OL



Value iteration

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

Action L



Value iteration

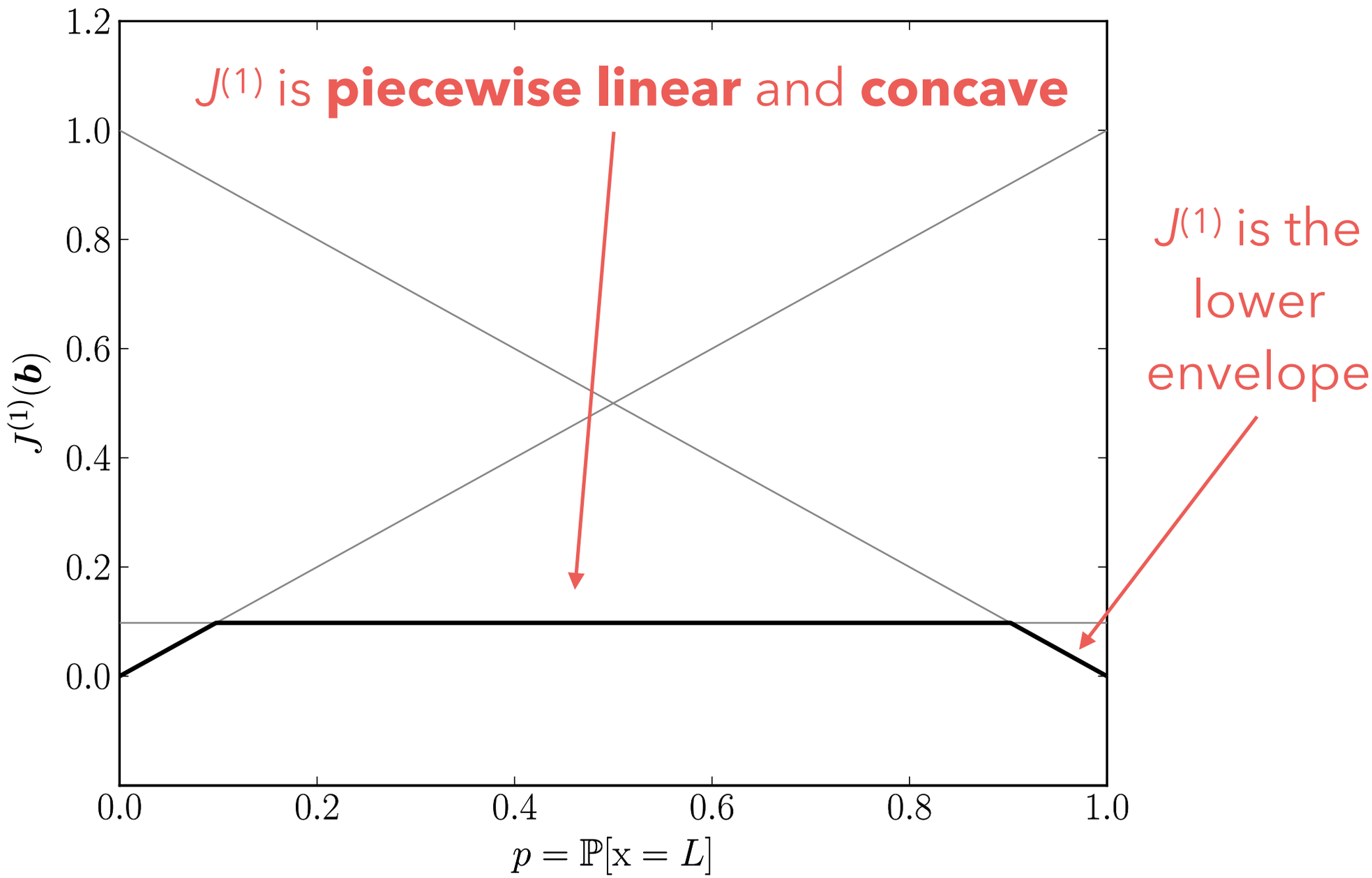
- Iteration 1:

$$J^{(1)}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x, a)$$

- We just computed, for every belief **b** and action *a*,

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

Value iteration



Representing $J^{(k)}$

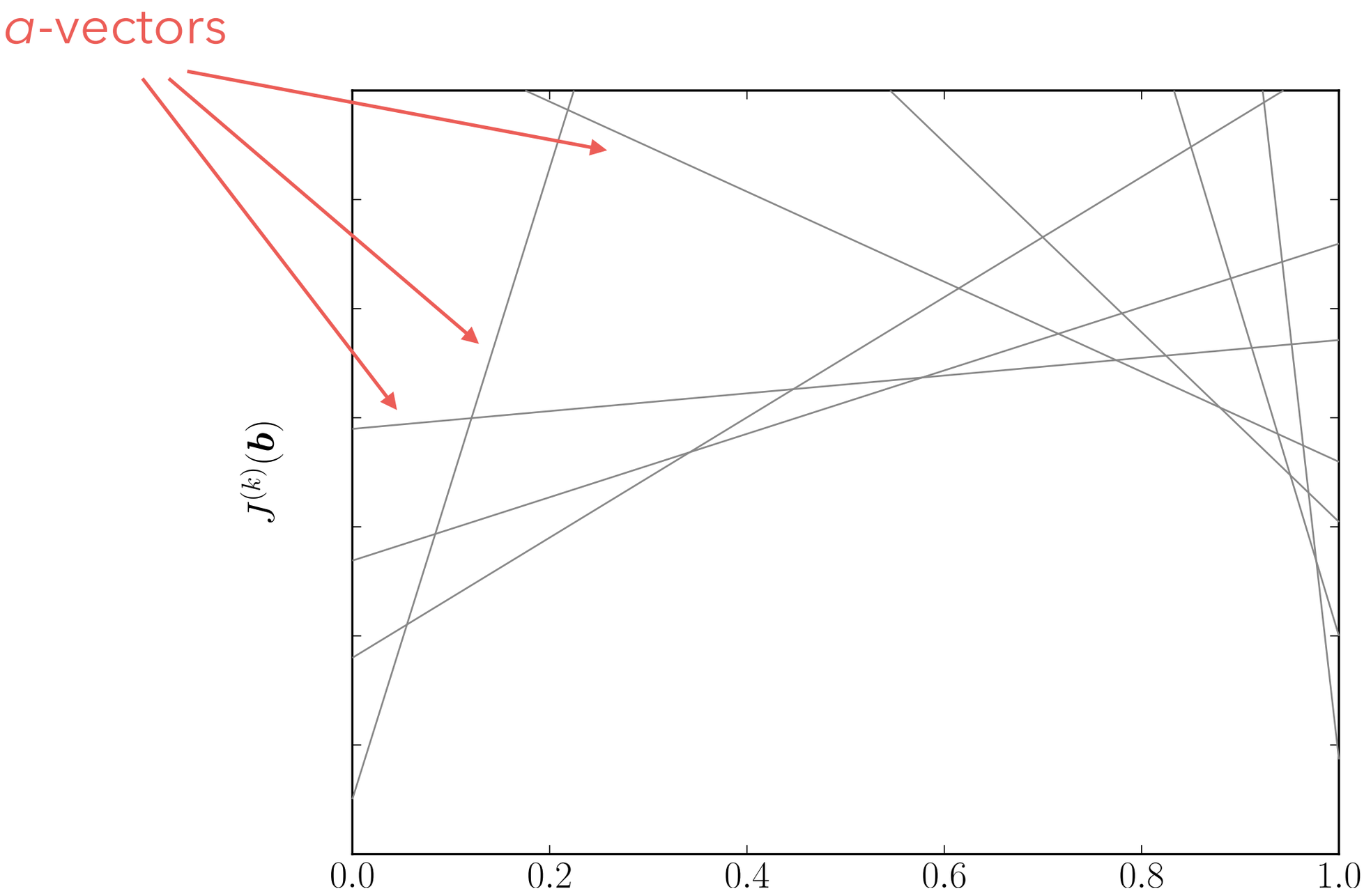
- The cost-to-go at each iteration of VI is always PWLC
  - Can always be written in the form

$$J^{(k)}(\boldsymbol{b}) = \min_{\alpha \in \Gamma} \boldsymbol{b} \cdot \boldsymbol{\alpha}$$

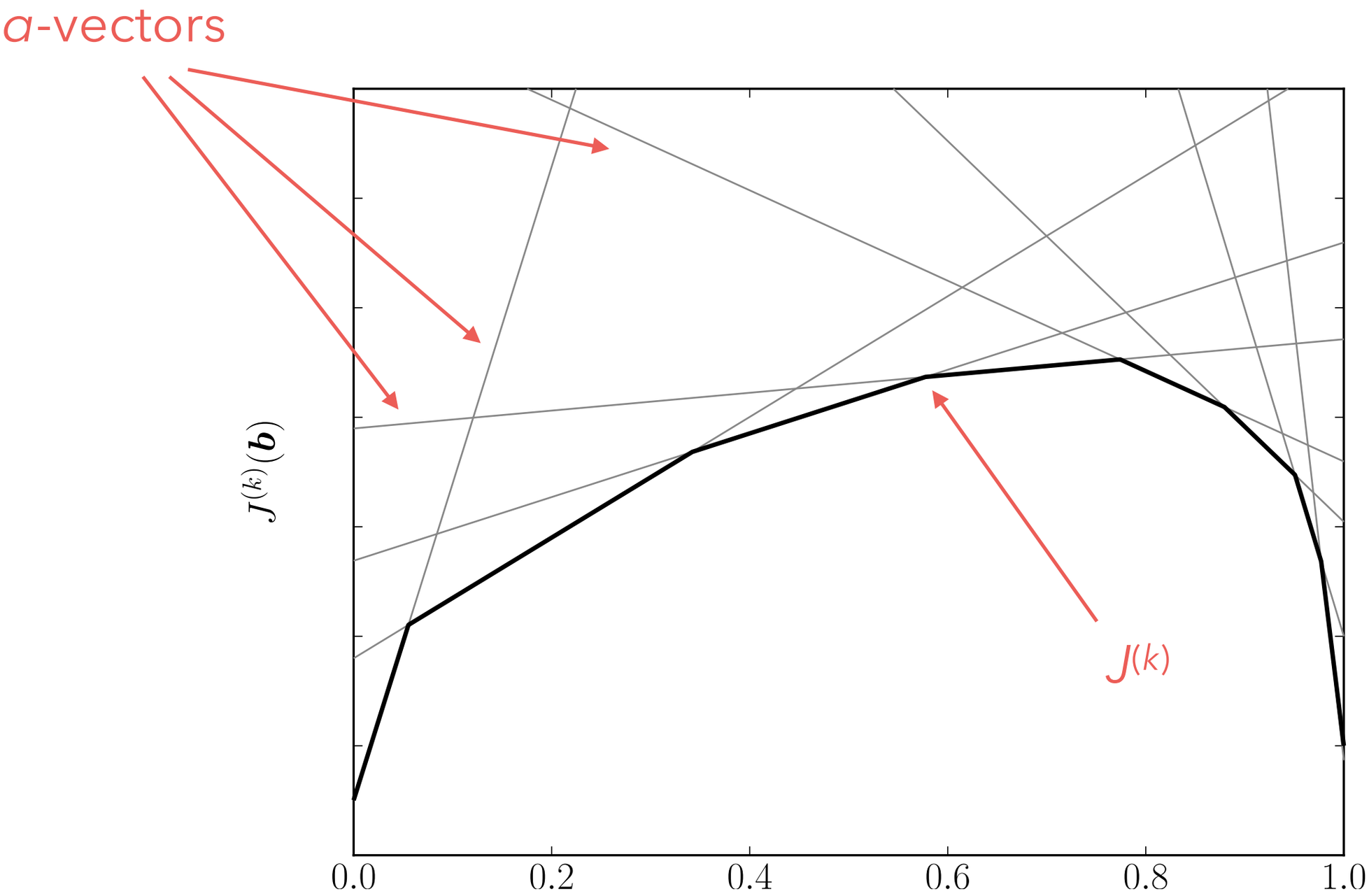
$$= \min_{\alpha \in \Gamma} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x)\boldsymbol{\alpha}(x)$$

Set of vectors used in the representation

Representing $J^{(k)}$



Representing $J^{(k)}$



Value iteration

- How do we compute these α-vectors?

$$J^{(k+1)}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x,a) \mathsf{O}(z \mid y,a) J^{(k)}(\boldsymbol{b}'_{z,a}) \right]$$

Replace representation

Value iteration

- How do we compute these α-vectors?

$$J^{(k+1)}(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x,a) \mathsf{O}(z \mid y,a) \min_{\boldsymbol{\alpha} \in \Gamma^{(k)}} \boldsymbol{b}'_{z,a} \cdot \boldsymbol{\alpha} \right]$$

Minimizing α-vector depends on ***b***

Value iteration

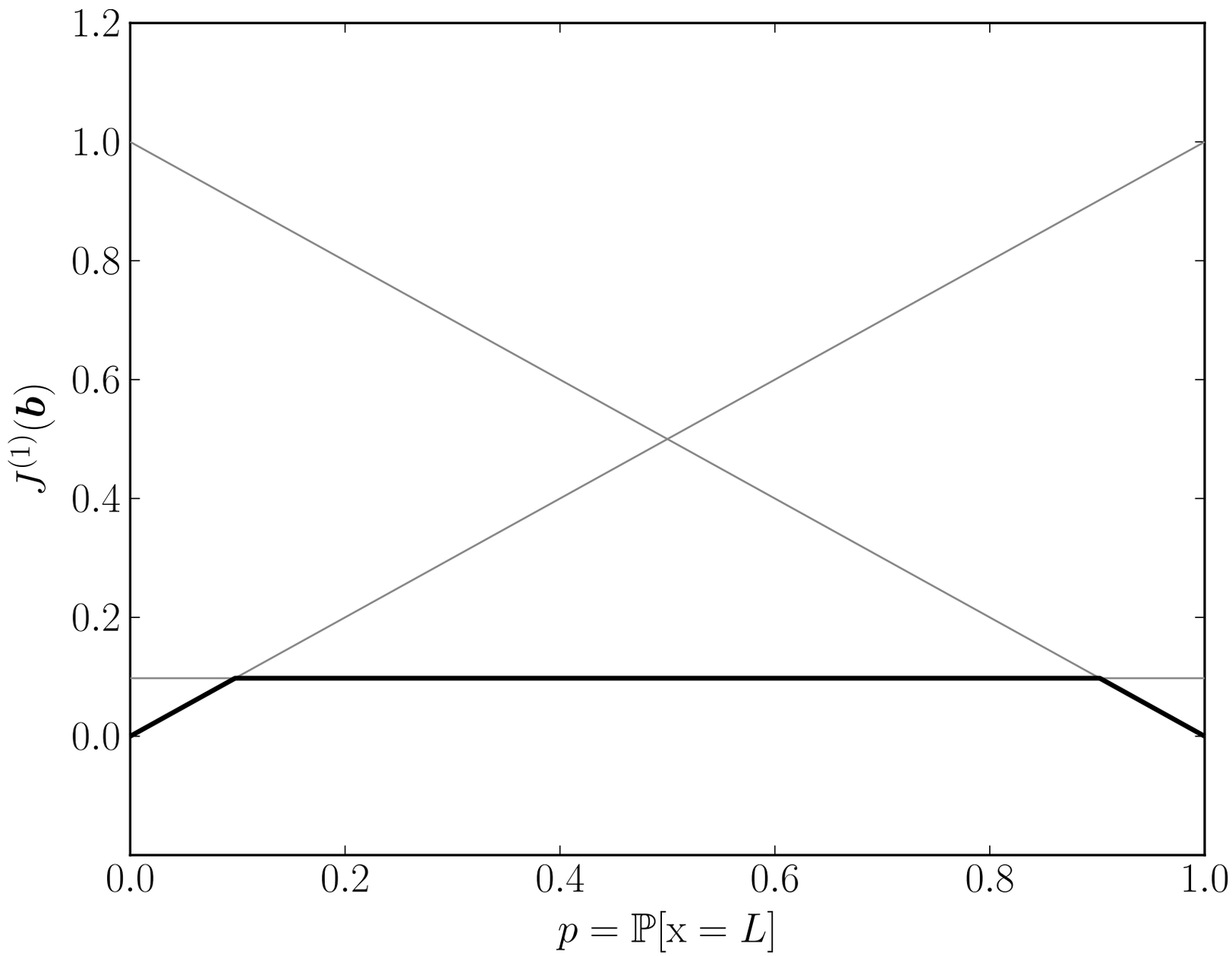
- Compute, at each iteration $k + 1$, the set $\Gamma^{(k+1)}$ from $\Gamma^{(k)}$
  - For each $\boldsymbol{\alpha} \in \Gamma^{(k)}$, compute

$$\boldsymbol{\alpha}_{a,z}^{(k)} = \frac{1}{|\mathcal{Z}|}\boldsymbol{C}_{:,a} + \gamma \mathrm{P}_a \mathrm{diag}(\boldsymbol{O}_{z,a})\boldsymbol{\alpha}$$

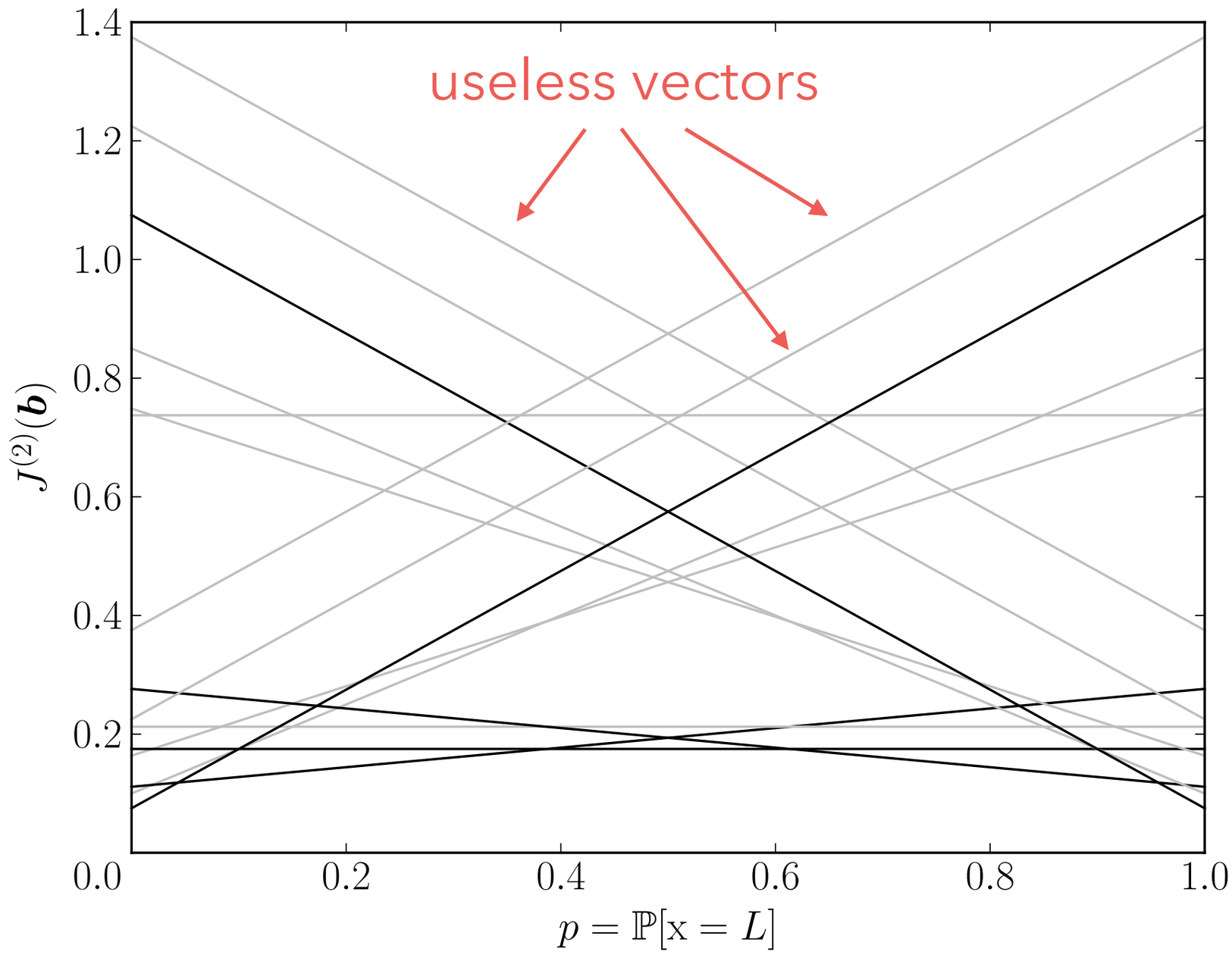
- Compute **all possible combinations** of $\boldsymbol{\alpha}_{a,z}^{(k)}$, for each z
- For each combination, let

$$\boldsymbol{\alpha}_{a}^{(k)} = \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}_{z,a}^{(k)}$$

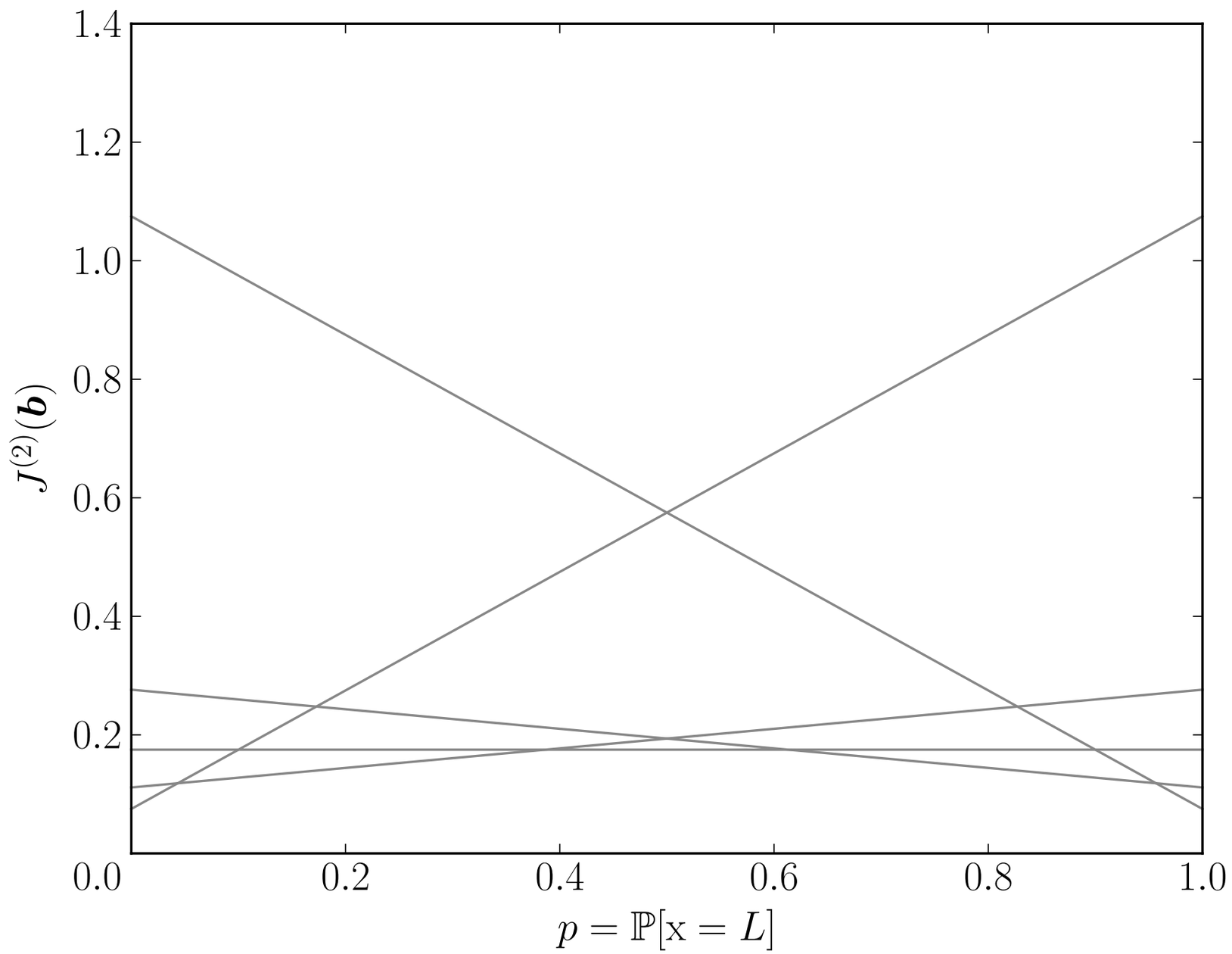
Value iteration



Value iteration



Value iteration



Value iteration

- Two approaches to build $\Gamma^{(k+1)}$ from $\Gamma^{(k)}$:
  - **Region based methods:** Start with empty $\Gamma^{(k+1)}$ and only add vectors that are necessary

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

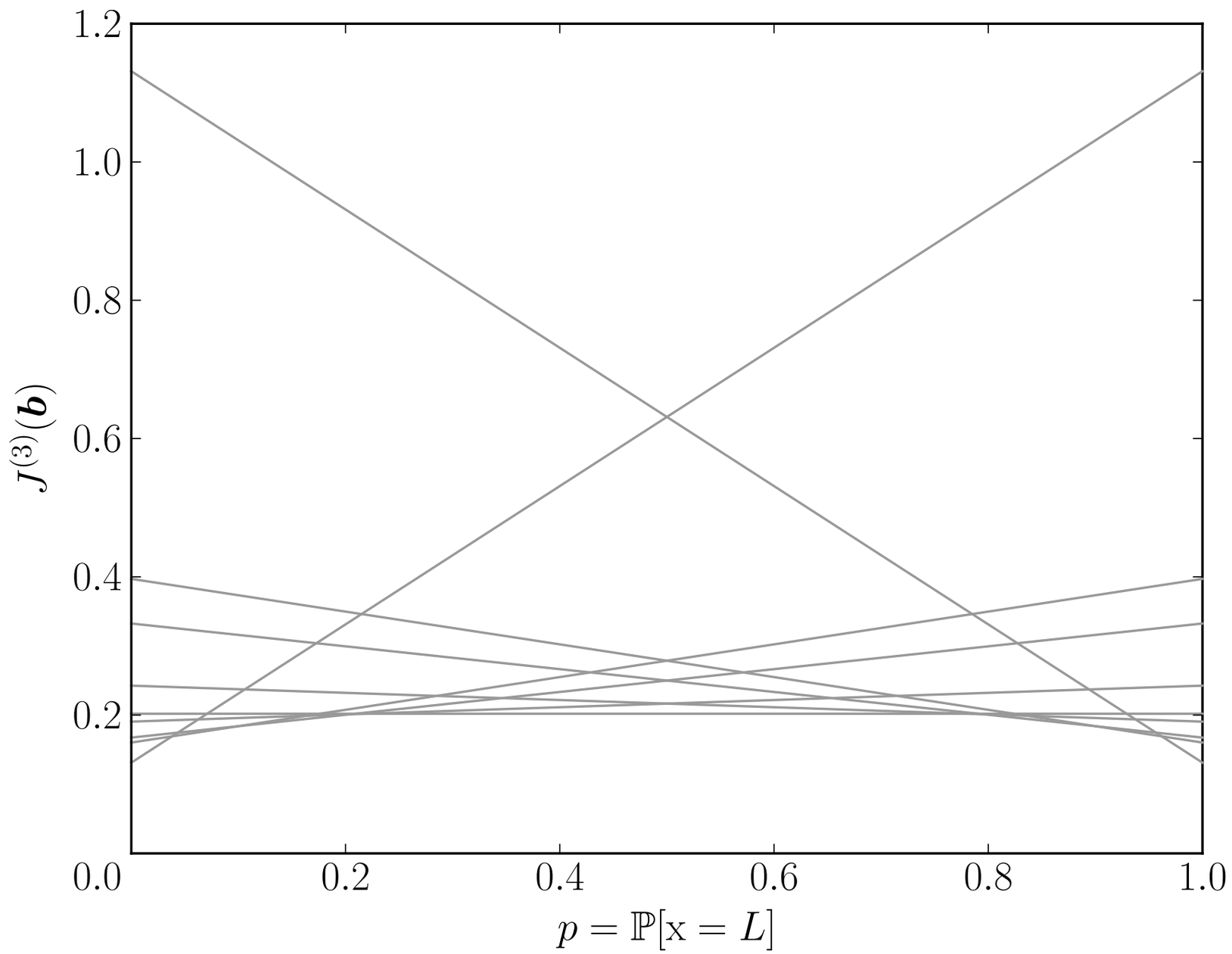
Value iteration

- Two approaches to build $\Gamma^{(k+1)}$ from $\Gamma^{(k)}$:
  - **Region based methods:** Start with empty $\Gamma^{(k+1)}$ and only add vectors that are necessary

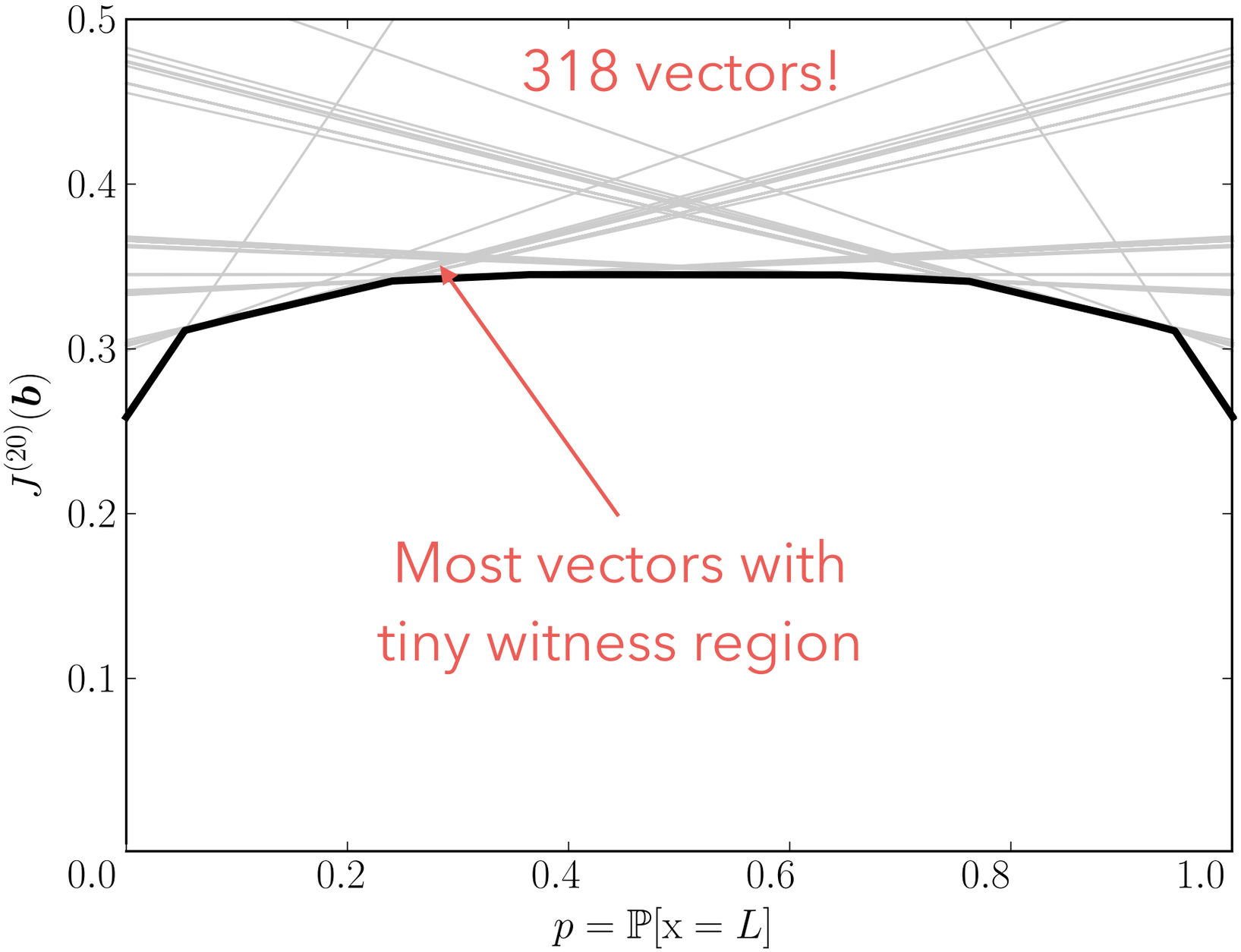
    Example: Witness algorithm
  - **Pruning-based methods:** Start with complete $\Gamma^{(k+1)}$ and remove vectors that are unnecessary

    Example: Incremental pruning

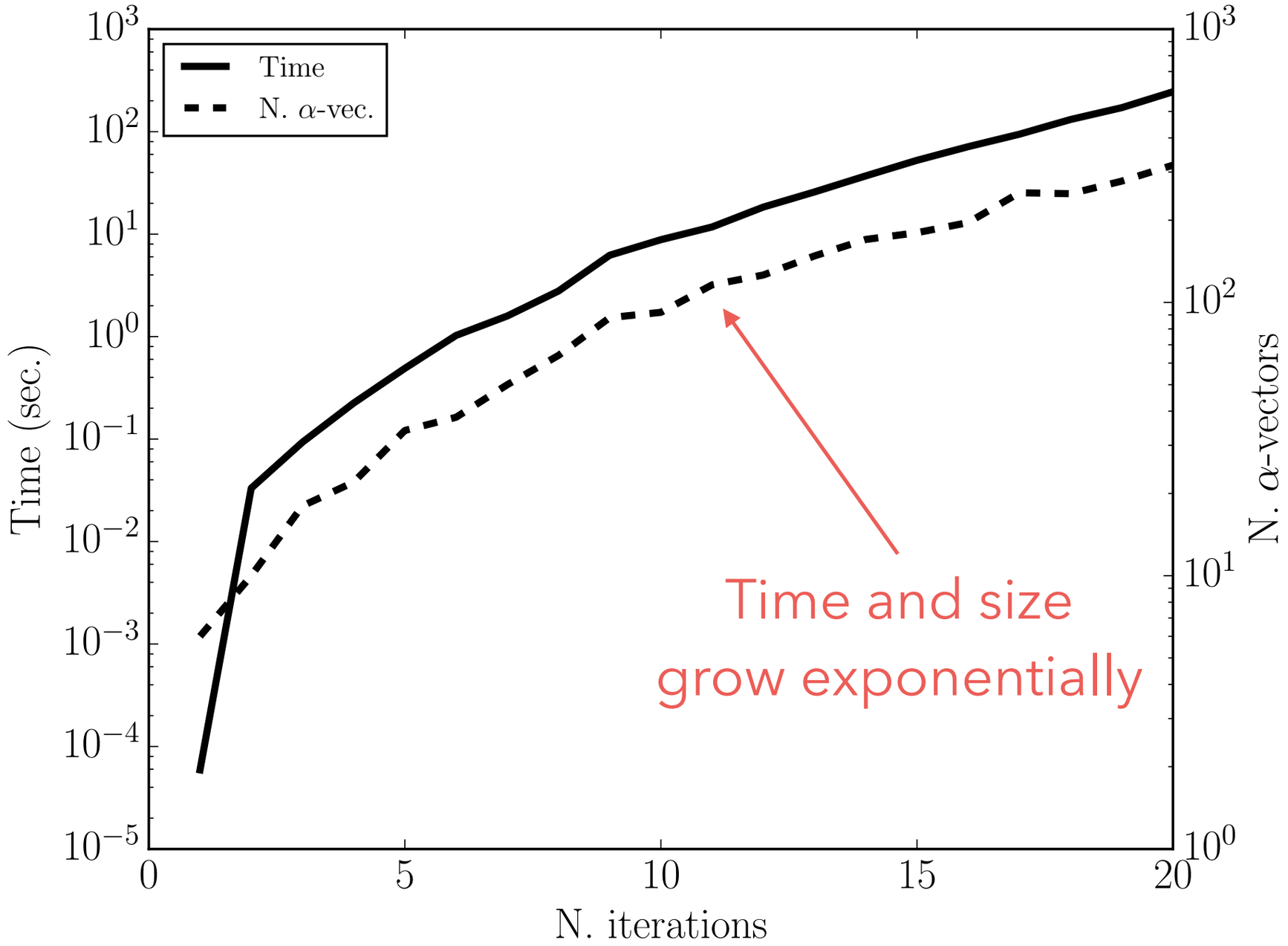
Value iteration



Value iteration



Computation time



Non-exact solutions

Idea n. 1 - Use the MDP

- "Under" a POMDP there is an MDP

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

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

MDP heuristics

MLS heuristic

- Suppose that π_{MDP} is 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

- Suppose that π_{MDP} is 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

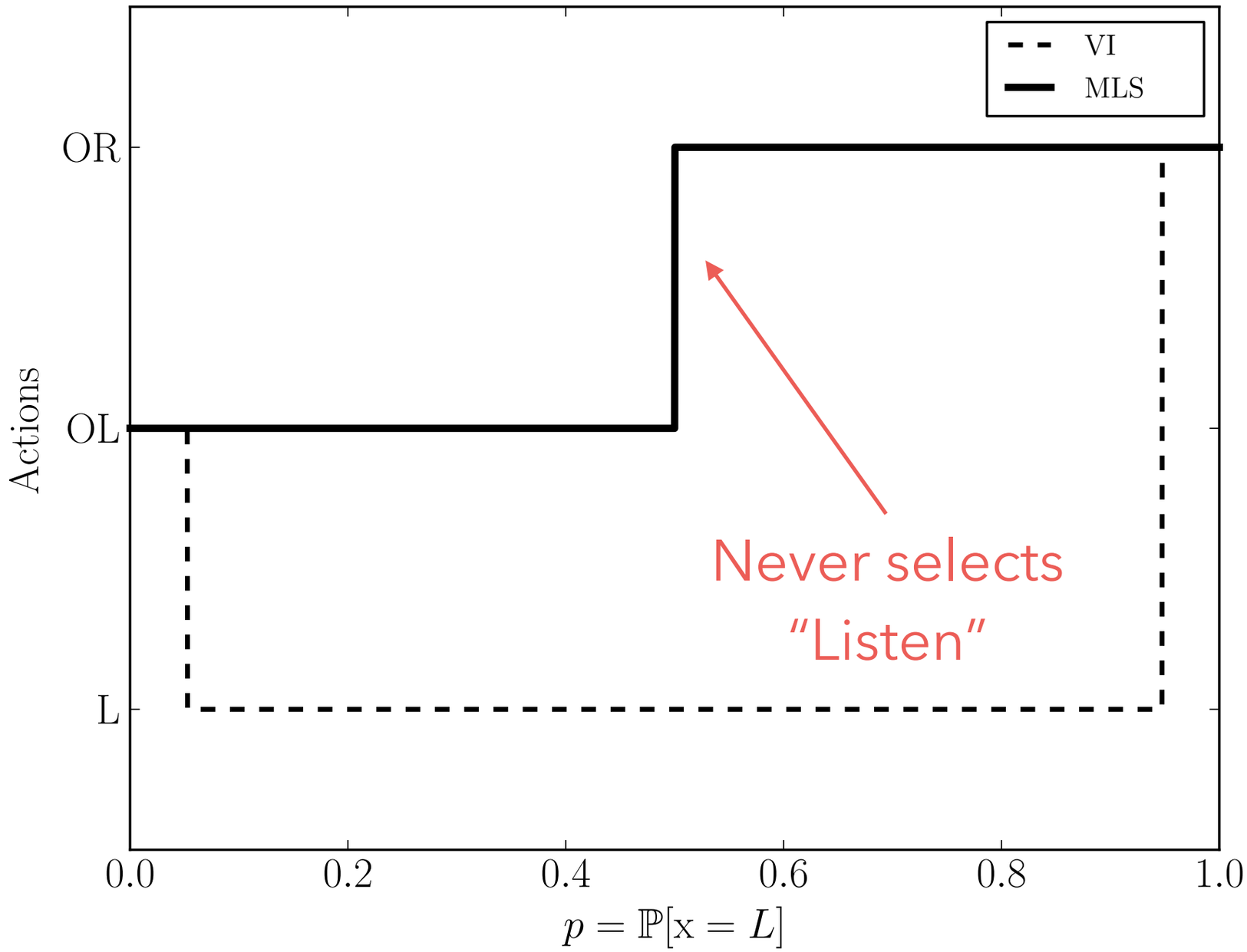
- Select the most likely state (state with largest probability)
- Execute corresponding action – $\pi_{\text{MDP}}(x_2)$

MLS heuristic

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

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

MLS heuristic



AV heuristic

- Suppose that π_{MDP} is 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

- Suppose that π_{MDP} is the optimal MDP policy
- At time step t, we have a belief

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

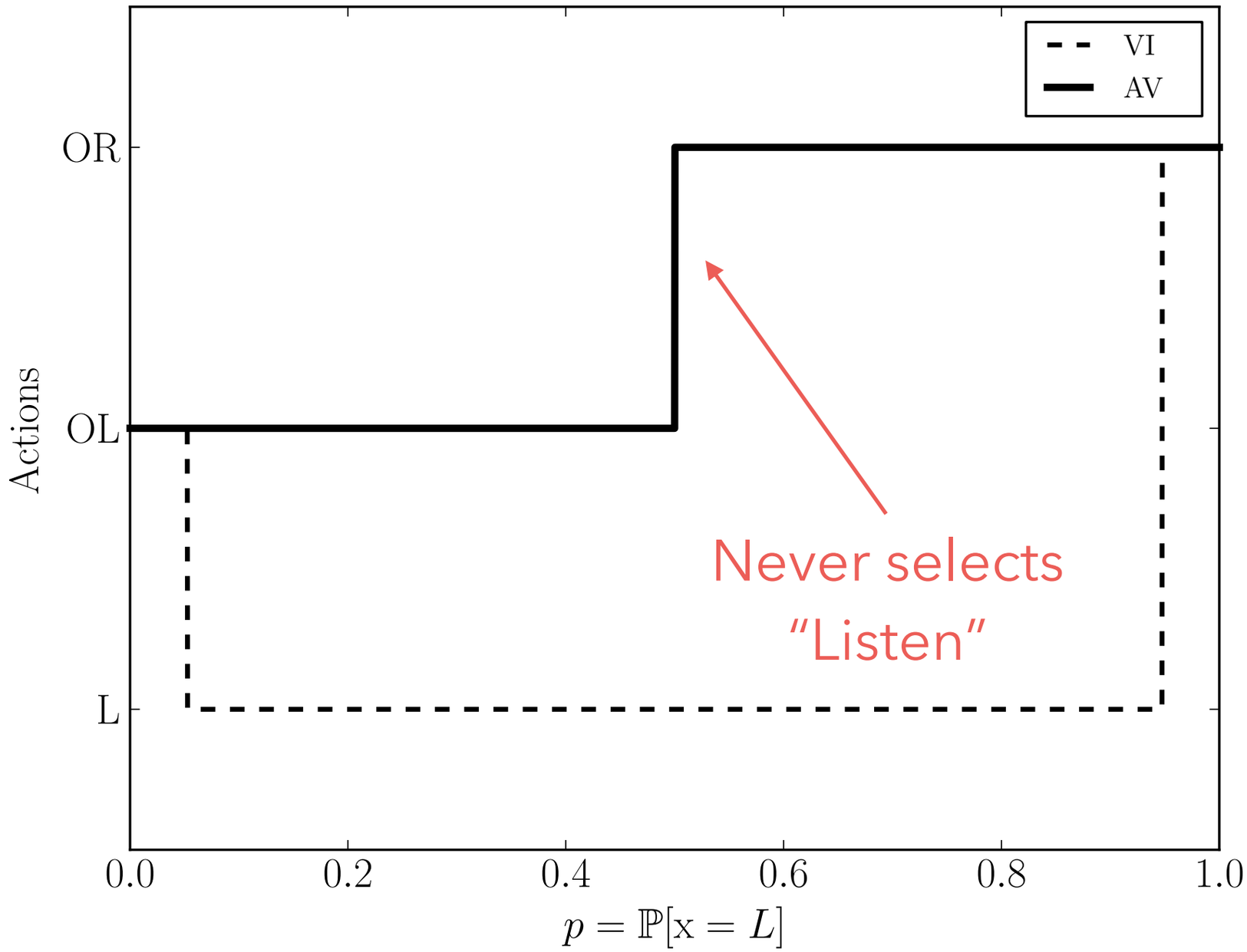
- States vote proportionally to their probability
- Execute most voted action

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



Q-MDP heuristic

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

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

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?

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

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?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \underbrace{\left[c(x, a) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x, a) J_{\mathrm{MDP}}(y) \right]}_{Q_{\mathrm{MDP}}(x, a)}$$

Q-MDP heuristic

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

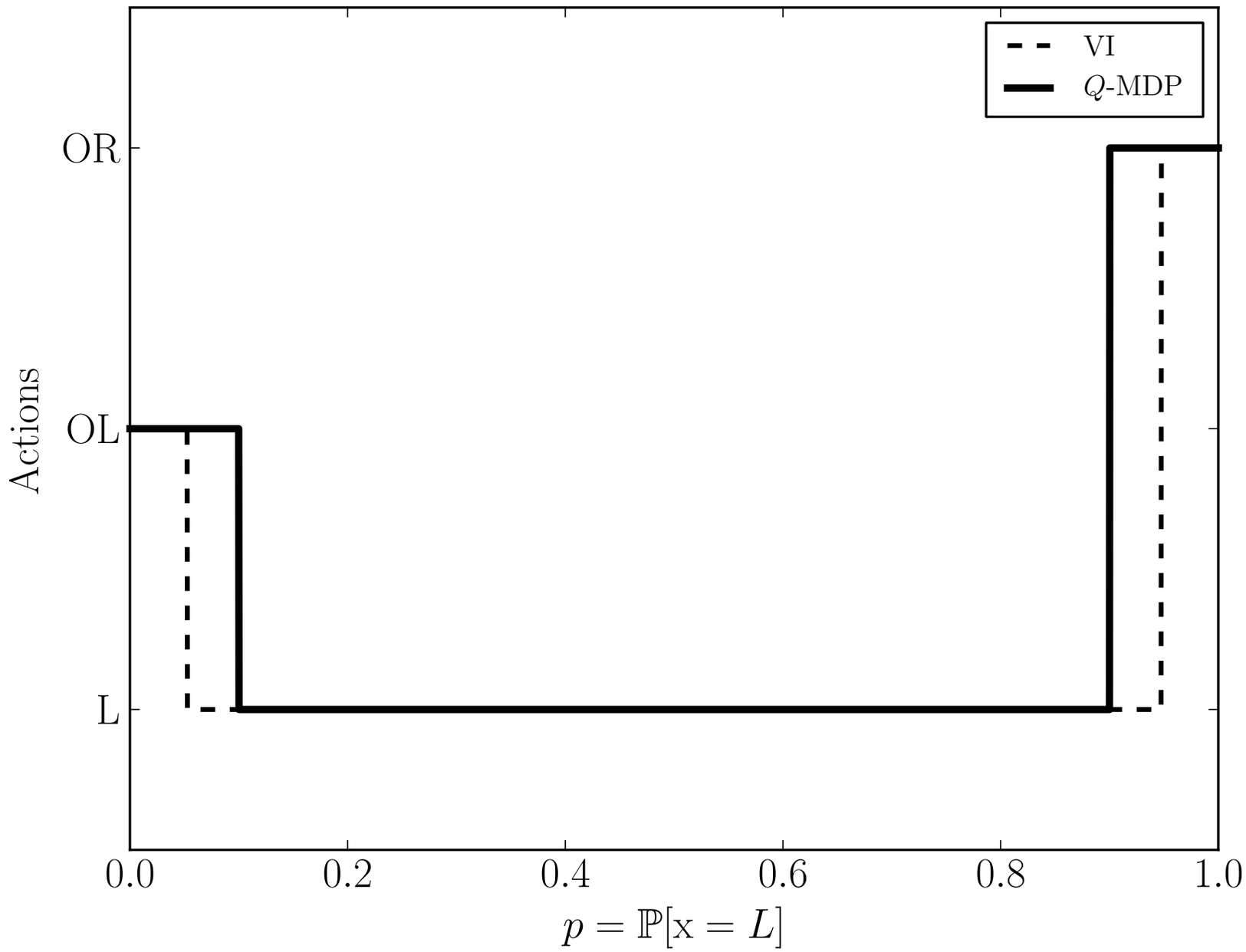
$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q_{\mathrm{MDP}}(x, a)$$

Q-MDP heuristic

- The Q-MDP heuristic is, then

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

Q-MDP heuristic



Q-MDP heuristic

- Q-MDP ignores partial observability from the next step on
- What does this mean in terms of cost-to-go approximation?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \left[c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x,a) \mathsf{O}(z \mid y,a) \min_{\boldsymbol{\alpha} \in \Gamma^*} \boldsymbol{b}'_{z,a} \cdot \boldsymbol{\alpha} \right]$$

Q-MDP heuristic

- Q-MDP ignores partial observability from the next step on
- What does this mean in terms of cost-to-go approximation?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \min_{\boldsymbol{\alpha} \in \Gamma^*} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \boldsymbol{b}(x) \mathsf{P}(y \mid x,a) \mathsf{O}(z \mid y,a) \boldsymbol{\alpha}_y \right]$$

Optimal

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \boldsymbol{b}(x) \mathsf{P}(y \mid x,a) \mathsf{O}(z \mid y,a) \min_{\boldsymbol{\alpha} \in \Gamma^*} \boldsymbol{\alpha}_y \right]$$

Q-MDP

Ignores partial observability

Q-MDP heuristic

- Q-MDP ignores partial observability from the next step on
- What does this mean in terms of cost-to-go approximation?

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \min_{\boldsymbol{\alpha} \in \Gamma^*} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \boldsymbol{b}(x) \mathsf{P}(y \mid x,a) \mathsf{O}(z \mid y,a) \boldsymbol{\alpha}_y \right]$$

Optimal

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \boldsymbol{b}(x) \mathsf{P}(y \mid x,a) \min_{\boldsymbol{\alpha} \in \Gamma^*} \boldsymbol{\alpha}_y \right]$$

Q-MDP

FIB heuristic

- FIB lies somewhere in the middle
- We include partial observability but factor out the belief

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \min_{\boldsymbol{\alpha} \in \Gamma^*} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} \boldsymbol{b}(x) \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a) \boldsymbol{\alpha}_y \right]$$

FIB heuristic

- FIB lies somewhere in the middle
- We include partial observability but factor out the belief

$$J^*(\boldsymbol{b}) = \min_{a \in \mathcal{A}} \left[\sum_{x \in \mathcal{X}} \boldsymbol{b}(x) c(x,a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) \min_{\boldsymbol{\alpha} \in \Gamma^*} \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x, a) \mathsf{O}(z \mid y, a) \boldsymbol{\alpha}_y \right]$$

FIB heuristic

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

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

FIB heuristic

- The FIB heuristic is, then

$$\pi_{\text{FIB}}(\boldsymbol{b}) = \underset{a \in \mathcal{A}}{\text{argmax}} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) Q_{\text{FIB}}(x, a)$$

FIB heuristic