Relatively Rational: Learning Utilities and Rationalities Jointly from Pairwise Preferences

Taku Yamagata [1] **Tobias Oberkofler** [2] **Timo Kaufmann** [2 3] **Viktor Bengs** [2 3] **Eyke Hüllermeier** [2 3]
Raúl Santos-Rodríguez [1]

Abstract

Learning utilities from preference feedback has become increasingly important, particularly in fine-tuning language models such as ChatGPT. Traditional methods often assume equal rationality among labellers, leading to inaccurate utility estimates. We propose an algorithm that jointly estimates trainer rationality and item utilities to enhance utility learning and gain additional insights from feedback. Our approach focuses on settings where feedback is received from multiple trainers, using the Boltzmann-rational model to relate choices to latent utilities while accounting for varying levels of rationality. Given shared utilities, our method identifies rationality ratios among trainers from observed choices without extra calibration data or assumptions. We analyse the theoretical impact of assuming equal rationality on utility accuracy and empirically show superior performance in an action-advice setting, where agents construct policies using the learned utilities as rewards. By accurately modelling trainer rationality, we can enhance high-quality feedback collection, potentially leading to better-aligned models and an improved understanding of human preferences.

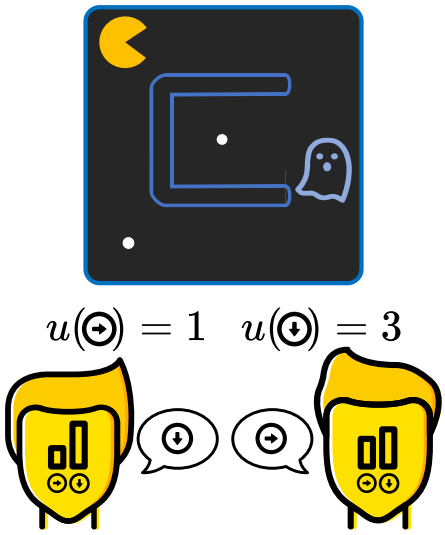

Figure 1. An illustration of the setting where multiple labellers provide preference feedback on a set of items. Both trainers share the same latent utilities, in this case, a higher utility for the action of moving down, but differ in their rationality coefficients. Consequently, they sample choices from different distributions, potentially leading to different observed choices. By jointly estimating the rationality coefficients and the utilities, we can improve utility learning and gain insights from the feedback.

1. Introduction

Learning utilities from preference feedback is a common machine learning problem, recently popularised through its application to fine-tuning language models such as ChatGPT (OpenAI, 2022): Given a preference for one alternative (behaviour) over another, the goal is to learn the utility of each option, which can then be used as an objective in approaches like reinforcement learning from human feedback (RLHF) (Kaufmann et al., 2023b). In practice, utilities are often learned from multiple trainers, each with varying levels of expertise, understanding of the task, and attention to detail. In this paper, we propose an algorithm that estimates the *rationality* of human trainers jointly with the utilities of the items, aiming to improve the accuracy of utility learning and gain additional insights from feedback.

This setting of receiving feedback from multiple labellers is common in practice, e.g. in crowdsourcing settings. Prior works often assume the labellers are equally rational, which is unrealistic. The goal may then be to learn individualised preference models or, assuming shared utilities, to leverage

[1]Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom [2]Institute for Informatics, LMU Munich, Munich, Germany [3]Munich Center for Machine Learning, Munich, Germany. Correspondence to: Taku Yamagata <taku.yamagata@bristol.ac.uk>.

The ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria. 2024.

all feedback to train a single model. In this paper, we focus on the latter setting.

Learning utilities from observed choices requires a model of how the choices are made. The model should relate the choices to utilities, i.e. latent values that represent the intrinsic desirability of the items, while also accounting for other influences on the choices, such as the rationality of the labellers. This is commonly formalised with the Boltzmann-rational model (e.g, Jeon et al., 2020), which models the probability of choosing an item c_i from a choice set $C = \{c_1, \dots, c_N\}$ as

$$P(c_i) = \frac{\exp(\beta u_i)}{\sum_{j=1}^{N} \exp(\beta u_j)},$$

where u_i is the latent utility of choice c_i and β is the labeller-specific *rationality coefficient*. Higher rationality leads to a more deterministic choice, while lower rationality leads to a more random choice.

The Boltzmann-rational model is insensitive toward (additive) shifts in utilities, making it impossible to identify all parameters, i.e. utilities and rationality, purely from observed choices (Train, 2009). The key insight of this paper is that, given feedback from multiple labellers with shared utilities, we can at least identify the *ratio of their rationalities*: If the utilities are shared, the observed differences in choices can be attributed to differences in rationality.

Without requiring additional data or assumptions on the feedback process, this enables us to identify the rationality of each labeller relative to each other in addition to learning the utilities. This rationality coefficient can then be used for downstream tasks, such as identifying labellers giving low-effort responses (*satisficing*) (Kaufmann et al., 2023a) or adaptively choosing a feedback modality (Ghosal et al., 2023). Through these means, we hope to simplify the process of collecting large amounts of high-quality feedback, leading to an enhanced understanding of human preferences and enabling us to train models more aligned with them.

Our contributions are the following:

1. Propose an algorithm that jointly learns utilities and rationality coefficients from pairwise preference feedback based on the joint likelihood of the utilities and the rationalities.

2. Theoretically analyse the impact of falsely assuming equal rationality among multiple labellers.

3. Empirically investigate utility learning and resulting downstream task performance in an action-advice setting where the agent learns a policy using the learned utilities as rewards.

Note that although RLHF is an exciting area of application with high practical relevance for our approach, our evaluations are in simpler ranking and action advice settings for now, and we leave the application to RLHF for future work.

2. Related Work

Our approach is focused on jointly learning shared utilities and labeler-specific rationality coefficients. Prior work has studied rationality learning both in isolation as well as RLHF-specific contexts, which we review in the following.

Learning Rationality There exists a long history in classical choice modelling literature of estimating rationalities of different labellers. To that extent Ben-Akiva & Morikawa (1990) and Swait & Louviere (1993) present algorithms to identify the *ratio* of the rationalities of labellers with equal underlying discrete choice models. These approaches however are not trivial to apply outside the discrete choice setting, while our method can readily be employed in the continuous domain.

The general problem of learning from noisy labels in non-reinforcement learning settings has, among others, been addressed by Dawid & Skene (1979) and Raykar et al. (2010) using the expectation maximisation algorithm (Dempster et al., 1977). More recently, learning from noisy labels has also been used in the context of supervised learning with promising results (Whitehill et al., 2009). Such approaches are also applicable in the RL setting as demonstrated by Yamagata et al. (2021), utilising binary feedback from multiple humans. However, such methods rely on global labels as feedback and do not naturally extend themselves to the setting with pairwise comparison data. Furthermore, these approaches are not easily generalizable beyond the discrete choice setting as well.

Rationality for RLHF Utilising human feedback has proven to be an effective strategy in reinforcement learning (RL) (Griffith et al., 2013; Knox & Stone, 2012). Learning from preference data has been a particularly successful paradigm in recent years (OpenAI, 2022; Kaufmann et al., 2023b). While our evaluations are in an action-advice setting, it is closely connected to these approaches of learning reward functions from human feedback and is directly applicable to that setting as well. Daniels-Koch & Freedman (2022) demonstrate query-based rationality learning using teacher selection, demonstrating the benefits of a rationality-adapted approach. Ghosal et al. (2023) have shown that estimating rationality coefficients of labellers can favourably influence reward learning. Their algorithm, however, requires costly (and potentially unobtainable) calibration data. Metz et al. (2023) shows that an initial calibration phase might still be insufficient to cope with the various influ-

ences on a labellers rationality and proposes to use multiple calibration phases continuously throughout an experiment. Within this work, we strive to estimate the rationality coefficients jointly with the utilities, completely eliminating the need for (ground truth) calibration data.

Freedman et al. (2023) outline an algorithm for active teacher selection for RLHF, in which they show that when the ground truth preference probabilities P are known and each labeller was given the same pair of objects (i, j), the rationalities can be analytically computed up to a scale factor $a = -\Delta_{ij}^{-1}$, which is the inverse of the utility difference between two objects: $\beta_m = a \cdot \ln\left(\frac{1}{P} - 1\right)$. They however do not address scenarios in which either no ground truth preference probabilities are known or not every labeller was presented the same object pair.

3. Method

We consider a scenario in which an RL agent receives pairwise preference feedback and learns a policy from it. There are several ways of incorporating such feedback in an RL setting. The most popular approach is comparing two trajectories (sequence of state and action pairs), where a trainer indicates which trajectory he believes to be better. The feedback is then used to train a reward model (Christiano et al., 2017) (or a policy directly (Rafailov et al., 2024)).

In this paper, we focus on the setting of comparing two actions for a given state. To produce such pairwise action comparisons, we asked trainers if the agent's action was correct and, if not, asked them to suggest a better action. These responses can then be converted to preference feedback. If the trainer expressed the agent's current action was right, we would generate multiple preferences between the agent's action and all the other actions by saying the agent's action was preferable. On the other hand, if the trainer suggested another action, we would generate a single preference pair between the trainer's suggested action and the agent's current action. It is worth noting that our method does not require trainers to give feedback at every time-step. Instead, they decide themselves when to give their feedback.

Other algorithms exist to elicit the preference feedback. However, our approach is straightforward yet effective as it obtains feedback on the agent's choice of action as right as it happens. Furthermore, the trainer does not necessarily need to suggest the best action, but just a *better* action, which alleviates the workload.

Human judgment tuples (a_1, a_2, s, k, w) are recorded in a database $\mathcal{D}$. Here, a_1 and a_2 are the pair of actions to compare, s is the state, k is the trainer's index, and w is an encoding of the judgement, $w = (w_1, w_2) = [1, 0]$ if the human selects a_1 and $w = [0, 1]$ if the human selects a_2.

3.1. Joint Learning of Utilities and Relative Rationalities

We use the Boltzmann-rational model to represent human preference feedback behaviour. The probability that trainer k indicates action a_1 is better than action a_2 in state s is given by

$$P(a_1 \succ_{k,s} a_2) = \frac{1}{1 + e^{-\tilde{\beta}(u_{s,a_1} - u_{s,a_2})}}. \tag{1}$$

Here, two groups of parameters need to be estimated:

$$\mathbf{u} = \{u_{s,a} \in \mathbb{R}\}_{s \in \mathcal{S}, a \in \mathcal{A}}$$

are utility parameters for each state/action pair (s, a), where $\mathcal{S}$ and $\mathcal{A}$ denote the state space and action space, respectively. These parameters indicate how good the action a is in state s, with higher numbers indicating a higher value. The second group,

$$\boldsymbol{\beta} = \{\tilde{\beta} \in \mathbb{R}\}_{k=1,\ldots,K},$$

consists of rationality parameters for the trainers $k \in \{1, \ldots, K\}$. These parameters indicate how rational the different trainers are. Trainers with a higher rationality parameter are more accurate in their feedback.

We employ the following negative log-likelihood function as a loss function and apply the stochastic gradient descent (SGD) algorithm to learn the utility and rationality parameters concurrently:

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\beta}) = -\sum_{(a_1,a_2,s,k,w) \in \mathcal{D}} w_1 \log P(a_1 \succ_{k,s} a_2) + w_2 \log P(a_2 \succ_{k,s} a_1)$$

From Eq. 1, it is evident that the probability depends only on the product of the rationality value and the difference of the utility values. Thus, the loss function remains unchanged when adding a constant to all utility values or when multiplying the rationality values by a constant and dividing the utilities by the same constant.

In order to stabilise the learning process, we implement two constraints. First, we fix the rationality for one of the trainers to one. Here, every trainer may, in principle, serve as a reference, except one that labels completely at random (and hence has rationality $\beta = 0$). In practice, such a selection could be facilitated by reviewing a labeller's history and ensuring to include at least one labeller whom we expect to behave reasonably rationally.

Secondly, we introduce a prior for the utilities. A range of priors have been proposed in the literature (Davidson & Solomon, 1973; Whelan, 2017). Here, we adopt a logistic prior, which can be introduced by having two preferences for each state-action pair against a virtual utility parameter

of zero value (Newman, 2023). As a result, the following regularisation term ($\mathcal{L}_r$) is added to the above loss function:

$$\mathcal{L}_r = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \log(1+e^{-u_{s,a}}) + \log(1+e^{u_{s,a}})$$

3.2. Deriving a Policy from Utilities

To decide which action to take in a given state, we employ greedy action selection based on the learned utility values, i.e. select action

$$a = \arg\max_{a\in\mathcal{A}} u_{s,a}\,.$$

Note that this formulation does not encourage exploration. We rely on the human trainer's feedback to guide the agent toward optimal actions instead. This amounts to learning the optimal action-value function Q^* in a reinforcement learning setting, requiring labellers to be aware of each action's future implications. This is in contrast to the RLHF setting in which we generally attempt to learn the (dynamics-independent) reward function and rely on a reinforcement learning algorithm to explore the environment dynamics. Extending our approach to the RLHF setting, which has reduced requirements on the human labeller, would be an exciting area for future work.

4. Information Theoretic Analysis

We conducted information-theoretic analyses based on mutual information to show the benefit of considering the trainers' different rationalities. Mutual information (MI) is a concept that quantifies the mutual dependencies between two random variables. It measures how much information one random variable provides about another.

Let i and j be the indices of two items and k the index of a trainer. Now consider the MI between a true ranking on a pair of items ($x_{i,j}$), i.e., a binary variable indicating whether or not the utility of item i truly is larger than the utility of item j, and an observed choice from the trainer ($y_{i,j}^{(k)}$). The true utility values of the two unknown objects, u_i and u_j can be seen as equally distributed random variables. The underlying true ranking on a pair (i,j) is, therefore, a derived random variable taking values 0 and 1 with equal probability, as the difference between the two equally distributed utilities is symmetric around 0. Similarly, the preference feedback $y_{i,j}^{(k)}$ can be seen as such a binary random variable as well.

We can model the relationship between these random variables with a binary symmetric channel (BSC) using the Boltzmann-rational model to compute the transition probabilities. Figure 2 shows the BSC model where $P_{i,j}^{(k)}$ denote a probability of observing a preference matching the true ranking (feedback accuracy) for the item i and j from trainer k. The feedback accuracy can be derived based on the Boltzmann-rational model with an assumed rationality parameter $\tilde{\beta}$ and the utility values as follows:

$$P_{i,j}^{(k)} = \frac{1}{1+e^{-\tilde{\beta}|u_j-u_i|}}.$$

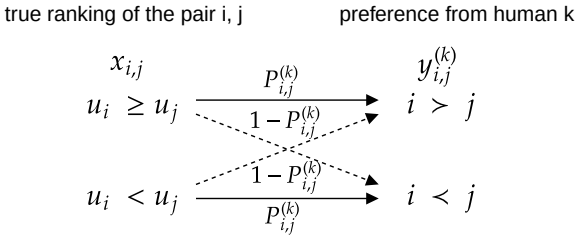

Figure 2. An illustration of the binary symmetric channel (BSC) model, a common communication channel model. The left-hand side variable ($x_{i,j}$) is the source information, and the right-hand side variable ($y_{i,j}^{(k)}$) is the received information. The received information is flipped with a probability of $1-P_{i,j}^{(k)}$, and otherwise correct.

Now, we can compute the MI between $x_{i,j}$ and $y_{i,j}^{(k)}$, which indicates how much information the preference feedback conveys about the true ranking:

$$\begin{aligned} I(x_{i,j}, y_{i,j}^{(k)}; P_{i,j}^{(k)}) = 1 + P_{i,j}^{(k)} \log_2(P_{i,j}^{(k)}) \\ + (1-P_{i,j}^{(k)}) \log_2(1-P_{i,j}^{(k)}). \end{aligned} \tag{2}$$

Note that the MI is parameterized by the feedback accuracy of the trainer, which determines the conditional probability $P(y_{i,j}^{(k)} \mid x_{i,j})$.[1]

Figure 3 illustrates the relationship between MI and feedback accuracy ($P_{i,j}$). It is obvious that a trainer with a higher probability of right feedback provides more information regarding the true ranking per feedback.

Next, we consider multiple trainers who may have different rationality values. As their rationalities are different, their feedback accuracies (denoted as $P_{i,j}^{(k)}$) will differ even for

[1] Equation (2) outlines the connection between the mutual information of the true ranking $x_{i,j}$ with the observed human preference $y_{i,j}^{(k)}$ and the entropy of the decision process. The derivation shows that the mutual information is equal to the negative entropy of $P_{i,j}^{(k)}$ up to addition of a constant. As the entropy is a measure of uncertainty inherit to a random variable's outcome, this is an intuitive connection to our understanding of rationality, which can be seen as a measure of the human's uncertainty in the choice. In this case, the mutual information can also be seen as a measure of "peakedness" of the distribution $P_{i,j}^{(k)}$.

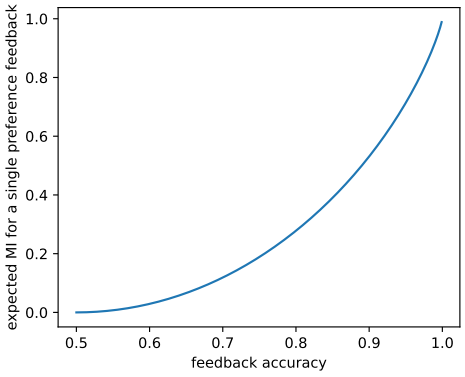

Figure 3. Mutual information for a binary symmetric channel (BSC) model with sweeping probabilities of the correct message.

the same pair of items. Hence, they all have different MI values for their feedback.

Let r be a probability distribution over the trainers, with the probabilities depending on the fraction of feedback given by the trainers. Concretely, let $r(k)$ be the fraction of feedback that trainer k contributed. We can now sample any given preference in a two-stage manner, by first sampling a trainer and then a preference given by that trainer. We now consider two different scenarios: one where we take trainer identities and the trainer's individual accuracies into account and one where we ignore trainer identities and assume average accuracy.

If we treat the feedback from different trainers separately and extract information for each trainer individually under consideration of their accuracy (assuming we know the rationality), then the expected MI for any given observed preference is given by $\mathbf{E}_{k\sim r}[I(x_{i,j}, y_{i,j}^{(k)}; P_{i,j}^{(k)})]$.

On the other hand, suppose that we aggregate all trainers' feedback without differentiating them. This is equivalent to a scenario with just a single trainer who represents the average opinion of all trainers, i.e. who casts a vote agreeing with the true ranking with probability $\bar{P}_{i,j} = \mathbf{E}_{k\sim r}[P_{i,j}^{(k)}]$. Hence, the expected MI for the feedback would be given by $I(x_{i.j}, y_{i,j}^{(k)}; \mathbf{E}_{k\sim r}[P_{i,j}^{(k)}])$.

Now, since Equation (2) is a strictly convex function of $P_{i,j}^{(k)}$ (see Cover & Thomas (2012) and Fig. 3), Jensen's inequality implies

$$\mathbf{E}_{k\sim r}[I(x_{i.j}, y_{i,j}^{(k)}; P_{i,j}^{(k)})] \geq I(x_{i.j}, y_{i,j}^{(k)}; \mathbf{E}_{k\sim r}[P_{i,j}^{(k)}])\,,$$

where the exact equality applies if and only if the $P_{i,j}^{(k)}$ are equal for all k (trainers), i.e. all trainers share the same rationality. This indicates that we always get equal or more information about the actual ranking by extracting information separately for each trainer under consideration of their individual rationalities.

It should be noted that the above analysis assumes knowledge of the true rationality parameters and does not account for the cost of estimating them. Despite this limitation, we observed that our algorithm achieves performance close to the values suggested by the MI analysis when the number of choice alternatives is much larger than the number of trainers. To account for estimation overhead, we plan to conduct an analysis using Fisher information, which naturally incorporates the interactions between estimation errors of multiple parameters. We consider this as a direction for future work.

5. Evaluation

We evaluated two types of tasks to test the benefits of estimating the rationality parameter. The first task is a basic ranking exercise to demonstrate whether estimating rationality can be advantageous in such minimal settings, which are, for example, crucial to evaluate the performance of labellers on crowd-sourcing platforms. The second task, set within a reinforcement learning (RL) environment, investigates whether the algorithm can provide any benefits over the standard assumption in Preference-based RL of assuming equal rationalities for all labellers.

For these evaluations, we simulated the human preference feedback using the Boltzmann-rational model with the given utility and rationality parameters. This allows us to control the reliability of each trainer and establish ground truth values to evaluate our algorithm and baselines.

5.1. Simple Ranking Task

Setup The first task is a simple ranking task involving four items. These items are ranked based on preference feedback from three trainers, each with different rationality values. We generate the pairwise preference feedback on randomly selected two items using the Boltzmann-rational model. The utility values for the four items are set at [1.0, 1.1, 1.2, 0.7], and the rationality values for the three trainers are [1.0, 0.5, 2.0]. We then evaluate three scenarios: 1. Estimating the rationality values using our approach. 2. Assuming all the rationality values to be 1.0 (conventional approach). 3. Assuming knowledge of all trainers' true rationality values (ideal). We run 200 trials and measured the ranking error rate for these three scenarios with varying numbers of feedback.

Results Table 1 shows ranking error rates for the simple ranking task involving four items with three trainers whose rationality parameters are 1.0, 0.5, and 2.0. The results show that our approach (estimate β) performs better than

number of FBs	β estimated ours	$\beta_i = 1.0$ fixed	$\beta = [1.0, 0.5, 2.0]$ ideal
100	<u>0.318</u>	0.330	**0.313**
200	0.203	<u>0.200</u>	**0.165**
500	<u>0.085</u>	0.100	**0.080**
1000	<u>0.020</u>	0.060	**0.005**

Table 1. Ranking error rates for the simple ranking task involving four items with three trainers. The best results are highlighted with boldface, and the second bests are highlighted underlined. The results show that our approach performs better than the conventional (fixed) approach in most cases and is close to the ideal scenario.

the conventional approach (assuming all $\beta = 1.0$) in most cases and is close to the ideal scenario (fixing β with the correct value).

It is to be noted that this must not hold in the asymptotic case with an unlimited number of samples where we expect the performance of all approaches to converge. Further details on the asymptotic behaviour can be found in Appendix A.

5.2. RL Task

Setup We assess the performance of our approach in an RL setting. In this setting, an agent interacts with a target environment and obtains preference feedback from trainers in order to learn the most suitable action for the current state. Unlike traditional settings, the agent does not receive rewards from the environment; instead, it relies on the feedback from the trainers to determine the best course of action.

We trained the agent repeatedly one hundred times over one thousand episodes and measured the total reward obtained from the environment for each episode. This shows how quickly the agent learns the right course of action from the feedback. Like in the previous task, we evaluate three scenarios: 1. Estimating the rationality values using our approach. 2. Assuming all the rationality values to be 1.0 (conventional approach). 3. Assuming knowledge of all trainers' true rationality values (ideal).

Environment We used a 5x5 grid world PacMan to evaluate our approach. The goal is to eat all the food pellets without being caught by the ghosts. Once the game is cleared (e.g. finished successfully), a +500 reward is given, while a -500 reward is given if PacMan is caught by ghosts. Also, each pellet awards +10 points, and each time-step costs 1 point (rewarded -1 point.) The state representation includes PacMan's position, the position and orientation of the ghost, and each food pellet's presence. Figure 1 shows the image of the grid world PacMan.

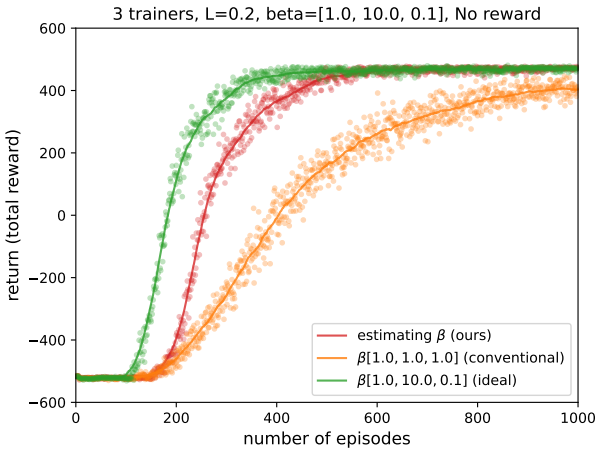

Figure 4. Evaluation results for RL task that learns a policy with preference feedback from three trainers. Red: estimating the rationality parameters (ours). Orange: assuming all trainers have the same rationality (conventional). Green: fixing the rationality to the true value (ideal). The results suggest a clear benefit of our approach over the conventional approach.

Preference Feedback Generation We use an Oracle to simulate human feedback. This allows us to sweep different parameters of feedback likelihood (L) that specify the rationality values and feedback frequency (how often feedback is provided by a trainer). The Oracle was created using Q-learning (Watkins, 1989) on the environment prior to the experiments. We scaled the learned Q function by a factor of 0.01, and used it as the utility value of each action in the given state. Then, we employ the Boltzmann-rational model to generate the preference feedback with the utility and rationality values.

In this experiment, we set the rationality parameters of the three trainers at 1.0, 10.0, 0.1, and the feedback likelihood at $L = 0.2$. This means a trainer provides feedback, on average, once every five time-steps.

Results Figure 4 shows the total reward vs number of training episodes. Each dot represents the average total reward across one hundred learning trials, and the lines come from applying a moving average to the dots over fifty time-steps. The results show that our approach (red) achieves high rewards faster than the conventional approach (which assumes all trainers have the same rationality values). Although it is still slower than the ideal case, representing the upper bound of the achievable performance.

These results show that our method of estimating the rationality parameters can enhance feedback efficiency and achieve better performance compared to the conventional

approaches (not estimating the rationalities) with the same amount of feedback.

6. Limitations

Though we have shown the advantages of estimating the rationality parameters, our approach is not without limitations, some of which can be alleviated in future work.

- We assume shared utilities among the trainers. This assumption may not hold in practice.
- We assume a Boltzmann-rational model, which attributes choices entirely to utilities and rationality. However, other factors, such as noise, biases, or context, may influence the choices.
- Our approach assumes that the trainers' rationality parameters are constant over time. However, the trainers' rationality may change over time due to various factors, such as fatigue, learning, or mood.
- Our mutual information based analysis assumes the knowledge of the true rationality parameters and does not account for the cost of estimating them. We need an analysis method that takes into account the rationality parameter estimation (e.g. Fishier information).

7. Conclusion

We present a pairwise preference learning approach that jointly learns the trainer's shared utilities and each trainer's individual rationality level. Our empirical evaluation is promising, indicating that we can match or improve utility learning performance while simultaneously learning the relative rationalities of the labellers without any extra calibration data. This has many potential downstream use-cases in settings such as crowdsourcing. We moreover provide an initial theoretical analysis of our approach, indicating benefits of our method with respect to preference learning efficiency. Our early results are not yet conclusive, but are a promising indicator for the success of our approach. Future work may alleviate limitations (Section 6) and further explore empirical and theoretical benefits while also applying our approach to more settings.

Acknowledgments

This work is supported by the UKRI Turing AI Fellowship EP/V024817/1, the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101073307 (LEMUR) and TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. It was further supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria.


Co-funded by
the European Union

References

Ben-Akiva, M. and Morikawa, T. Estimation of switching models from revealed preferences and stated intentions. *Transportation Research Part A: General*, 24(6):485–495, 1990. doi: 10.1016/0191-2607(90)90037-7.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2012. ISBN 978-1-118-58577-1.

Daniels-Koch, O. and Freedman, R. The Expertise Problem: Learning from Specialized Feedback. In *NeurIPS 2022 Workshop on ML Safety*, 2022. URL https://openreview.net/forum?id=I7K975-H1Mg.

Davidson, R. R. and Solomon, D. L. A bayesian approach to paired comparison experimentation. *Biometrika*, 60(3): 477–487, 1973.

Dawid, A. P. and Skene, A. M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Freedman, R., Svegliato, J., Wray, K., and Russell, S. Active teacher selection for reinforcement learning from human feedback, 2023. URL http://arxiv.org/abs/2310.15288. preprint.

Ghosal, G. R., Zurek, M., Brown, D. S., and Dragan, A. D. The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023. doi: 10.1609/aaai.v37i5.25740.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., and Thomaz, A. L. Policy Shaping: Integrating Human Feedback with Reinforcement Learning. In *Advances*

in Neural Information Processing Systems (NIPS), volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html.

Jeon, H. J., Milli, S., and Dragan, A. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2f10c1578a0706e06b6d7db6f0b4a6af-Abstract.html.

Kaufmann, T., Ball, S., Beck, J., Kreuter, F., and Hüllermeier, E. On the Challenges and Practices of Reinforcement Learning from Real Human Feedback. In *ECML PKDD 2023 Workshop Towards Hybrid Human-Machine Learning and Decision Making*, 2023a.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A Survey of Reinforcement Learning from Human Feedback, 2023b. URL http://arxiv.org/abs/2312.14925. preprint.

Knox, W. B. and Stone, P. Reinforcement learning from simultaneous human and MDP reward. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, 2012.

Metz, Y., Lindner, D., Baur, R., Keim, D. A., and El-Assady, M. RLHF-Blender: A Configurable Interactive Interface for Learning from Diverse Human Feedback. In *ICML 2023 Workshop on Interactive Learning with Implicit Human Feedback*, 2023. URL https://openreview.net/forum?id=JvkZtzJBFQ.

Newman, M. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25, 2023.

OpenAI. Introducing ChatGPT, 2022. URL https://openai.com/index/chatgpt/. (accessed 2024-05-28).

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning From Crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010. ISSN 1533-7928. URL http://jmlr.org/papers/v11/raykar10a.html.

Swait, J. and Louviere, J. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. *Journal of Marketing Research*, 30(3):305–314, 1993. doi: 10.2307/3172883.

Train, K. E. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511805271.

Watkins, C. J. C. H. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge United Kingdom, 1989.

Whelan, J. T. Prior distributions for the bradley-terry model of paired comparisons. *arXiv preprint arXiv:1712.05311*, 2017.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J., and Ruvolo, P. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems (NIPS)*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/hash/f899139df5e1059396431415e770c6dd-Abstract.html.

Yamagata, T., McConville, R., and Santos-Rodríguez, R. Reinforcement Learning with Feedback from Multiple Humans with Diverse Skills. In *NeurIPS 2021 Workshop on Safe and Robust Control of Uncertain Systems (SafeRL)*, 2021.

A. Asymptotic behaviour of maximum likelihood estimation under wrong rationality assumptions

Here, we further want to investigate how assuming wrong rationalities impacts the maximum likelihood estimation of the utility differences in the asymptotic case. We assume the Bradley-Terry model, linking (latent) utilities θ_1, θ_2 to observed preferences

$$P(a_1 \succ a_2) = \frac{\theta_1}{\theta_1 + \theta_2}\,, \tag{3}$$

which we can connect to the Boltzmann-rational model with fixed rationality $\tilde{\beta}$ using $\theta_1 = e^{\tilde{\beta} u_1}$ and $\theta_2 = e^{\tilde{\beta} u_2}$. Then, the likelihood of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ given the data can be computed as

$$\mathcal{L}(\boldsymbol{\theta}) = \left(\frac{\theta_1}{\theta_1 + \theta_2}\right)^{n_1} \cdot \left(\frac{\theta_2}{\theta_1 + \theta_2}\right)^{n_2},$$

with n_1 and n_2 being the number of times choice alternative one or two were preferred, respectively. The maximum likelihood estimate of Equation (3) is found at the extreme of the PDF, e.g. where the first derivative of the log-likelihood function

$$ln(\mathcal{L}(\boldsymbol{\theta})) = n_1 \cdot \ln \frac{\theta_1}{\theta_1 + \theta_2} + n_2 \cdot \ln \frac{\theta_2}{\theta_1 + \theta_2} \tag{4}$$

equals 0, i.e.

$$\frac{d}{d\theta_1} ln(\mathcal{L}(\boldsymbol{\theta})) = 0 \iff \frac{n_1 \cdot \theta_2}{\theta_1 \cdot (\theta_1 + \theta_2)} - \frac{n_2}{\theta_1 + \theta_2} = 0 \iff \frac{\theta_1}{\theta_2} = \frac{n_1}{n_2}\,. \tag{5}$$

We can assume without loss of generality that $\theta_2 = 1 - \theta_1$, since in that case $\theta_1 + \theta_2 = 1$ and $P(a_1 \succ a_2) = \theta_1$. Combining this with Equation (5) yields the maximum likelihood estimate

$$\frac{\hat{\theta}_1}{1 - \hat{\theta}_1} = \frac{n_1}{n_2} \iff \hat{\theta}_1 = \frac{n_1}{n_2 + n_1}\,. \tag{6}$$

Relating this back to utilities and rationalities with $\hat{\theta}_i = e^{\tilde{\beta} \hat{u}_i}$ results in

$$\hat{u}_1 = \frac{\ln n_1 - \ln (n_2 + n_1)}{\tilde{\beta}} \quad \text{and} \quad \hat{u}_2 = \frac{\ln n_2 - \ln (n_2 + n_1)}{\tilde{\beta}}\,. \tag{7}$$

As $n = n_1 + n_2$ goes to infinity, the sample mean approaches the true Bernoulli parameter and it follows that $n_1 = P(a_1 \succ a_2) \cdot n$. Due to our assumption on $\theta_2 = (1 - \theta_1)$, it further follows that $n_1 = \theta_1 \cdot n = e^{\beta \cdot u_1} \cdot n$ and analogously $n_2 = e^{\beta \cdot u_2} \cdot n$. Thus

$$\hat{u}_1 = \frac{\ln n_1 - \ln (n_2 + n_1)}{\tilde{\beta}} = \frac{\ln(e^{\beta \cdot u_1} \cdot n) - \ln (n_2 + n_1)}{\tilde{\beta}} = \frac{\beta \cdot u_1 + \ln n - \ln n}{\tilde{\beta}} = \frac{\beta}{\tilde{\beta}} u_1 \tag{8}$$

and analogously $\hat{u}_2 = (\beta / \tilde{\beta}) u_2$.

We see that the learned utilities are scaled by the true rationality over the assumed rationality $\beta / \tilde{\beta}$, i.e. compared to the case where the true rationalities are known (resulting in $\hat{u}_i = u_i$), the estimated utilities in the asymptotic case will only be rescaled by a constant factor. As a consequence, irrespective of the distribution over the possible outcomes, the policy maximizing expected utility remains unchanged. Therefore, assuming the wrong rationality should not change the optimal policy, at least in the asymptotic setting.

This can be related to the setting of multiple trainers by aggregating them to a hypothetical 'merged' trainer with an intermediate rationality. Then, the same reasoning applies, resulting in no difference in the optimal policy in the asymptotic setting.

Note that while this study may initially suggest that estimating rationalities offers no benefits for utility estimation for the purposes of policy learning, it does not allow us to reach any immediate conclusions regarding the case of limited data and imperfect learning algorithms. The empirical results in the main body of the paper (Section 5) seem to suggest the potential for improvement in this setting. However, a more thorough evaluation is necessary to further understand the effect of learning rationalities.