



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Multi-Dimensional Reward Learning from Preference Feedback

LAUREA MAGISTRALE IN COMPUTER SCIENCE AND ENGINEERING - INGEGNERIA INFORMATICA

Author: LEONARDO BIANCONI

Advisor: PROF. ALBERTO MARIA METELLI

Co-advisors: SIMONE DRAGO, MARCO MUSSI

Academic year: 2024-2025

1. Introduction

Reinforcement Learning (RL) is a well-known subfield of Machine Learning (ML) in which an agent learns to solve sequential decision-making problems from data. The agent interacts with an environment, usually defined as a Markov Decision Process (MDP), by, at each time step, observing the state of the environment (s_t), playing an action (a_t), and receiving a scalar *reward* (r_t).

The goal of the agent is finding the *policy* (i.e., a probabilistic mapping between a state and the action to play) maximizing the long-term sum of collected rewards. While this framework's theoretical and applicative research grows exponentially over time, a criticality in the RL setting is the dependence on the reward signal. Such reward encodes the notion of the goal for the task, needs to be manually selected between an a possibly infinite set of plausible choices, and is heavily problem-dependent. Moreover, a misspecification in the reward function generates unwanted behaviors, the most important one being *reward hacking*, where the agent maximizes the specified reward, while not pursuing the real goal of the task.

To overcome these challenges, several alternative frameworks have been proposed.

Preference-based Reinforcement Learning [PbRL, 4] tries to solve the problem by introducing the possibility to query a human expert for preferences between pairs of trajectories (i.e., played state, action sequences on the MDP, formally $\tau = (s_t, a_t)_{h \in [H]} \in \mathcal{T}$). The expert is modeled probabilistically, defining a preference (or expert) model, i.e. $\rho(\tau_1, \tau_2) := P(\tau_1 \succ \tau_2)$ (preference of τ_1 being preferred over τ_2). PbRL algorithms try to find the optimal policy by either (a) directly learning it from preference data, (b) learning a surrogate preference model $P(a_i \succ a_j | s)$ or (c) learning a surrogate utility, also called *reward model* $U : \mathcal{T} \rightarrow \mathbb{R}$, recover a *Markovian reward* $R(s, a)$ and then apply standard algorithms for solving MDPs.

This last method is the most used and analyzed in the literature. The de-facto standard preference model is the ubiquitous Bradley-Terry [BT, 1] model, where:

$$P(\tau_1 \succ \tau_2) := \sigma(U(\tau_1) - U(\tau_2)), \quad (1)$$

with $\sigma(z) := (1 + \exp(-z))^{-1}$ being the *logistic sigmoid* function. In the case of a linear reward model $U(\tau) := \mathbf{w}^\top \phi(\tau)$ (where $\phi : \mathcal{T} \rightarrow \mathbb{R}^k$ is a known *trajectory feature mapping*), this model is optimizable by Maximum Likelihood Estimation (MLE), via minimizing the negative log-

likelihood (NLL) function, i.e.:

$$-\sum_{i=1}^N \log \sigma(U(\tau_{i1}) - U(\tau_{i2})), \quad (2)$$

where $\zeta = \{\tau_{i1} \succ \tau_{i2}\}_{i \in [N]}$ is a collected dataset of preferences.

Another framework coping with problems in the reward specification is Multi-Objective Reinforcement Learning [MORL, 3]. In MORL, the reward signal is modeled as multi-dimensional. This allows to specify multiple, possibly contrasting objectives, thereby alleviating the burden of selecting an appropriate scalarization. Algorithms can find a set of *Pareto-optimal* policies, i.e., policies that obtain a performance not improvable in all objectives.

Recently, [2] analyzed the computational issues in PbRL, focusing on the utility representations of pre-order relations between trajectories. The authors conjecture the problem of assessing policy (Pareto-)optimality with respect to partial orders (that is, a relation in \mathcal{T} in which trajectories can be *incomparable*), to be computationally intractable. This motivates the shift towards (multi-dimensional) rewards, that allow for more friendly computational properties.

Original Contribution. To the best of our knowledge, no work has ever dealt with reward model learning in the presence of an expert that can mark two trajectories as incomparable. This possibility is intrinsically bound to the presence of a trade-off between underlying true utilities of the trajectories. This thesis investigates this research direction that links two already developed research areas, namely PbRL and MORL. This working direction is also noted in [4].

The contribution can be summarized as follows:

- In Section 2 we define the novel setting of the problem, we postulate desirable properties (Definition 2.1) for an expert model in this setting and we present a negative result concerning the impossibility of defining a model complying with such properties that is also optimizable via MLE with a convex loss (Theorem 2.1).
- In Section 3 we define two choices of expert model complying with the properties of Definition 2.1 (Model A, in Section 3.1 and Model B, in Section 3.2). Furthermore,

we present an expert model complying partially with such conditions but enjoying a convex loss function (Model C, in Section 3.3), expressing its expressiveness limitations.

2. Setting, Desiderata and Negative Results

We formalize this setting employing a probabilistic model for the expert, as classical in PbRL literature [4]. In these works, the expert has been modeled using a probability function $\rho : \mathcal{T} \times \mathcal{T} \rightarrow [0, 1]$, $\rho(\tau_1, \tau_2) = P(\tau_1 \succ \tau_2)$. In this novel setting, the problem is a 4-class one. In fact, the expert is able to, given (τ_1, τ_2) , output a label in $\mathcal{Y} := \{\succ, \prec, \asymp, \parallel\}$ (i.e., strict preferences, indifference or incomparability, respectively).

Moreover, following PbRL literature, we assume the presence of a real utility function $\hat{U} : \mathcal{T} \rightarrow \mathbb{R}^d$, unknown to the learner, that the expert uses to label trajectory pairs, i.e., given (τ_1, τ_2) , a label is generated as $y \sim \gamma(\Delta \hat{U}(\tau_1, \tau_2))$ (where $\Delta \hat{U}(\tau_1, \tau_2) := \hat{U}(\tau_1) - \hat{U}(\tau_2)$), for a suitable $\gamma : \mathbb{R}^d \rightarrow \Delta(\mathcal{Y})$. The learner observes a dataset $\zeta := \{(\tau_{i1}, \tau_{i2}, y_i)\}_{i \in [N]}$ where the labels y_i are sampled from γ .

2.1. Desiderata for γ

In the following, we postulate a set of properties that γ should satisfy.

Definition 2.1 (Rational Multi-Objective Expert Model). A probability function $\gamma : \mathbb{R}^d \rightarrow \Delta(\mathcal{Y})$ is said to be a rational multi-objective expert model if and only if the following conditions are satisfied:

1. $\gamma(\Delta \mathbf{U}) \rightarrow (1, 0, 0, 0)^\top$ as $\Delta \mathbf{U} \rightarrow (+\infty, \dots, +\infty)^\top$.
2. $\gamma(\Delta \mathbf{U}) \rightarrow (0, 1, 0, 0)^\top$ as $\Delta \mathbf{U} \rightarrow (-\infty, \dots, -\infty)^\top$.
3. $\gamma(\Delta \mathbf{U}) \rightarrow (0, 0, 1, 0)^\top$ as $\Delta \mathbf{U} \rightarrow \mathbf{0}_d$.
4. $\forall \mathbf{v} \in \{-1, 1\}^d$ s.t. $\mathbf{v} \neq \pm \mathbf{1}_d$: $\gamma(\Delta \mathbf{U}) \rightarrow (0, 0, 0, 1)^\top$ as $\Delta \mathbf{U} \rightarrow \mathbf{v} \odot (+\infty, \dots, +\infty)^\top$.
5. $\forall t \in \mathbb{R} : \gamma_4(t \cdot \mathbf{1}_d) = 0$.
6. Let $\mathcal{V}_\succ := \{(+\infty, \dots, +\infty)^\top\}$, $\mathcal{V}_\prec := \{(-\infty, \dots, -\infty)^\top\}$, $\mathcal{V}_\asymp := \{\mathbf{0}_d\}$, $\mathcal{V}_\parallel := \{\mathbf{v} \odot (+\infty, \dots, +\infty)^\top, \mathbf{v} \in \{-1, 1\}^d \text{ s.t. } \mathbf{v} \neq \pm \mathbf{1}_d\}$. For all possible choice of vectors \mathbf{x}, \mathbf{y} chosen from distinct sets between $\mathcal{V}_\succ, \mathcal{V}_\prec, \mathcal{V}_\asymp$ and \mathcal{V}_\parallel , the value of γ for the non-zero

components are monotone on the line segment joining \mathbf{x} and \mathbf{y} , i.e., on $\{\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}, \lambda \in [0, 1]\}$.

Condition 1 (resp. 2) is associated with class \succ (resp. \prec): it requires that, if all the components of the utility are greater for τ_1 (resp. τ_2), the probability mass should be shifted towards \succ (\prec). Condition 3 states that, if all the objectives are similar for τ_1 and τ_2 , the probability mass should be placed on \asymp . Condition 4 requires a high probability to the incomparability class for significant trade-offs in the utility values for τ_1 and τ_2 . Condition 5 states that, if all the objectives are equal (therefore a case reducible to the scalar utility case), $P(\tau_1 \parallel \tau_2) \rightarrow 0$. Condition 6 enforces a weak notion of monotonicity of γ .

2.2. Non Convexity of NLL

As in classic PbRL, a possible solution for this problem is postulating a parametric reward model, i.e. $\mathbf{U}(\tau) = h(\mathbf{W}, \phi(\tau_1), \phi(\tau_2))$ and learning the optimal value of \mathbf{W} via MLE, i.e.:

$$\hat{\mathbf{W}}^{MLE} = \arg \min_{\mathbf{W}} - \sum_{i=1}^N \log(\gamma_{y_i}(\Delta\mathbf{U}(\tau_{i1}, \tau_{i2}))). \quad (3)$$

In the following, an important negative result concerning every γ compliant with Definition 2.1 is presented.

Theorem 2.1. *There exists no function $f : \mathbb{R} \rightarrow [0, 1]$ such that $\lim_{x \rightarrow -\infty} f(x) = 1 \wedge \lim_{x \rightarrow +\infty} f(x) = 0 \wedge \lim_{x \rightarrow 0} f(x) = 1$ and such that the function $g : [0, 1] \rightarrow \mathbb{R}$ defined as $g(x) = -\log(f(x))$ is convex in x .*

Remark 2.1 (Non-Convexity of Negative Log-Likelihood). *Let $\mathbf{v} \in \{-1, 1\}^d$ s.t. $\mathbf{v} \neq \pm \mathbf{1}_d$. We note that f is the restriction on $t \cdot \mathbf{v}, t \in \mathbb{R}$ of the incomparability component of a multi-objective expert model $\gamma : \mathbb{R}^d \rightarrow \Delta(\mathcal{Y})$ that is compliant with Conditions 4 and 5 of Definition 2.1. Moreover, we note that g is the restriction on $t \cdot \mathbf{v}, t \in \mathbb{R}$ of the negative log-likelihood function for the incomparability class, i.e.:*

$$\begin{aligned} l(\Delta\mathbf{U}) &:= -\log(\gamma_{\parallel}(\Delta\mathbf{U})), \\ g(t) &= l(t \cdot \mathbf{v}). \end{aligned} \quad (4)$$

Therefore, $l(\Delta\mathbf{U})$ is not convex in $\Delta\mathbf{U}$. Since $\Delta\mathbf{U}$, i.e., the reward model, is a function of

some optimization parameter \mathbf{W} , in the general case, l is not convex in \mathbf{W} . It follows from this result that there exists no model γ that is optimizable via MLE with convex optimization algorithms (differently, for example, from the Bradley-Terry model) and compliant with Definition 2.1.

3. Models Definitions

As stated in Remark 2.1, there exists no expert model that simultaneously (a) complies with all the properties of Definition 2.1 and (b) has a convex negative log-likelihood. However, one could trade expressivity (defined as degree of compliance with the properties of Definition 2.1) with convexity of the negative log-likelihood. In what follows, we first try to achieve full expressiveness (Models A and B), disregarding optimization guarantees, and then define a model that trades off the least degree of expressiveness in favor of a convex NLL (Model C).

A final desiderata for every model is that they should reduce to the BT model in the standard, scalar case ($d = 1$). To do this, the natural choice is to use the multinomial logit (or softmax) model, that generalizes the BT model to the case of $n \geq 3$ output classes (in our case $n = 4$):

$$\gamma_y(\Delta\mathbf{U}) = \frac{\exp f_y}{\sum_{y' \in \mathcal{Y}} \exp f_{y'}}, y \in \mathcal{Y}, \quad (5)$$

The main objective can then be reduced to finding suitable logits (that need to depend from the utility difference $\Delta\mathbf{U}$, as the BT model) such that the conditions of Definition 2.1 are satisfied.

3.1. Model A

For simplicity, we reason in the $d = 2$ case, focusing on the $\Delta U_1 / \Delta U_2$ plane. Considering the four quadrants, we need a high logit for the \succ class in the I quadrant since $\Delta U_1 > 0 \wedge \Delta U_2 > 0 \iff (U_1(\tau_1) > U_1(\tau_2)) \wedge (U_2(\tau_1) > U_2(\tau_2))$, in other words, τ_1 has a higher utility under all objectives with respect to τ_2 .

The same reasoning can be done, with opposite signs, for the \prec class, obtaining that the logit for the \prec class needs to be high in the III quadrant. Incomparability means that the 2 objectives are contrasting, therefore the two differences ΔU_1

and ΔU_2 have different sign, implying a high logit for the \parallel class in the II and IV quadrants. Indifference (\asymp) means that the two trajectories are almost equivalent along the 2 objectives, i.e., $\Delta U_1 \simeq 0 \wedge \Delta U_2 \simeq 0$. This implies that the logit for the \asymp class needs to be only high around the origin.

We notice how the bisector $\Delta U_1 = \Delta U_2$ can ease the computation of such logit functions. For any $\Delta \mathbf{U} \in \mathbb{R}^2$, its projection on the bisector gives us, ignoring \asymp and \parallel classes, a view on the percentage of belonging to the \prec and \succ classes. Indeed, if its projection falls on the I quadrant, $P(\tau_1 \succ \tau_2)$ surely needs to be greater than $P(\tau_1 \prec \tau_2)$, and viceversa for the III quadrant.

Now, the \parallel class logit needs to capture the closeness to the II and IV quadrants. Moreover, as Condition 5 of Definition 2.1 requires, this function needs to be minimum along the bisector. An elegant choice is the *standard deviation* of $\Delta \mathbf{U}$, since $std((u, u)^\top) = 0$ and $\arg \max_{\Delta \mathbf{U} \in \mathbb{R}^2} \{std(\Delta \mathbf{U})\} = \{(-C, +C)^\top, (+C, -C)^\top\}$, $C \rightarrow +\infty$.

Lastly, we utilize a (learnable) constant K as logit for the indifference class (such that $P(\tau_1 \asymp \tau_2)$ is high only when all other logits are much smaller than K , i.e., around the origin).

We now formally define what we just discussed, also generalizing in \mathbb{R}^d :

Model A Definition.

$$\gamma_y(\Delta \mathbf{U}) = \frac{\exp f_y(\Delta \mathbf{U})}{\sum_{y' \in \mathcal{Y}} \exp f_{y'}(\Delta \mathbf{U})}, y \in \mathcal{Y} \quad (6)$$

$$f_{\succ}(\mathbf{p}) := \frac{1}{\sqrt{d}} \sum_{i=1}^d p_i = -f_{\prec}(\mathbf{p}) \quad (7)$$

$$f_{\parallel}(\mathbf{p}) = \sqrt{d} \cdot std(\mathbf{p}). \quad (8)$$

$$f_{\asymp}(\mathbf{p}) := K \quad (9)$$

3.2. Model B

We can obtain a model that is conceptually similar to Model A by utilizing 2^d linear functions of $\Delta \mathbf{U}$, each having maximum value in one orthant of \mathbb{R}^d . In \mathbb{R}^2 , such functions are the four planes $\Delta U_1 + \Delta U_2$, $\Delta U_1 - \Delta U_2$, $-\Delta U_1 + \Delta U_2$ and $-\Delta U_1 - \Delta U_2$.

We note how $\Delta U_1 + \Delta U_2$ ($-\Delta U_1 - \Delta U_2$) is a measure of \succ (\prec) - indeed these functions are already employed in Model A. On the other hand, $\Delta U_1 - \Delta U_2$ and $-\Delta U_1 + \Delta U_2$ are a measure of incomparability (in the case of a generic d , we have $2^d - 2$ functions associated with \parallel).

The idea is to use 2^d classes (instead of 4, as in Model A) and group probability outputs for the scores associated with the \parallel class (i.e., $f_{\parallel}^{(j)}$).

Below is the formal definition of Model B:

Model B Definition.

$$f_{\succ}(\Delta \mathbf{U}) := \mathbf{1}_d^\top \cdot \Delta \mathbf{U} = \sum_{i=1}^d \Delta U_i, \quad (10)$$

$$f_{\prec}(\Delta \mathbf{U}) := -\mathbf{1}_d^\top \cdot \Delta \mathbf{U} = -\sum_{i=1}^d \Delta U_i, \quad (11)$$

$$f_{\asymp}(\Delta \mathbf{U}) := K, \quad (12)$$

$$f_{\parallel}^{(j)}(\Delta \mathbf{U}) := \mathbf{v}^{(j)\top} \cdot \Delta \mathbf{U} = \sum_{i=1}^d v_i^{(j)} \Delta U_i. \quad (13)$$

$$\gamma_y(\Delta \mathbf{U}) = \frac{\exp f_y(\Delta \mathbf{U})}{DEN(\Delta \mathbf{U})}, y \in \{\succ, \prec, \asymp\}, \quad (14)$$

$$\gamma_{\parallel}(\Delta \mathbf{U}) = \frac{\sum_{j=1}^{2^d-2} \exp f_{\parallel}^{(j)}(\Delta \mathbf{U})}{DEN(\Delta \mathbf{U})}, \quad (15)$$

where $\{\mathbf{v}^{(j)}\}_{j \in [2^d-2]} = \{-1, 1\}^d \setminus \{\mathbf{1}_d, -\mathbf{1}_d\}$ and DEN denotes the denominator of the model, i.e., the sum of all exponentiated logit functions.

3.3. Model C

We notice that the conditions in Definition 2.1 that prevent the NLL to be convex are Condition 4 and 5 (the one concerning the \parallel class). We note that, if we are able to define 4 functions f_{\succ} , f_{\prec} , f_{\asymp} and f_{\parallel} such that they are linear in $\Delta \mathbf{U}$, then convexity of the NLL is guaranteed.

f_{\succ} , f_{\prec} and f_{\asymp} as defined in Model A and B are already linear in $\Delta \mathbf{U}$; we need to linearize the function $f_{\parallel}(\Delta \mathbf{U}) = std(\Delta \mathbf{U})$ (or any equivalent function). However, this is a non-trivial task, as no linear approximator g of $std(\Delta \mathbf{U})$ has a bounded error $\|std(\Delta \mathbf{U}) - g(\Delta \mathbf{U})\|$ over \mathbb{R}^d . As it can be demonstrated, the best choice is a simple constant: $f_{\parallel}(\Delta \mathbf{U}) = \alpha$.

However, this choice is problematic as $f_{\succ}(\Delta U) = K$ is now equivalent to f_{\parallel} . This is an intrinsic limitation, and one choice is to discard indifferences and only limit to model classes \succ , \prec and \parallel .

The formal definition of Model C follows exactly from this reasoning:

Model C Definition.

$$\gamma_y(\Delta U) = \frac{\exp f_y(\Delta U)}{\sum_{y' \in \{\succ, \prec, \parallel\}} \exp f_{y'}(\Delta U)}, \quad (16)$$

$$f_{\succ}(\Delta U) := \mathbf{1}_d^\top \cdot \Delta U = \sum_{i=1}^d \Delta U_i, \quad (17)$$

$$f_{\prec}(\Delta U) := \mathbf{1}_d^\top \cdot \Delta U = \sum_{i=1}^d \Delta U_i, \quad (18)$$

$$f_{\parallel}(\Delta U) := \alpha \quad (19)$$

3.4. Considerations

Once a suitable reward model $\Delta U(\tau_1, \tau_2) = h(\mathbf{W}, \phi(\tau_1), \phi(\tau_2))$ is chosen (e.g., a linear reward model $\Delta U(\tau_1, \tau_2) = \mathbf{W}(\phi(\tau_1) - \phi(\tau_2))$), parametrized by \mathbf{W} , the optimal parameters can be found minimizing the NLL function (Equation 3). This optimization is guaranteed to be convex in the sole case of Model C, for linear reward models.

As already seen, Model C has limited expressivity, since its decision surfaces are hyperplanes in \mathbb{R}^d ; therefore, it cannot isolate specific orthants like Model A and B. This can be seen by looking at the *argmax* of the predictions for Model A (Figure 1) and Model C (Figure 2). As for expressivity, Model A and B are exactly equivalent. It can be shown that Model C is equivalent to another already known extension of BT, known as the Bradley-Terry with Ties model, in which class \asymp is replaced by \parallel . It can be shown that, for $d = 1$, all models reduce to the BT model, for an appropriate choice of their parameters.

4. Experiments

In this section, we present the most significant results of our numerical experiments.

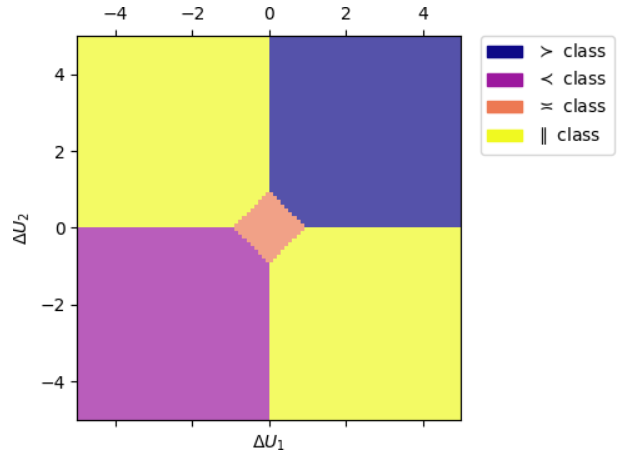


Figure 1: $\arg \max_y \gamma_y((\Delta U_1, \Delta U_2)^\top)$, Model A

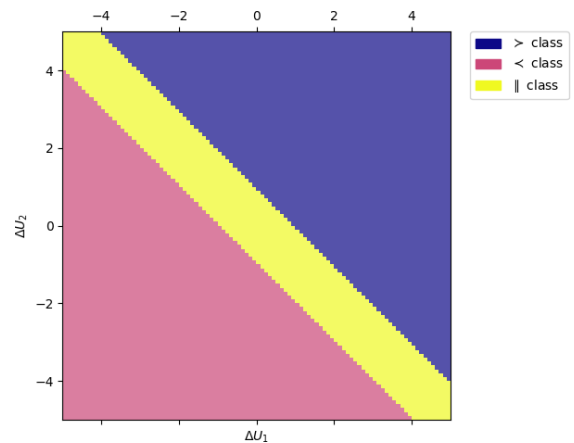


Figure 2: $\arg \max_y \gamma_y((\Delta U_1, \Delta U_2)^\top)$, Model C

Preliminary Validation on Synthetic Datasets. We fitted Models A, B and C to a dataset of size $N = 5000$, obtained sampling directly the $\phi(\tau)$ space. The labels were given by a fictitious expert model chosen between Models A, B and C, employing a linear reward model with parameter $\hat{\mathbf{W}} \in \mathbb{R}^{2 \times 3}$. Of particular importance is the analysis of the average *Jensen-Shannon divergence* between the real expert model outputs and the learner's one. Results are shown in Table 1.

| | Expert A | Expert B | Expert C |
|-----------|----------|----------|----------|
| Learner A | 0.0048 | 0.0072 | 0.0332 |
| Learner B | 0.0082 | 0.0043 | 0.0303 |
| Learner C | 0.0792 | 0.1347 | 0.0032 |

Table 1: Average Jensen-Shannon divergence for all combinations of learner/expert models (lower is better).

As can be seen in Table 1, Models A and B can

learn their parameters such that they adapt to a real expert model chosen between Models A and B. As for Model C, this table suggests, once again, its limitations: the high JS divergence values show that this model is not capable of learning decision surfaces that match the ones of Models A and B. Due to these theoretical limitations, backed by experimental evidence, Model C has been excluded from subsequent experiments.

Multi-Objective Gridworld. We further validate the models with a series of experiments on a 5×5 gridworld, with a starting state, a goal state, 2 obstacles and a three dimensional true reward. The dimensions of the problem are $d = 3$, $k = 25$. A set of $N_T = 300$ trajectories played by a random policy has been fully labeled by an expert chosen between Models A and B, obtaining $N = 44850$ trajectory pairs.

The focus of these experiments is placed on the robustness of the optimization procedure. Since the NLL is non-convex for Models A and B, a state-of-the-art non-convex optimizer tackles the minimization of the NLL. Each experiment has been repeated $N_{eval} = 10$ times, obtaining very low variance between found minima (see Table 2), indicating that, for more realistic and substantial datasets, the NLL has properties similar to convex functions.

| $l(\mathbf{W}^{MLE})$ | mean | std | min | max |
|-----------------------|----------|-------|----------|----------|
| Model A | 5202.66 | 1.08 | 5201.78 | 5204.74 |
| Model B | 13277.83 | 23.96 | 13264.60 | 13345.03 |

Table 2: Negative log-likelihood statistics for Models A and B.

5. Conclusions

In this thesis, we analyzed the problem of learning a multi-dimensional reward model from a dataset comprising preferences, indifferences and incomparabilities between pairs of trajectories. We first formalized the setting, extending the already established PbRL framework, accounting for a probabilistic expert outputting labels of the four aforementioned categories. We then postulated reasonable properties that such expert model should satisfy, defining the *rational multi-objective expert model*; we presented an important negative result: the impossibility

for any model complying with these properties to have a convex negative log-likelihood. We showed the derivations of some choices of models complying with such desiderata and reducing to the Bradley-Terry model in the scalar reward case (namely, Model A and Model B). We then achieved a convex-loss model that violates the least amount of properties defined above (Model C), and highlighted its expressiveness limitations and its inadequacy in correctly learning the various dimensions of the true underlying reward. Finally, we validated these models in small to sufficiently large datasets, highlighting, in these cases, the robustness of the optimization of Models A and B, even if they do not exhibit a convex loss.

References

- [1] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL <http://www.jstor.org/stable/2334029>.
- [2] S. Drago, M. Mussi, A. M. Metelli, et al. Towards theoretical understanding of sequential decision making with preference feedback. In *42nd International Conference on Machine Learning, ICML 2025*, pages 1–16, 2025.
- [3] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1), Apr. 2022. ISSN 1573-7454. doi: 10.1007/s10458-022-09552-y. URL <http://dx.doi.org/10.1007/s10458-022-09552-y>.
- [4] C. Wirth, R. Akrou, G. Neumann, and J. Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017. URL <http://jmlr.org/papers/v18/16-634.html>.