

SETTING AND GOAL

ENVIRONMENT: Agent interacts with an MDP $\mathcal{M}$:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r^*, \mu, H \rangle,$$

generates trajectories $\tau := (s_1, a_1, \dots, s_H, a_H) \in \mathcal{T}$

The agent does not observe the **reward function** r^* !

PREFERENCE GENERATION: Agent shows two trajectories $\tau, \tau' \in \mathcal{T}$ to a human expert, who generates a preference via a **probabilistic model** g :

$$\mathbb{P}(\tau > \tau') = g(u^*(\tau) - u^*(\tau')),$$

where $u^*(\tau) = \sum_{h=1}^H r^*(s_h, a_h)$, with $r^*(s, a) \in [0, R_{\max}]$

COMPATIBLE REWARD SET: Set of reward functions that induce the expert's preference probability $\mathbb{P}$ under a fixed model g :

$$\mathcal{R}(\mathbb{P}, g) := \{r \in \mathcal{R} : \mathbb{P}(\tau > \tau') = g(u_r(\tau) - u_r(\tau')), \forall (\tau, \tau') \in \mathcal{T} \times \mathcal{T}\}$$

Clearly $r^* \in \mathcal{R}(\mathbb{P}, g)$ for every choice of g
Due to invariance, many other reward functions can be present

GOAL: Gather preferences from the human expert to identify a **reward function** $\hat{r}$ that lies in (an approximation of) $\mathcal{R}(\mathbb{P}, g)$

HOW TO ACHIEVE THE GOAL: The agent selects a **query set** $\hat{\mathcal{Q}} \subseteq \mathcal{T} \times \mathcal{T}$, i.e., a set of trajectory pairs to show the human expert. The agent can then recover a reward in the $\hat{\mathcal{Q}}$ -compatible reward set:

$$\mathcal{R}(\mathbb{P}, g; \hat{\mathcal{Q}}) := \{r \in \mathcal{R} : \mathbb{P}(\tau > \tau') = g(u_r(\tau) - u_r(\tau')), \forall (\tau, \tau') \in \hat{\mathcal{Q}}\}$$

A query set is called:

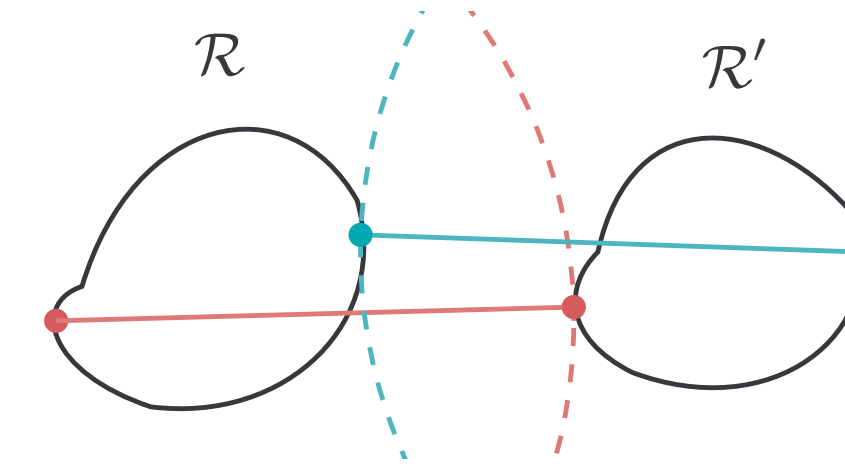
- **sufficient** if $\mathcal{R}(\mathbb{P}, g; \hat{\mathcal{Q}}) = \mathcal{R}(\mathbb{P}, g)$ (allows recovering a reward function equivalent to r^*)
- **minimal** if it is sufficient and if there exists no $\hat{\mathcal{Q}}' \subset \hat{\mathcal{Q}}$ that is sufficient
- **minimum** if it is sufficient and if there exists no $\hat{\mathcal{Q}}'$ with $|\hat{\mathcal{Q}}'| < |\hat{\mathcal{Q}}|$ that is sufficient

LEARNING PROBLEM

HAUSDORFF DISTANCE

Greatest among the shortest distances between two sets

$$\mathcal{H}(\mathcal{R}, \mathcal{R}') := \max \left\{ \sup_{r \in \mathcal{R}} \inf_{r' \in \mathcal{R}'} \|r - r'\|_{\infty}, \sup_{r' \in \mathcal{R}'} \inf_{r \in \mathcal{R}} \|r - r'\|_{\infty} \right\}$$



PAC ALGORITHM

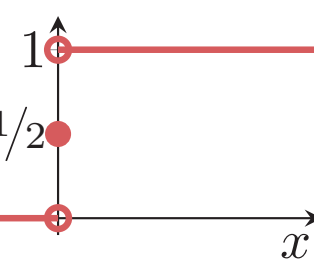
An algorithm is (ε, δ) -PAC, with $\varepsilon \geq 0$ and $\delta \in [0, 1]$, if for every problem instance $(\mathbb{P}, g)$, it generates a query set $\hat{\mathcal{Q}}$ such that:

$$\mathbb{P} \left(\mathcal{H} \left(\mathcal{R}(\mathbb{P}, g), \mathcal{R}(\mathbb{P}, g; \hat{\mathcal{Q}}) \right) > \varepsilon \right) \leq \delta$$

DETERMINISTIC PREFERENCES

PROBABILISTIC MODEL

$$g(x) = \mathbb{H}(x) := \begin{cases} 1 & x > 0 \\ 0.5 & x = 0 \\ 0 & x < 0 \end{cases}$$



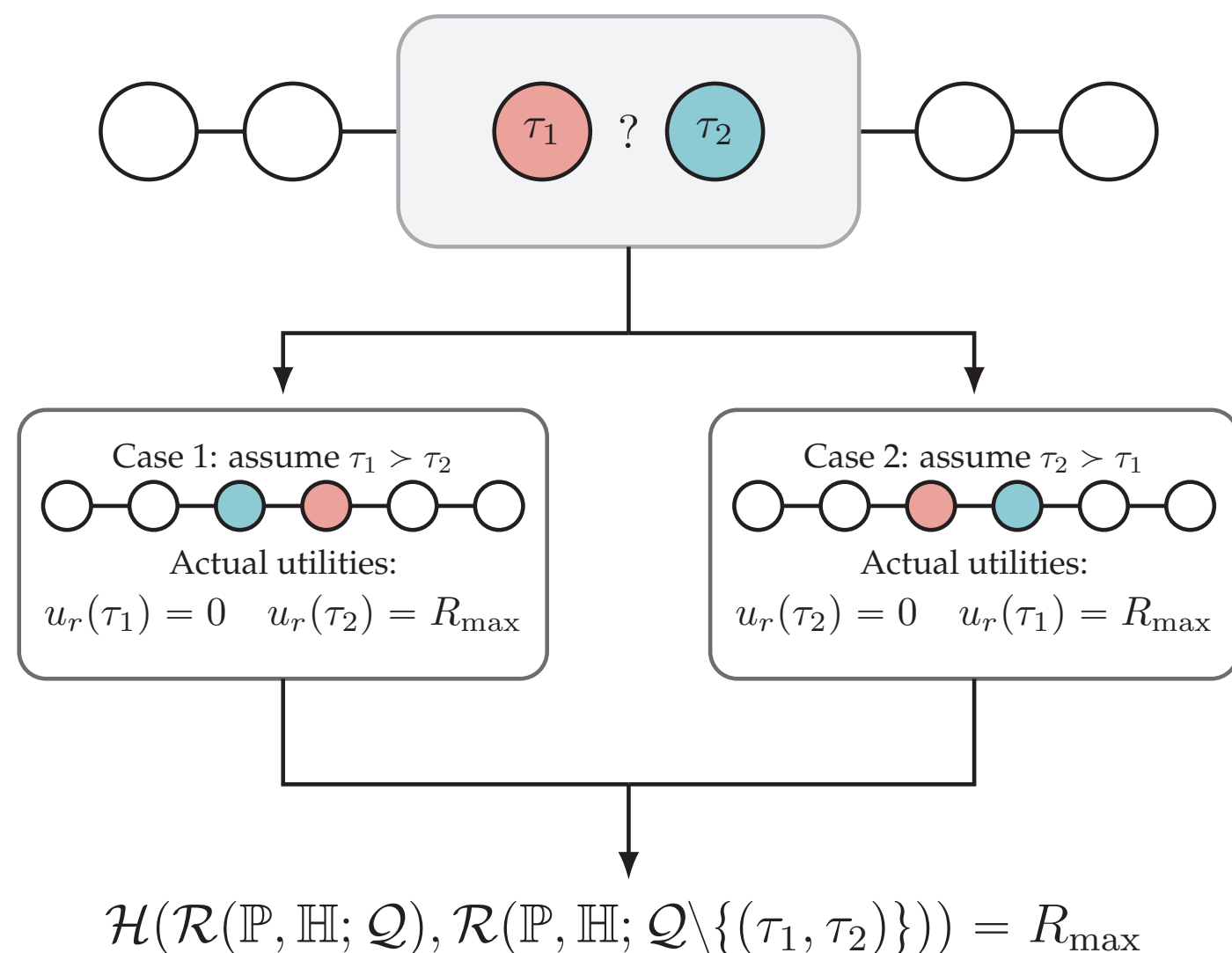
NEGATIVE RESULT

INAPPROXIMABILITY

For every **minimal** query set $\mathcal{Q}$ and $(\tau, \tau') \in \mathcal{Q}$, there exists an MDP such that:

$$\mathcal{H}(\mathcal{R}(\mathbb{P}, \mathbb{H}; \mathcal{Q}), \mathcal{R}(\mathbb{P}, \mathbb{H}; \mathcal{Q} \setminus \{(\tau, \tau')\})) = R_{\max}$$

Consider the case in which every query is labeled except for two: τ_1 and τ_2



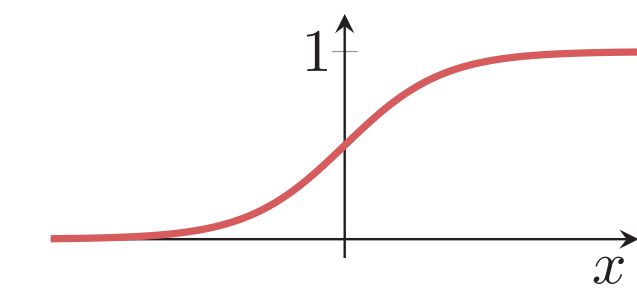
With $g = \mathbb{H}$, arbitrarily assuming an unknown label can lead to **maximum Hausdorff distance** in the worst case!

STOCHASTIC PREFERENCES

PROBABILISTIC MODEL

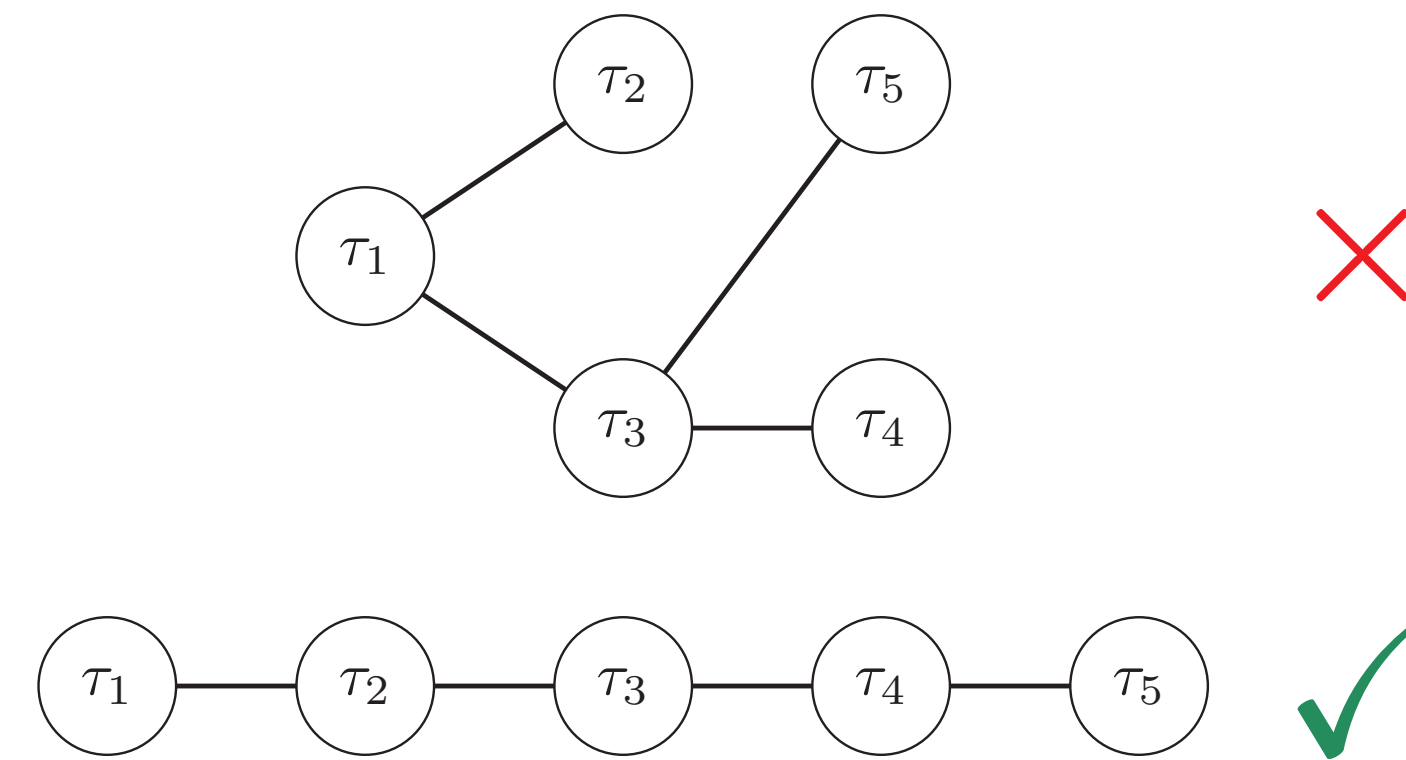
Standard
Bradley-Terry
model

$$g(x) = \sigma(x) := \frac{1}{1 + \exp(-x)}$$



QUERY CHAIN

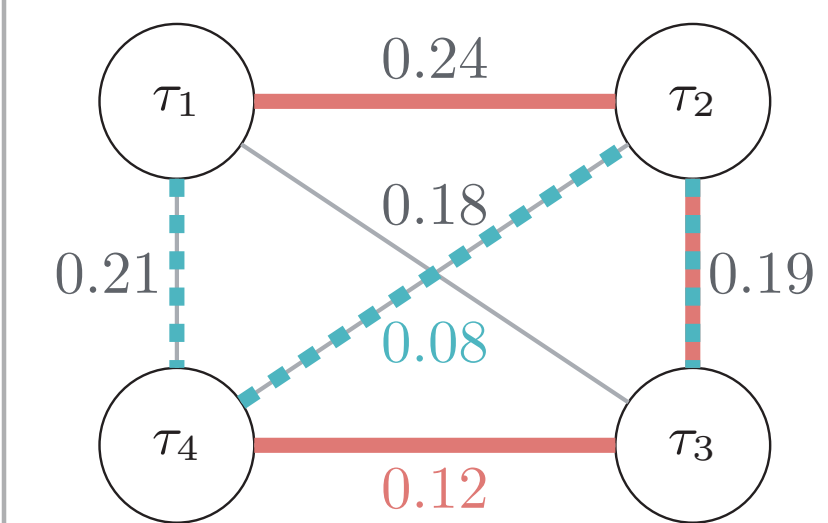
A **query chain** at stage $h \in \llbracket H \rrbracket$ is a **minimum** query set whose graph representation is a chain



QUERY CHAIN INDEX

Minimum variance of the probability distribution of the queries in a query set:

$$V(\mathcal{Q}) := \min_{(\tau, \tau') \in \mathcal{Q}} (\mathbb{P}(\tau > \tau')(1 - \mathbb{P}(\tau > \tau')))$$



BEST BASIS

$$\mathcal{Q}^* \in \arg \max V(\mathcal{Q})$$

QUERY COMPLEXITY UPPER BOUND

$\forall h \in \llbracket H \rrbracket$, λ -SQE finds a λ -optimal basis (i.e., $V(\mathcal{Q}^*) \leq (1 + \lambda)V(\mathcal{Q})$) via *successive elimination*, and finds a ε -optimal reward in:

$$\tilde{\mathcal{O}} \left(\frac{S^4 A^4 H \ln(1/\delta)}{\varepsilon^2 V(\mathcal{Q}^*)^2} \right)$$

Query complexity is minimized if basis comprises only *almost equivalent* trajectories ($u_r(\tau) - u_r(\tau') \approx 0$)!

REFERENCES

- Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 1952.
- Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.
- Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.
- R. Tyrrell Rockafellar and Roger J-B. Wets. *Variational analysis*. Springer Science & Business Media, 2009.