TÉCNICO LISBOA

Planning, Learning and Decision Making

Lecture 16. Learning and decision-making

Approaches to learning

- "Symbolic" approach → Learn rules (DT)
- Probabilistic approach → Learn probabilities (LR, NB)
- Similarity-based approach → Learn by similarity (kNN, SVM)
- Neural approach → Brain-like learning (NN)

Approaches to learning

- Each approach assumes specific structure:
  - Symbolic → classes can be described by small number of rules
  - Probabilistic → classes can be described probabilistically
  - Similarity based → classes can be described "geometrically"
  - Neural based → classes can be described "geometrically" by bending the space

What about MDPs?

- What if we assume that the classes can be described as an MDP?
  - The examples come from the optimal policy of an MDP
  - The examples may be noisy

… then…

- We can:

1. … try to bring the MDP structure to one of the approaches used

→ MDP-induced metric

or

2. … "invert" the MDP

→ Inverse reinforcement learning

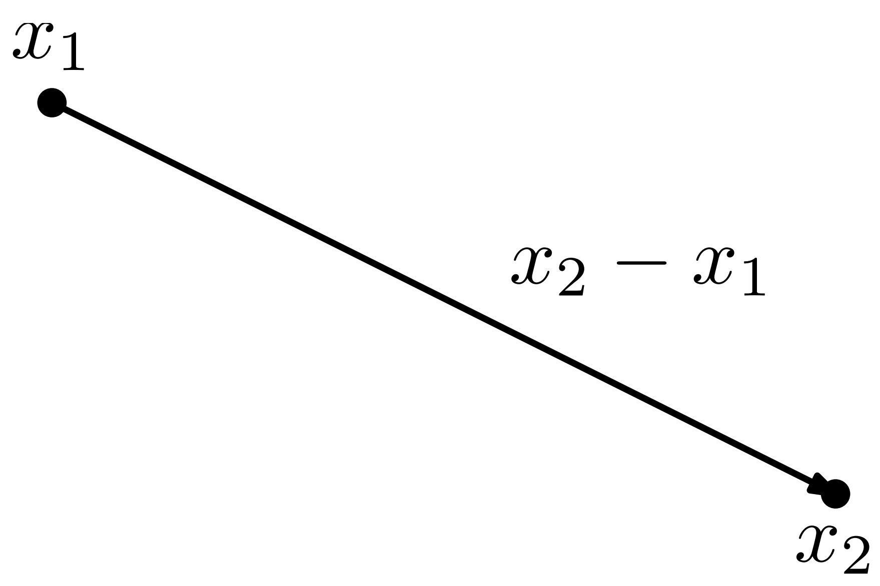
MDP-induced metric

What is a “metric”?

- A way to measure distances
- Example: Euclidean distance (ℓ_2)

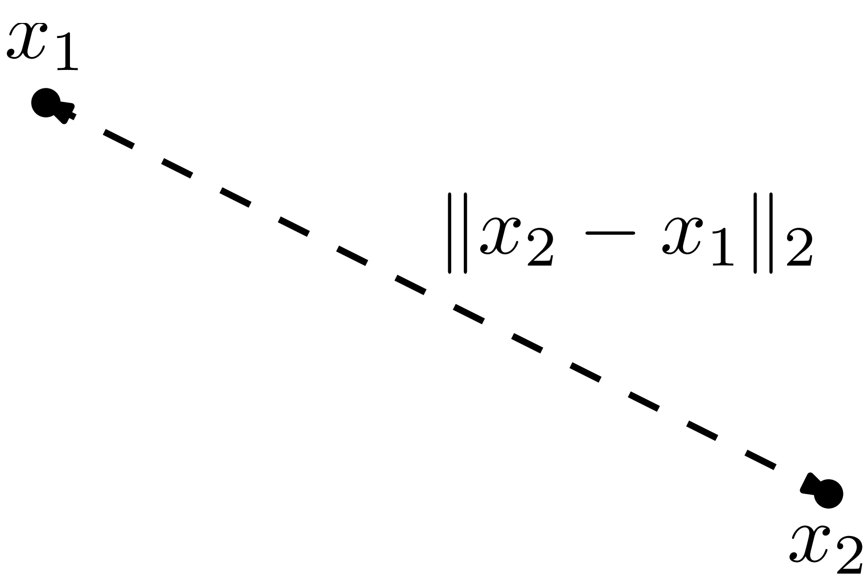
What is a “metric”?

- A way to measure distances
- Example: Euclidean distance (ℓ_2)



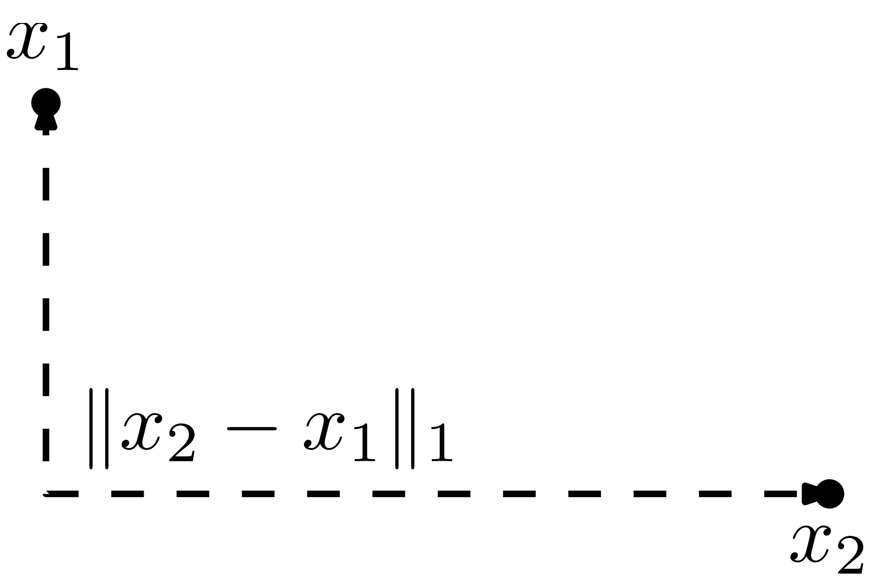
What is a "metric"?

- A way to measure distances
- Example: Euclidean distance (ℓ_2)



What is a "metric"?

- A way to measure distances
- Example: Manhattan distance (ℓ_1)



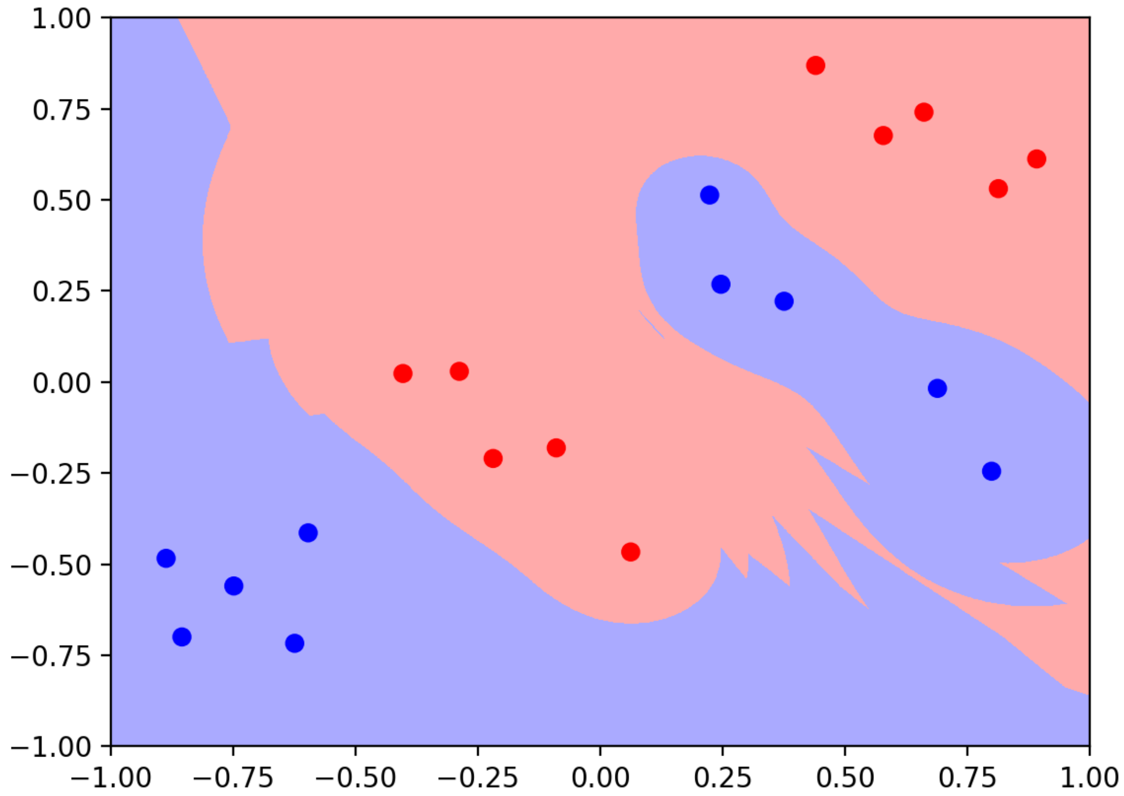
What is a “metric”?

- A way to measure **similarity**
  - Points that are close are similar
  - Points that are far are dissimilar

TÉCNICO LISBOA

Why do we care?

- Metrics are at the core of similarity-based methods:
  - kNN relies critically on the notion of metric
  - It selects the action of a point based on nearby points



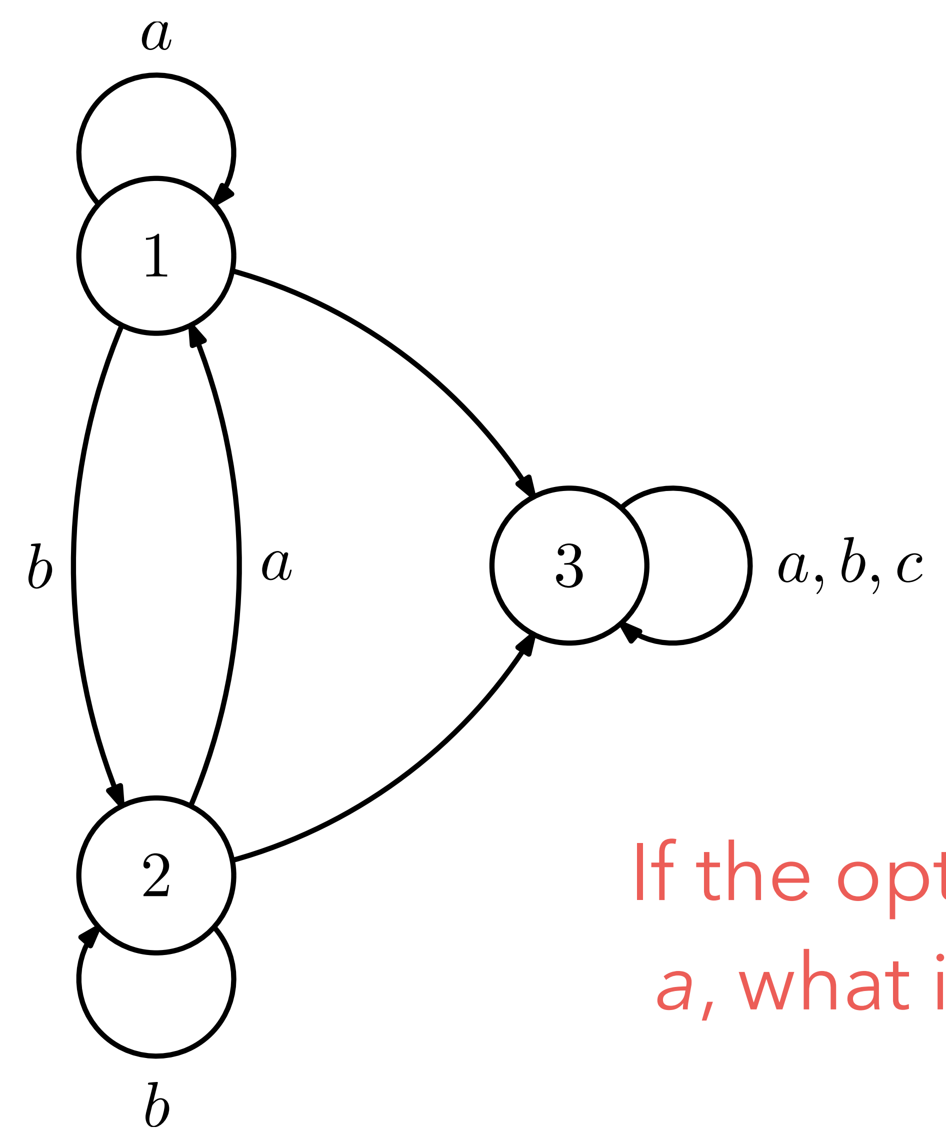
Can we define a metric between states in an MDP?

Metric for MDPs

- Given a metric for MDPs…
    - … examples of actions in one state can be generalized to “nearby” states
- Then, if the “teacher” follows the optimal policy…
    - … the learner can learn the optimal policy without planning

Measuring similarity

- Consider the MDP (without costs):



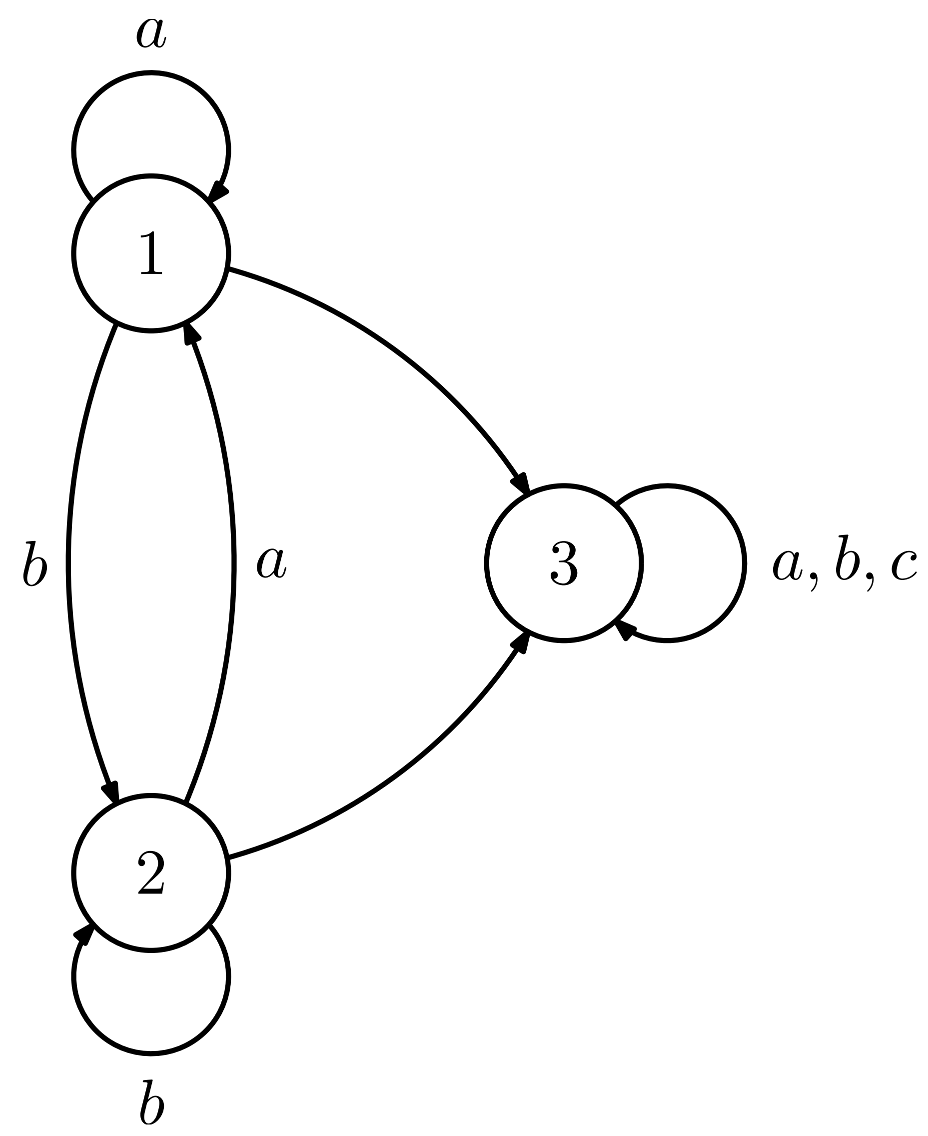
If the optimal action in state 1 is *a*, what is the optimal action in state 2?

Measuring similarity

- If all costs are allowed, no generalization is possible

Why?

All policies are possible, so no generalization between states is possible



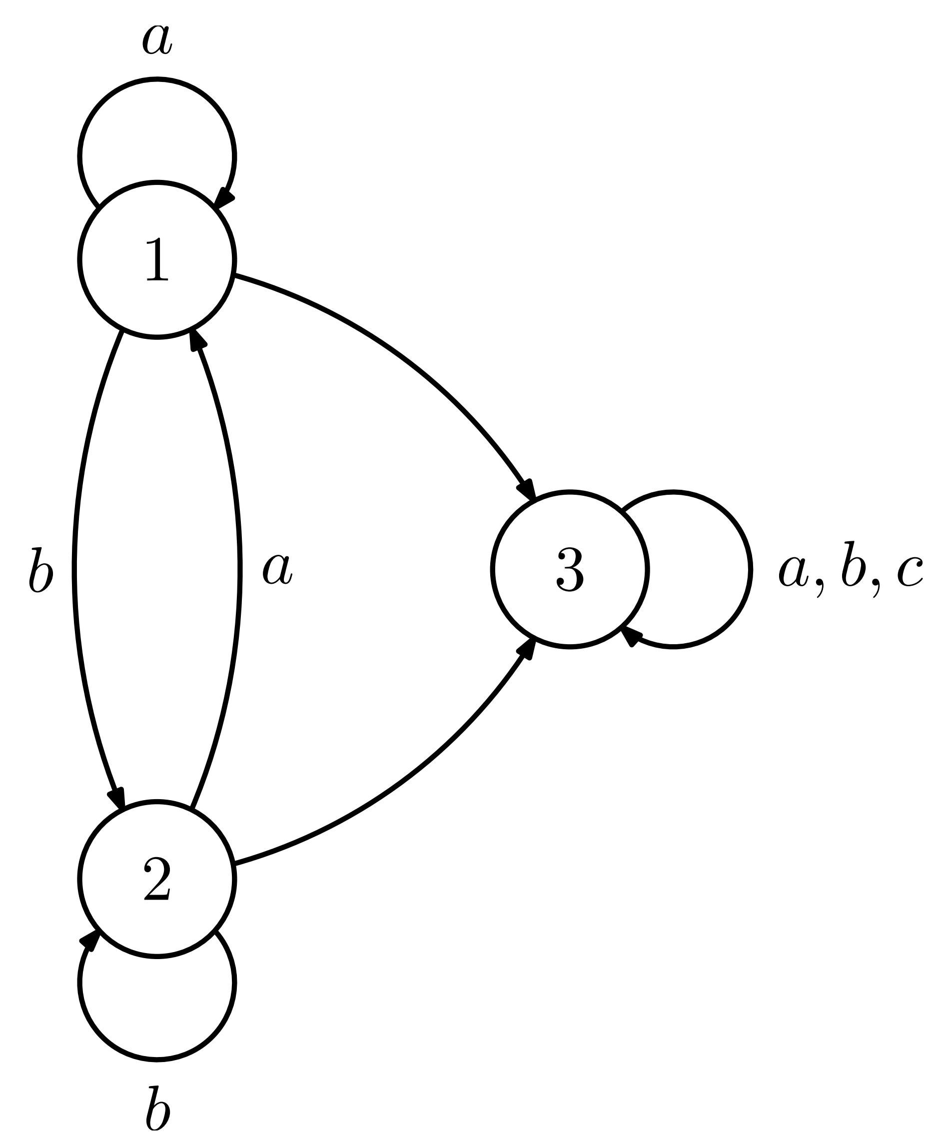
Measuring similarity

- Inductive bias assumption: **costs are the same for all actions** (only depend on state)
- If the optimal action in state 1 is *a*, what is the optimal action in state 2?

↓

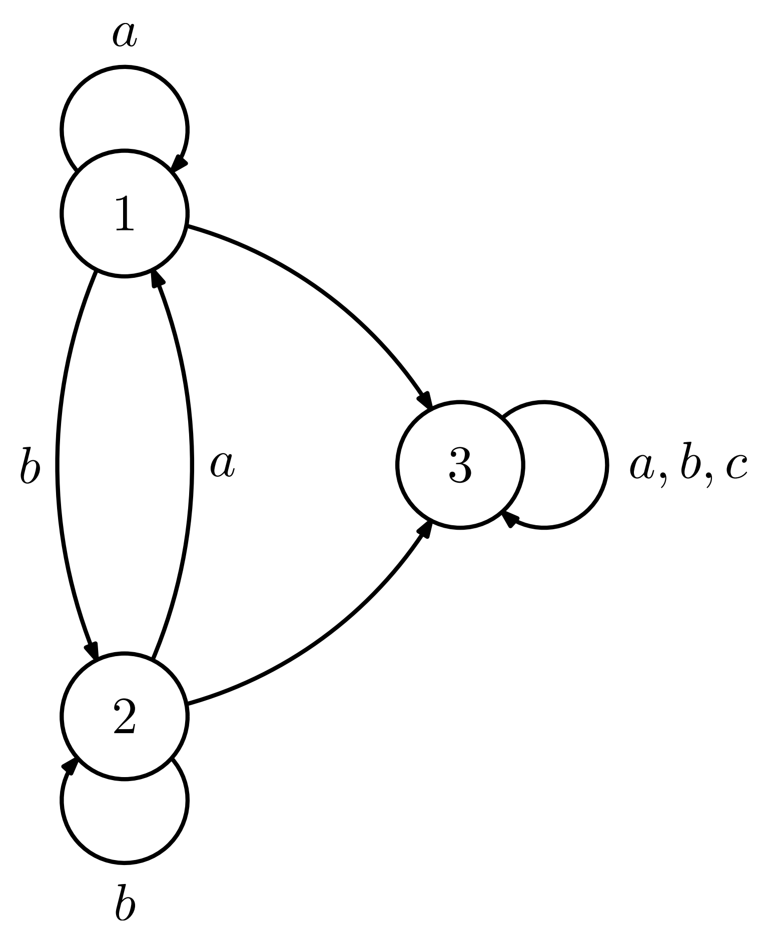
Also *a*

Actions are alike in states 1 and 2



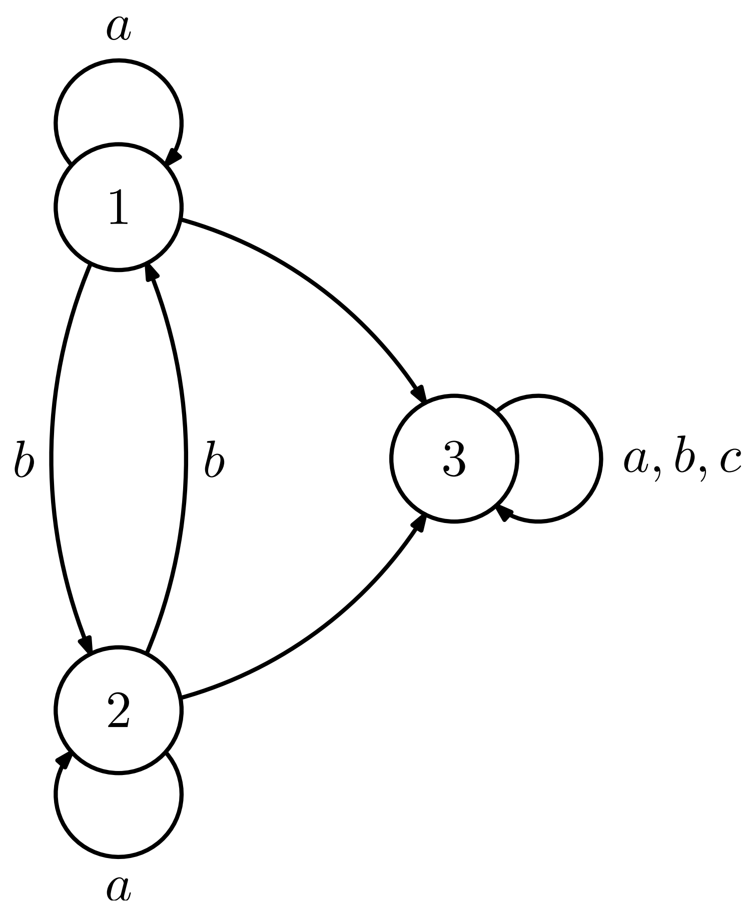
Measuring similarity

- How to compare states in MDPs?
    - Same actions have same outcomes → states are similar



Measuring similarity

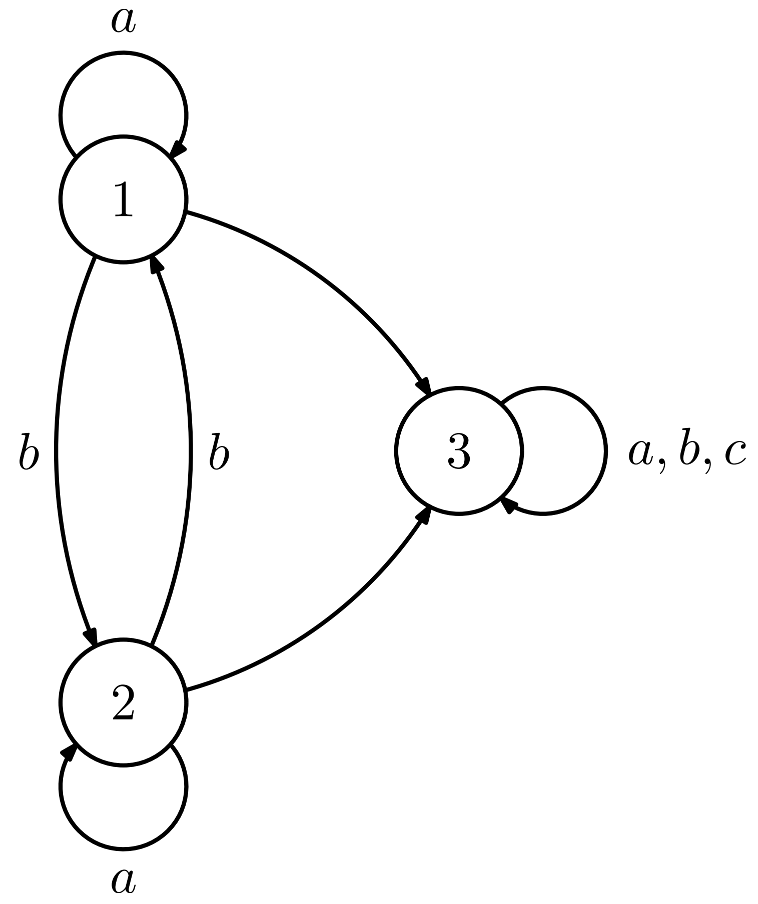
- How to compare states in MDPs?
  - Same actions have same outcomes → states are similar



What if we relabel actions?

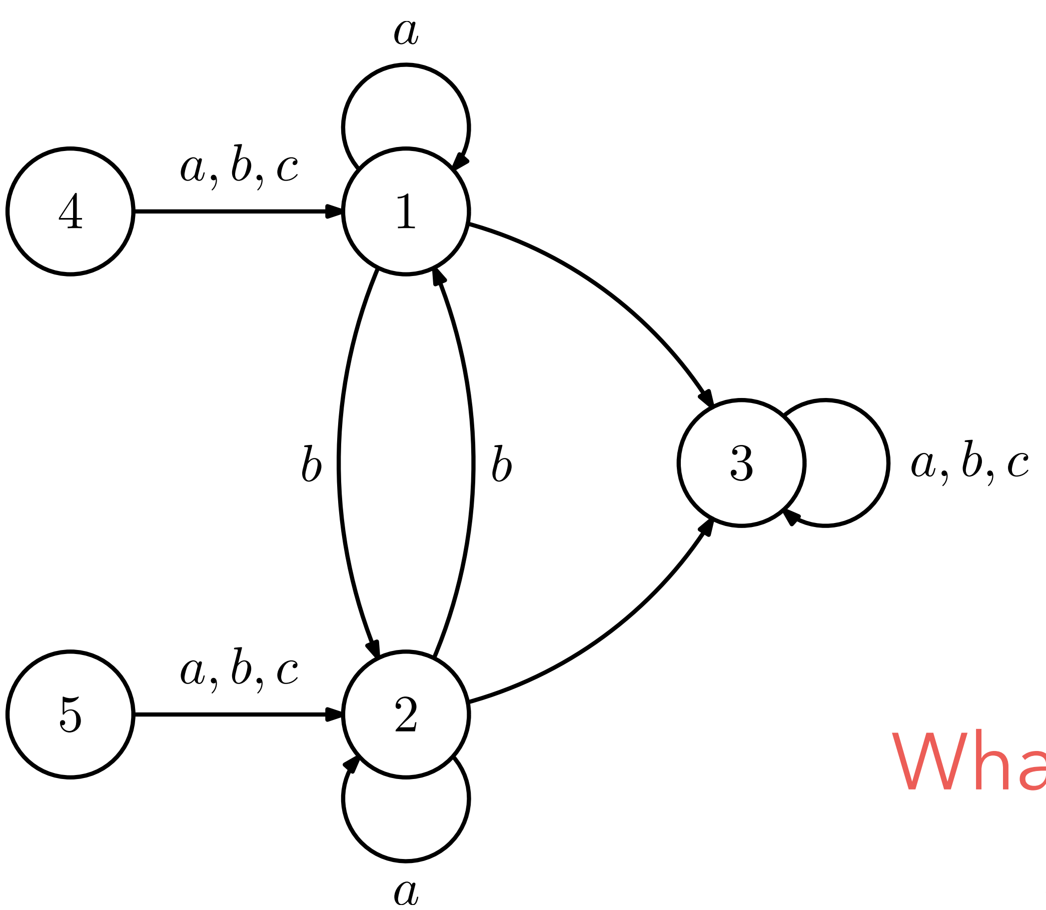
Measuring similarity

- How to compare states in MDPs?
  - Actions have similar outcomes → states are similar



Measuring similarity

- How to compare states in MDPs?
  - Actions have similar outcomes → states are similar



What if actions lead to similar states?

Measuring similarity

- How to compare states in MDPs?
  - We need some notion of “long term similarity”

↓

Bissimulation
metric

Measuring similarity

- We depart from a distance between states *d*
  - Build a distance between distributions (EMD)

More distant

Bissimulation metric

- Using the distance *d* between **distributions…**
- We define a 1-step dissimilarity between a pair (*x*, *a*) and a pair (*y*, *b*) as

$$\delta(x, a; y, b) = kd(\boxed{\mathrm{P}(\cdot \mid x, a)}, \boxed{\mathrm{P}(\cdot \mid y, b)})$$

Dissimilarity between transition probabilities in (*x*, *a*) and (*y*, *b*)

Bissimulation metric

- From δ, we remove the dependence on the actions:

$$d'(x,y) = \boxed{\max\left\{\max_{a\in\mathcal{A}}\min_{b\in\mathcal{A}}\delta(x,a;y,b), \max_{b\in\mathcal{A}}\min_{a\in\mathcal{A}}\delta(x,a;y,b)\right\}}$$

New distance between states

Dissimilarity of transition probabilities of most similar actions

Bissimulation metric

- We take this new distance, *d'*, and repeat the process, to get d''
- We repeat the process until the distance is the same in two consecutive steps

↓

Bissimulation metric

Bissimulation metric

- We can use bissimulation metric…
  - … and apply kNN to MDPs
  - … and apply SVM to MDPs
  - … etc.
- Disadvantage:
  - The bissimulation metric is computationally intensive

INVERSE

Inverse Reinforcement Learning

Going back to MDPs

- The optimal policy for an MDP can be computed as

$$
\begin{aligned}
\pi^*(x) &= \operatorname*{argmin}_{a \in \mathcal{A}} Q^*(x, a) \\
&= \operatorname*{argmin}_{a \in \mathcal{A}} \left[c(x, a) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}(y \mid x, a) J^*(y) \right]
\end{aligned}
$$

What if someone gives us the optimal policy, can we recover the task (cost)?

Going back to MDPs

- If we are given the optimal policy, then for all $x \in \mathscr{X}$ and all $a \in \mathscr{A}$,

$$J^*(x) \leq Q^*(x, a)$$

Going back to MDPs

- If we are given the optimal policy, then for all $x \in \mathscr{X}$ and all $a \in \mathscr{A}$,

$$\boldsymbol{c}_\pi + \gamma \mathsf{P}_\pi \boldsymbol{J}^* \leq \boldsymbol{c}_a + \gamma \mathsf{P}_a \boldsymbol{J}^*$$

Ignoring dependence on *a*

Going back to MDPs

- If we are given the optimal policy, then for all $x \in \mathcal{X}$ and all $a \in \mathcal{A}$,

$$\boldsymbol{c} + \gamma \mathsf{P}_\pi \boldsymbol{J}^* \leq \boldsymbol{c} + \gamma \mathsf{P}_a \boldsymbol{J}^*$$

$$\downarrow$$

$$(\mathsf{P}_\pi - \mathsf{P}_a)\boldsymbol{J}^* \leq 0$$

$$\downarrow$$

$$(\mathsf{P}_\pi - \mathsf{P}_a)(\boldsymbol{I} - \gamma \mathsf{P}_\pi)^{-1}\boldsymbol{c} \leq 0$$

$$\downarrow$$

We're computing cost from policy

Inverse Reinforcement Learning (IRL)

However…

• • •

- All policies are optimal if c ≡ 0…

$$(\mathsf{P}_\pi - \mathsf{P}_a)(\boldsymbol{I} - \gamma \mathsf{P}_\pi)^{-1}\boldsymbol{c} \leq 0$$

$$\downarrow$$

Ill-defined problem

IRL

- Original approaches select among possible solutions using heuristics
  - E.g., the difference between the value of best action and second best action is as large as possible

↓

Cumbersome

Arbitrary

Doesn't handle imperfect teachers

A probabilistic approach

- We adopt a probabilistic model for the teacher
  - We do not assume that the teacher is optimal
  - We assume that, given cost c, it selects each action *a* in state *x* with probability

Optimal Q for cost c

$$\pi_c(a \mid x) = \frac{e^{-\eta Q_c^*(x,a)}}{\sum_{a' \in \mathcal{A}} e^{-\eta Q_c^*(x,a')}}$$

Confidence on expert

A probabilistic approach

- We adopt a probabilistic model for the teacher
- Assume that the examples are independent
- Inferring the cost can then be done…
  - … using Bayesian inference (assume cost over costs):

$$\mathbb{P}\left[\mathrm{c} = c \mid \mathcal{D}\right] \propto \prod_{n=1}^{N} \pi_c(a_n \mid x_n)\mathbb{P}\left[\mathrm{c} = c\right]$$

Prior

A probabilistic approach

- We adopt a probabilistic model for the teacher
- Assume that the examples are independent
- Inferring the cost can then be done…
  - … using Bayesian inference (assume cost over costs)
  - … using simple maximum-likelihood:

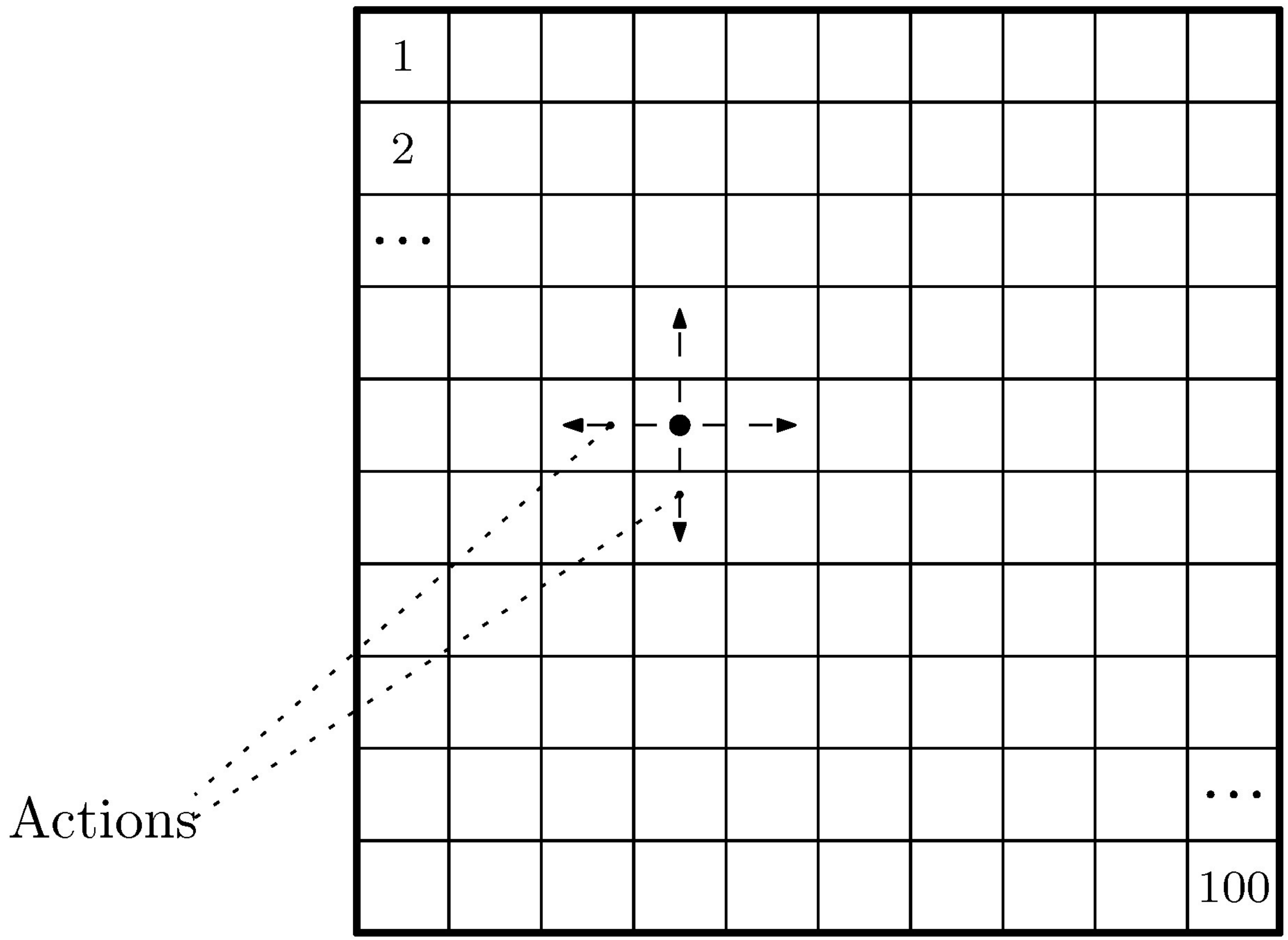
$$c^* = \operatorname*{argmax}_c \sum_{n=1}^{N} \log \pi_c(a_n \mid x_n)$$

Use gradient ascent

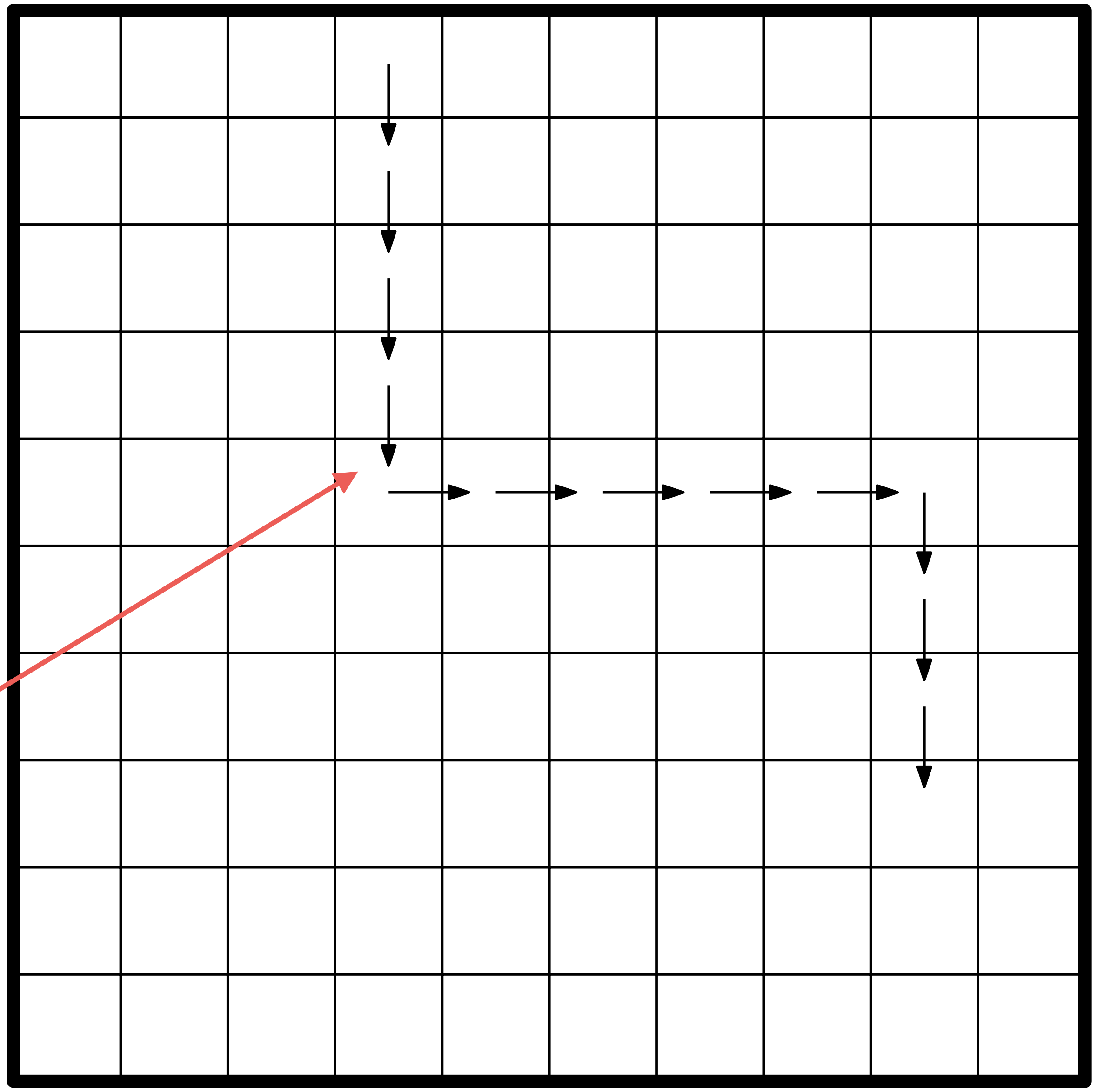
A probabilistic approach

- We adopt a probabilistic model for the teacher
- Assume that the examples are independent
- Inferring the cost can then be done…
    - … using Bayesian inference (BIRL)
    - … using simple maximum-likelihood with gradient ascent (GIRL)

Example



Example



Demonstrated actions

Example

- Learned policy moves along demonstration