Do we use the Right Measure?
Challenges in Evaluating Reward Learning Algorithms
–Supplementary Material–

Nils Wilde
Department of Cognitive Robotics
Delft University of Technology Netherlands
n.wilde@tudelft.nl

Javier Alonso-Mora
Department of Cognitive Robotics
Delft University of Technology Netherlands
n.wilde@tudelft.nl

Keywords: Human Robot Interaction, Reward Learning

A Proofs

Theorem 1 (Unbounded Reward Difference)**.** Let $\boldsymbol{w}^{\text{user}}$ be a user weight, and $\boldsymbol{w}'$ be an estimate, where the alignment is $\delta \leq \alpha(\boldsymbol{w}', \boldsymbol{w}^{\text{user}}) < 1$ for some $\delta < 1$. The difference in reward $R(\mathcal{T}^{\text{user}}, \boldsymbol{w}^{\text{user}}) - R(\mathcal{T}', \boldsymbol{w}^{\text{user}})$ is unbounded.

Proof. We consider a discrete planning problem with two features. Let there be only two solutions $\mathcal{T}^A$ and $\mathcal{T}^B$ with features $\boldsymbol{\phi}(\mathcal{T}^A) = [-1\ -M]$ and $\boldsymbol{\phi}(\mathcal{T}^B) = [-N\ 0]$ where $M > 0$ and $N > 1$. Further, let the user weights be $\boldsymbol{w}^{\text{user}} = [1\ 0]$, and the estimate $\boldsymbol{w}' = [1\ \epsilon]$, where $\epsilon > 0$. We notice that $\alpha(\boldsymbol{w}', \boldsymbol{w}^{\text{user}}) \geq \delta$ as $\epsilon \to 0$ for any $\delta < 1$, that is, the alignment becomes arbitrarily close to 1 for small ϵ.

First, we calculate $\boldsymbol{w}^{\text{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^A) = -1$ and $\boldsymbol{w}^{\text{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^B) = -N$. Since $N > 1$ the trajectory $\mathcal{T}^A$ collects the higher reward and hence is the optimal solution for $\boldsymbol{w}^{\text{user}}$. Further, we have $\boldsymbol{w}' \cdot \boldsymbol{\phi}(\mathcal{T}^A) = -1 - M\epsilon$ and $\boldsymbol{w}' \cdot \boldsymbol{\phi}(\mathcal{T}^B) = -N$. We construct the case where $\mathcal{T}^B$ is the optimal solution for the estimate $\boldsymbol{w}'$, i.e., the estimated weights result in a different, suboptimal trajectory:

$$-1 - M\epsilon < -N. \tag{1}$$

The difference in reward is $R(\mathcal{T}^{\text{user}}, \boldsymbol{w}^{\text{user}}) - R(\mathcal{T}', \boldsymbol{w}^{\text{user}}) = \boldsymbol{w}^{\text{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^A) - \boldsymbol{w}^{\text{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^B) = -1 + N$. This is unbounded if we can pick an arbitrarily large N such that (1) is satisfied as $\epsilon \to 0$. Choosing $M = {N^2}/{\epsilon}$ simplifies (1) to $N^2 > N - 1$ which is satisfied for any $N > 1$. Hence, N has no upper bound, making the reward difference unbounded. $\square$

Theorem 1 shows that even when the alignment is arbitrarily close to 1, it does not allow for claims on how much reward is collected, compared to optimal. Following the same proof the result extends to the second parameter-based measure, the MSE. Next we study the multi-scenario case to show that the reward-based measures do not translate from test to training scenarios. We use a similar construction as for Theorem 1 to establish an analogous result. Let $I^{\texttt{Train}}$ be a training and $I^{\texttt{Test}}$ a test instance. For weights $\boldsymbol{w}'$, $\boldsymbol{w}^{\text{user}}$, we use $R^{\texttt{rel}}_{\texttt{train}}(\boldsymbol{w}', \boldsymbol{w}^{\text{user}})$ to denote the relative reward collected in $I^{\texttt{Train}}$ and similarly $R^{\texttt{rel}}_{\texttt{test}}(\boldsymbol{w}', \boldsymbol{w}^{\text{user}})$ for the test instance $I^{\texttt{Test}}$.

Theorem 2 (Unbounded Test Error)**.** Let $\boldsymbol{w}^{\text{user}}$ be a user weight, and $\boldsymbol{w}'$ be an estimate. Further, let $I^{\texttt{Train}}$ be a training instance, where the relative reward $R^{\texttt{rel}}_{\texttt{train}}(\boldsymbol{w}', \boldsymbol{w}^{\text{user}})$ is taking values in $[\delta, 1]$ for some $\delta < 1$. There exist test instances $I^{\texttt{Test}}$ where the relative reward $R^{\texttt{rel}}_{\texttt{test}}(\boldsymbol{w}', \boldsymbol{w}^{\text{user}})$ has no tighter lower bound than 0.

Proof. Again we consider a discrete planning problem with two features. We construct a training instance $I^{\texttt{Train}}$ with only two solutions $\mathcal{T}^A$ and $\mathcal{T}^B$ with features $\boldsymbol{\phi}(\mathcal{T}^A) = [-1\ -2]$ and $\boldsymbol{\phi}(\mathcal{T}^B) = [-5\ 0]$. Further, let the user weights be $\boldsymbol{w}^{\text{user}} = [1\ 1]$, and the estimate $\boldsymbol{w}' = [1\ 0]$. For both weights,

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

$\mathcal{T}^A$ achieves the higher reward, i.e., is the respective optimal solution. Thus, we have relative reward $R^{\texttt{rel}}_{\texttt{train}}(\boldsymbol{w}', \boldsymbol{w}^{\texttt{user}}) = 1$, the estimated weights yields an optimal solution.

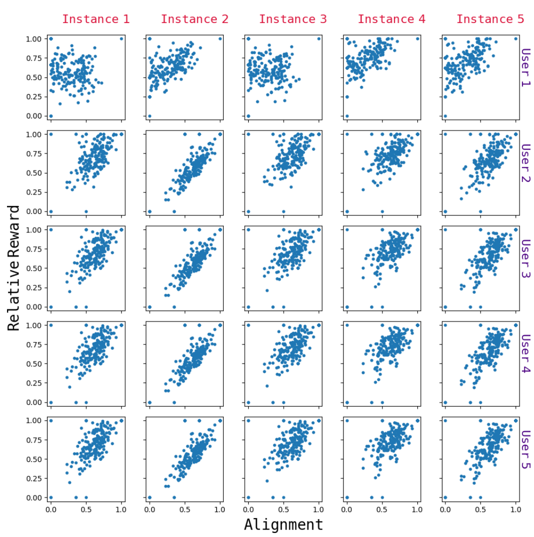
Now, consider the test instance $I^{\texttt{Test}}$ where we again have only two solutions $\mathcal{T}^C$ and $\mathcal{T}^D$ with features $\boldsymbol{\phi}(\mathcal{T}^C) = [-2\ 0]$ and $\boldsymbol{\phi}(\mathcal{T}^D) = [-1\ \ -N]$ for some $N > 1$. Thus, $\boldsymbol{w}' \cdot \boldsymbol{\phi}(\mathcal{T}^C) = -2$ and $\boldsymbol{w}' \cdot \boldsymbol{\phi}(\mathcal{T}^D) = -1$; the solution $\mathcal{T}^D$ is optimal for weights $\boldsymbol{w}'$. Further, $\boldsymbol{w}^{\texttt{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^C) = -2$ and $\boldsymbol{w}^{\texttt{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^D) = -1 - N$, implying that $\mathcal{T}^C$ is always the optimal solution for weights $\boldsymbol{w}^{\texttt{user}}$. Hence, the relative reward is

$$R^{\texttt{rel}}_{\texttt{test}}(\boldsymbol{w}', \boldsymbol{w}^{\texttt{user}}) = \frac{\boldsymbol{w}^{\texttt{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^C)}{\boldsymbol{w}^{\texttt{user}} \cdot \boldsymbol{\phi}(\mathcal{T}^D)} = \frac{2}{1+N}.$$

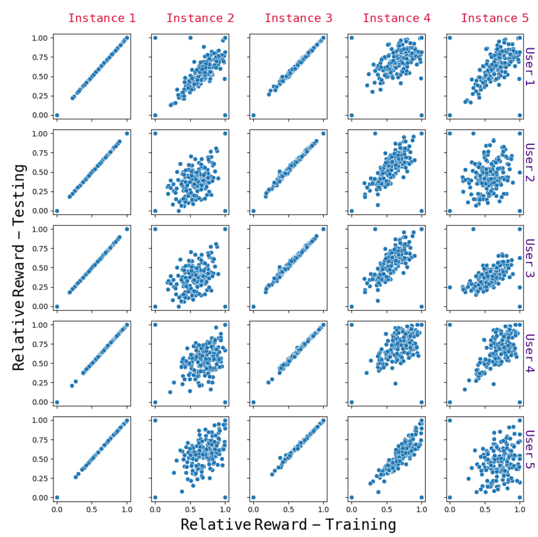
By picking an arbitrarily large N the relative reward in the test scenario can become arbitrarily small, implying that no lower bound greater than 0 exists. □

B Additional Simulation Results

Figure 1 provides example plots for the numerical experiments with the Server task, showing alignment against relative reward, as well as relative reward in training against testing.



(a) Alignment-reward relationship.



(b) Training-Testing Relationship.

Figure 1: Examples for Server task.