# Learner-aware Teaching:

## Inverse Reinforcement Learning with Preferences and Constraints

Sebastian Tschiatschek[1*], Ahana Ghosh[2*], Luis Haug[3*], Rati Devidze[2], Adish Singla[2]

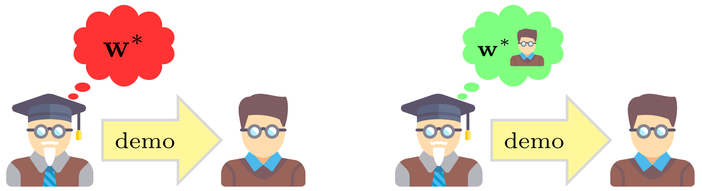
[1]Microsoft Research, [2]ETH Zurich, [3]Max Planck Institute for Software Systems, [*]equal contribution

## Highlights

**Problem setting**

- Teaching a *learner with preferences* via demonstrations
- Studied two teaching strategies

*Learner-agnostic teaching* | *Learner-aware teaching*




Ignoring learner's preferences | Considering learner's preferences

**Main results**

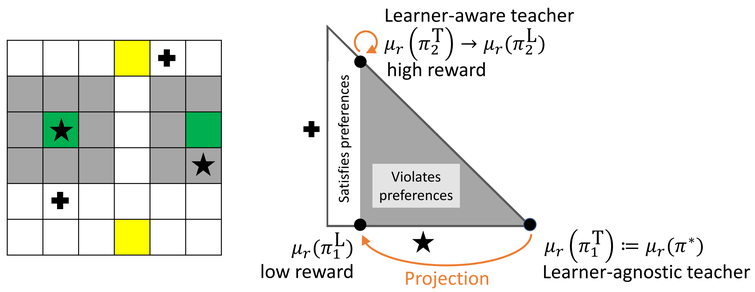
- Learner-agnostic teaching can be arbitrarily bad
- New algorithms for learner-aware teaching achieving high performance

## A Teacher and an IRL Learner Without Preferences

- MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma, R)$ with rewards $R(s) = \langle w^*, \phi_r(s) \rangle$
- Teacher T provides demonstrations using policy $\pi^{\mathsf{T}}$ in $\mathcal{M}$
- Policy $\pi$ has reward $R(\pi) = \langle w^*, \mu_r(\pi) \rangle$, where $\mu_r(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi_r(s) \mid \pi\right]$
- Learner receives demonstrations and outputs $\pi^{\mathsf{L}}$ s.t. $\|\mu_r(\pi^{\mathsf{L}}) - \mu_r(\pi^{\mathsf{T}})\| \leqslant \epsilon$
- This ensures that $R(\pi^{\mathsf{L}}) \geqslant R(\pi^{\mathsf{T}}) - \epsilon$

## Challenges in Teaching Learners With Preferences

- Object-world gathering game
- $\star$ yields reward $1.0$, $+$ yields reward $0.9$
- Learner's preference: Avoid frequent proximity of green cells ($\leqslant$ 1-cell distance)




- *Providing demonstrations from optimal behavioral policy $\pi^*$ can lead to arbitrarily bad learner's performance!*

## Learner Models

- Learner's preferences are captured by features $\phi_c(s)$
- Formalized as constraints on $\mu_c(\pi)$, where $\mu_c(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi_c(s) \mid \pi\right]$

**Standard maximum causal entropy IRL learner**

$$\begin{aligned} \max_{\pi} \quad & H(\pi) && \textit{causal entropy} \\ \text{s.t.} \quad & \|\mu_r(\pi) - \mu_r(\pi^{\mathsf{T}})\| = 0 && \textit{feature matching} \end{aligned}$$

**Learner with hard preferences**

$$\begin{aligned} \min_{\pi} \quad & \|\mu_r(\pi) - \mu_r(\pi^{\mathsf{T}})\| && \textit{feature matching} \\ \text{s.t.} \quad & g(\mu_c(\pi)) \leqslant 0 && \textit{hard preferences} \end{aligned}$$

**Learner with soft preferences**

$$\begin{aligned} \max_{\pi, \delta_r^{\text{soft}}, \delta_c^{\text{soft}}} \quad & H(\pi) - C_r \|\delta_r^{\text{soft}}\|_p - C_c \|\delta_c^{\text{soft}}\|_p \\ \text{s.t.} \quad & \|\mu_r(\pi) - \mu_r(\pi^{\mathsf{T}})\| \leqslant \delta_r^{\text{soft}} && \textit{feature matching} \\ & g(\mu_c(\pi)) \leqslant \delta_c^{\text{hard}} + \delta_c^{\text{soft}} && \textit{hard+soft preferences} \end{aligned}$$

- Learner trades-off reward-feature matching and its own preferences

## Learner-Aware Teaching for Known Constraints

**Learner-aware teaching for hard preferences: AWARE-CMDP**

- Define a set of feasible reward feature expectations $\Omega_r^{\mathsf{L}} = \{\mu_r(\pi) \mid g(\mu_c(\pi)) \leqslant 0\}$
- Optimal teaching policy = solution of constrained MDP:

$$\max_{\pi^{\mathsf{T}}} \ \langle w^*, \mu_r(\pi^{\mathsf{T}}) \rangle \ \text{ s.t. } \ \mu_r(\pi^{\mathsf{T}}) \in \Omega_r^{\mathsf{L}}$$

- *Theorem.* The value of learner-aware teaching can be arbitrarily high, given by

$$\max_{\pi \text{ s.t. } \mu_r(\pi) \in \Omega_r^{\mathsf{L}}} \langle w^*, \mu_r(\pi) \rangle - \langle w^*, \text{Proj}_{\Omega_r^{\mathsf{L}}}(\mu_r(\pi^*)) \rangle$$

- For linear $g(\cdot)$, the above problem can be solved via linear programming

**Learner-aware teaching for soft preferences: AWARE-BIL**

- Optimal teaching problem can be formulated as a bi-level optimization:

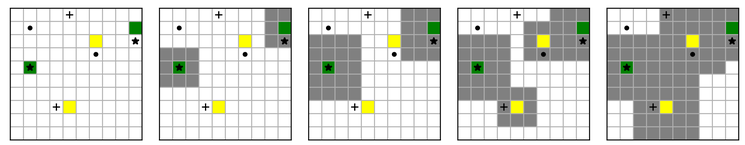
$$\max_{\pi^{\mathsf{T}}} \ \langle w^*, \mu_r(\pi^{\mathsf{L}}) \rangle \quad \text{s.t. } \ \pi^{\mathsf{L}} \in \arg\max_{\pi} \text{IRL}(\pi, \mu(\pi^{\mathsf{T}}))$$

- Here $\text{IRL}(\pi, \mu(\pi^{\mathsf{T}}))$ stands for the IRL problem solved by the learner
- Optimal teaching policy is a softmax policy satisfying the learner's constraints
- A challenging non-convex optimization problem
- Proposed a gradient-based optimization approach

## Experimental Results

**Experimental setup**

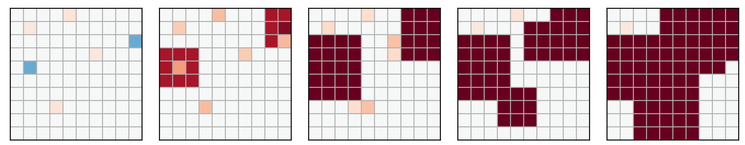
- Object-world gathering environment:
  - Rewards: $\star$ yields $1.0$, $+$ yields $0.9$, $\bullet$ yields $0.2$
  - Two "green" distractors at 0-cell and 1-cell distance to the $\star$ objects
  - Two "yellow" distractors at 1-cell and 2-cell distance to the $+$ objects
  - Discount factor $\gamma = 0.99$
- Learners with soft preferences ($C_r = 5$, $C_c = 10$) and $\delta_c^{\text{hard}} = 0$
- Environment and learners' preferences for 5 different learners L1, …, L5



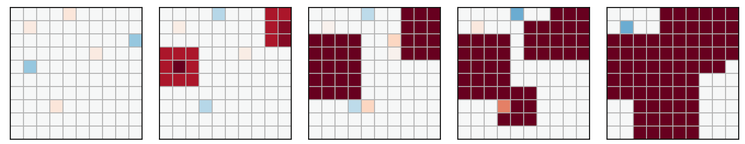
L1 | L2 | L3 | L4 | L5

- For instance, L2 has two preference features indicating whether there is a green cell at a distance of 0-cells or 1-cell, respectively

**Learner-aware teaching for known constraints**



Learners' rewards inferred from learner-agnostic teacher (AGNOSTIC)



Learners' rewards inferred from learner-aware teacher (AWARE-BIL)

| Teacher | L1 | L2 | L3 | L4 | L5 |
|---|---|---|---|---|---|
| AGNOSTIC | $7.99 \pm 0.02$ | $0.01 \pm 0.00$ | $0.01 \pm 0.00$ | $0.01 \pm 0.00$ | $0.00 \pm 0.00$ |
| AWARE-BIL | $8.00 \pm 0.02$ | $7.20 \pm 0.01$ | $4.86 \pm 0.30$ | $3.15 \pm 0.27$ | $1.30 \pm 0.07$ |

## Further Results

- Algorithms for learner-aware teaching with unknown constraints
- Additional experimental results
- Formal statements, proofs, and derivations

Microsoft | MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS | ETH zürich