Preference Learning in Assistive Robotics: Observational Repeated Inverse Reinforcement Learning

Bryce Woodworth, Francesco Ferrari, Teofilo E. Zosa, Laurel D. Riek
Computer Science and Engineering
University of California San Diego

Abstract

As robots become more affordable and more common in everyday life, there will be an ever-increasing demand for adaptive behavior that is personalized to the individual needs of users. To accomplish this, robots will need to learn about their users' unique preferences through interaction. Current preference learning techniques lack the ability to infer long-term, task-independent preferences in realistic, interactive, incomplete-information settings. To address this gap, we introduce a novel preference-inference formulation, inspired by assistive robotics applications, in which a robot must infer these kinds of preferences based only on observing the user's behavior in various tasks. We then propose a candidate inference algorithm based on maximum-margin methods, and evaluate its performance in the context of robot-assisted prehabilitation. We find that the algorithm learns to predict aspects of the user's behavior as it is given more data, and that it shows strong convergence properties after a small number of iterations.

1. Introduction

Robots are becoming more ubiquitous in everyday life, moving from factory floors to our homes, roads, and workplaces. This shift has the potential to revolutionize the way we think about transportation, healthcare, home care, and many other fields. Personal robots could be used to prepare nutritious meals for users with mobility impairments, perform physical tasks for users with injuries, or provide wellness and social support for older adults.

Despite their great potential, the widespread adoption of robots in personal settings could be hindered by their limited understanding of humans and long-term human preferences. In order to successfully infer the preferences of a user, robots will have to model the desired behaviour of a user through observation. In fact, demonstrating desired behavior is easier than formally specifying a desirable one (Abbeel and Ng (2004); Rothkopf and Dimitrakakis (2011)). Inferring task-dependent goals and preferences of each user will enable better collaboration with humans and faster learning on new unseen tasks (Wirth and Fürnkranz (2013); Christiano et al. (2017)). This is particularly true for several important application areas in robotics such as healthcare, which will require robots to work with people across a multitude of tasks, including providing physical and cognitive support to stakeholders including people with disabilities, clinicians, and caregivers (Riek (2017, 2015); Luxton and Riek (2018); Lee and Riek (2018); Moosaei et al. (2017, 2014)).

Consider an assistive robot tasked with facilitating a rehabilitative therapy regimen with a user. The robot's goal in this case is to make sure that the user completes their

regimen of exercises to speed up recovery. One possible approach would be to determine an optimal goal, such as ensuring the user completes their prescribed stretches each day. In standard reinforcement learning, such a goal is specified as maximizing a pre-defined reward function. However, such directly-specified reward functions fail to take into account the unique preferences of the user. Instead, the robot can observe the user performing a routine alone or with an expert therapist, and infer the user's goals and preferences with regards to that task - for instance, which muscle groups the user wants to target or whether they prefer motions that can be done while sitting down. Then, when the robot is facilitating therapy on its own, it will be better able to use this information to engage in an enhanced, individualized interaction with the user.

There is existing work in preference learning (Evans et al. (2016); Erkin et al. (2010)), which helps set the stage for the current work. Evans et al. (2016) analyses how false beliefs and suboptimal policies can be included in the learning of a preference inference algorithm. Erkin et al. (2010) focused on a healthcare application, where they inferred a patient's preferences for liver transplant based on their health state history.

While promising, this work has several gaps. First, preference learning requires complete knowledge of the environment, which is often impractical in real world robot teaming scenarios. Second, a majority of the preferences are short term in nature and may not endure across time, making it necessary to relearn preferences in the future. Finally, the preferences learned are specific to the tasks in which they are learned, so preferences that may be generalizable must be relearned across each task, no matter how similar the tasks are.

We address these gaps by introducing a novel preference-inference formulation, inspired by the needs of assistive robotics applications, and by demonstrating an algorithm that effectively performs this inference in a real-world scenario. In this formulation, which we call Observational Repeated Inverse Reinforcement Learning (ORIRL), the robot observes the user completing multiple tasks in which the user selects a set of actions.

The robot is given some partial information about tasks, such as what constitutes task completion, but it does not have information about the user's preferences when selecting each action. For example, the robot may observe a user performing a rehabilitation exercise involving shoulder's stretching, and a cooking task where the user needs to reach utensils and mix different ingredients together. In realistic cases, knowledge of the tasks is insufficient to understand and predict the user's behavior, as different users may have different motor skill limitations different preferences about the ordering of steps cooking steps, and many other details. In order to best assist the user, the robot should be able to infer many of these preferences based on observing the user's behavior in other tasks; for example inferring that the user might have shoulder limitations based on the previous stretching exercise might affect the suggestions that the robot might give about the location of cooking utensils which are not easy to reach.

In particular, we do not assume to have complete information about the tasks, including the preference of a user towards each task. The goal of the robot is to infer the user's preferences in a task-independent manner, as well as to understand how these preferences interact with the various tasks to produce the observed behavior. Previous work on inferring users' task-independent preferences makes unrealistic assumptions on how feedback can be obtained, assuming the existence of an expert that provides optimal demonstrations

whenever the robot makes a suboptimal decision (Amin et al. (2017)). In our formulation the robot cannot assume a user will provide such oversight, and must learn task-independent preferences exclusively from observation.

Clinical Relevance Many potential applications of assistive robotics will require understanding human preferences across tasks to a degree that is not currently feasible in real-world environments. Existing preference-inference techniques suffer from unrealistic requirements on the degree of available supervision and interactivity, on the amount of available training data, or on the feasibility of retraining from scratch for each task. Constructing task-independent user-preference models allows us the flexibility to model multiple disparate tasks, while also providing the ability to utilize information learned in previous tasks to quickly generalize to new tasks. This affords an improved ability to handle tasks where it is difficult to gather a large number of human demonstrations, while also allowing faster generalization to new tasks. For example, if the therapy facilitation robot mentioned earlier was tasked with a new objective, such as helping prepare a meal, it could use information learned in the facilitation task to more effectively satisfy the user's preferences in the cooking task. The robot might have learned that the user has difficulty bending over and lifting objects, which can then be transferred to the meal preparation task space.

This information reflects the user's task-independent preferences, and utilizing that information will allow the robot to better provide cooking assistance even with sparse demonstration information - such as by knowing ahead of time to fetch objects stored near the ground for the user. This transfer of knowledge from one task to another (transfer learning), relies on a sufficiently expressive model of a user's task-independent preferences and a minimal description of the task (such as which food is being prepared), Transfer learning allows the robot to effectively perform a personalized version of the task without ever observing a user demonstration for that task (Chao et al. (2011)).

In robotic healthcare assistance, one concrete application of interest for human-robot interaction is in facilitating prehabilitation activities. For instance, "active" breaks, wherein users partake in exercises that strengthen muscles implicated with repetitive strain injury, have been shown to provide significant health benefits (Abdelhameed and Abdel-aziem (2016)). This prehabilitative approach has the potential to offer much greater long-term health benefits as compared to non-active breaks (i.e, breaks where users simply rest affected muscle groups) due to injury-susceptibility reduction. In addition to being a desirable application area in its own right, the prehabilitation setting enables the performance of many tasks with correlated underlying preferences, allowing us to gather long-term data about the choices a user makes, and exposing relationships between choices made on different days. This makes prehabilitation an ideal context for evaluating candidate algorithms for our preference-inference formulation.

Technical Significance The contributions of this paper include: a new preference-learning formulation, the presentation of an algorithm for performing this inference, and a validation of this approach in a realistic, long-term robot-assisted interaction study. The formulation builds upon Repeated Inverse Reinforcement Learning (Amin et al. (2017)) with relaxed assumptions that allow for incomplete information and non-expert users. The proposed algorithm for performing this inference is an application of the maximum-margin framework, one of the most common classes of approaches to Inverse Reinforcement Learn-

ing (IRL). We demonstrate that, using our new max-margin approach, we can successfully infer a user's task independent preferences and predict features of a user's actions for unseen tasks, facilitating personalized workflows for each user.

The outline of the paper is as follows. First, we discuss the ideas and existing literature behind preference learning in Secion 2. Then, in Section 3, we formalize the aforementioned proposal for incomplete-information task-independent preference inference, and introduce the new max-margin algorithm. Next, we present an empirical evaluation of the algorithm (see Section 4). Our Results, described in Section 5, show that we are able to predict features of unseen tasks and infer the user's preferences across different tasks. Finally, we discuss the implications of these findings for the robotics community in Section 6.

2. Preference Learning

Preference learning is a subfield of machine learning concerned with learning individuals' proclivities. This allows a system to make sensible predictions based on the users' historical choices. Some examples in AI include recommendation systems which use other users' preferences or products' features to recommend products that the user might like Schafer et al. (2007), adaptive user interfaces which change according to the user's preferences, and autonomous agents which adjust their suggestions based on previous responses by the user (Wirth and Fürnkranz (2013)).

In robotics, examples of preference learning include robots that collaborate directly with humans (Nikolaidis et al. (2017); Saunders et al. (2016); Munzer et al. (2017)). For example, Munzer et al. (2017) show how learning the user's preferences can be beneficial during a toolbox making task. In this task, the robot passed the human the pieces needed to build a toolbox and it adapted at each iteration based on whether it provided the right piece to the human or not.

There are several methods used to infer preferences from observed user behavior.

One approach is recommender systems, which attempt to infer which products a user will like based on how they have felt about other options (Schafer et al. (2007)). Recommender systems typically use collaborative filtering or content filtering to perform this inference.

Collaborative filtering approaches tackle this problem by imposing a similarity score on users, which is based on whether users expressed similar responses for the same products. Unfortunately, this approach requires having a large corpus of user data. In fact predicting how a user will feel about a new product requires a sufficiently large number of similar users who have themselves provided feedback on the new product. By comparison, content filtering seeks to build a model of the user's preferences in relation to features of another entity (Schafer et al. (2007)).

Finally, another approach worth mentioning is meta-learning. Finn et al. (2017) has shown how meta-learning can successfully build algorithms that are model-agnostic and which are applicable to a wide variety of tasks.

While these methods offer great results, they are ill-suited toward longitudinal robot preference learning as they require a large volume of labeled data, only work in the short term, and rely on problems which depend only on the present state, a property also known as the markov property. Many applications in robotics are in longer-term interactive settings in which the user's choices influence the future state and are influenced by the past.

Furthermore in HRI, since the user interacts with the environment and the robot in a complex manner, it is more relevant to build a predictive model based on inverse reinforcement learning (IRL) (Ng and Russell (2000)).

IRL uses observation to derive the reward function, and hence also preferences, of a user (Argall et al. (2009); Chernova and Thomaz (2014)). IRL has been used with great success in many robotic applications (Hadfield-Menell et al. (2016); Abbeel and Ng (2004); Jin et al. (2015)) Abbeel and Ng (2004) spearheaded the efforts in IRL by demonstrating how a car in a simulation can learn a reward function simply by observing an expert. Further work by Hadfield-Menell et al. (2016) explores the implications of IRL not only in an isolated environment but also in a cooperative one where human actors change their behaviour when interacting with artificial agents.

Within IRL, there are three main approaches: maximum-margin methods, feature expectation matching, and methods that treat the policy as being parameterized by the reward function (Abbeel and Ng (2011)). Max-margin methods address the problem by optimizing for a reward function that makes the expert's observed policy as good as possible compared to alternatives, while also selecting for simpler rewards (Ratliff et al. (2006)). Feature expectation matching attempts to find a policy that generates features similar to those generated by the expert's policy, without emphasising inference of the true reward function. The last class of methods assume the expert's policy is a function of the reward, allowing solution with methods such as gradient descent and approximate Bayesian inference (Rothkopf and Dimitrakakis (2011)).

While IRL has been successful in many domains, it traditionally involves inferring a single reward function for a single task, and does not allow robots to take advantage of similarities to generalize to new tasks. This is because such approaches model a task's reward function as a single atomic entity, independent of the reward functions for other tasks. Because the robot observes different behavior when the user completes different tasks, it therefore must either throw away all known information and model a new task from scratch, or it must attempt to model a single unified reward function that explains all behavior in all tasks. The former method suffers from inefficiency and the inability to build up general models of the user's preferences over time, while the latter suffers from incredibly high complexity and the need for huge amounts of demonstrations over a wide variety of tasks. Thus, we must use a cross-task method which is able to transfer knowledge gained for historical tasks on new unseen ones.

There has been some existing work on extending IRL outside the single-task single-reward model. Inferring multiple reward functions has been studied in contexts where observations are generated from multiple (unknown) experts (Choi and Kim (2012)), as well as the case where multiple reward functions are stochastically interchanged (Slivkins and Upfal (2008)). However, these cases are different than the example described in Section 1, in which a single user is observed completing multiple different tasks and we must build up a unified and coherent model of their overarching preferences.

Rothkopf and Dimitrakakis (2011) present a Bayesian formulation for inferring reward functions for multiple related tasks, which assumes tasks are drawn randomly from a prior which must be inferred. This is along the same line as other work which attempts to infer priors in Bayesian settings (Evans et al. (2016)). In contrast, in our work we do not make any assumptions about the possible range of tasks; instead we focus on how the user's individual

preferences affect a given task. For example, we would like to infer what actions and how much time a user will spend on a new task based on their previous global preferences. In addition, we utilize existing partial information about task structure (for instance, that a cleaning task involves a higher reward after the room is cleaned) to improve the inference.

Recent work on repeated inverse reinforcement learning (RIRL) has also extended IRL to settings in which a user is observed performing different tasks (Amin et al. (2017)). The goal in RIRL is the same as the goal in our domain: to infer user-specific, task-independent reward terms that the user attempts to satisfy in all settings, For example, Amin et al. (2017) focused on autonomous driving and inferred user preferences for safety that may not be explicitly specified in the given task rewards.

However, work by Amin et al. (2017) assumes that the agent possesses complete prior information not only about task reward structure, but also about the interactions between the various task rewards and the user's overall preferences, which is unrealistic in most applications of interest. For example, the robot may know that a particular user just left rehabilitation and has some shoulder mobility issues, and it may have some partial information about what it means to complete a cooking task. However, the robot does not assume to know ahead of time how the user weights the task reward relative to their overall preferences; some users may be able to reach items on a high shelf when cooking, while others are not, and this uncertainty only compounds as more tasks are added. Furthermore, Amin's work was not tested in a real-world environment with a physical robot, limiting the potential impact that their findings might have in HRI.

The existing literature covers a large range of diverse applications as described above, but there is a gap in modelling preferences in the repeated setting when only partial information is known about task rewards. Our work addresses this gap by introducing a method for preference inference in the more realistic ORIRL setting with only partial task-reward knowledge. Furthermore we focus on a real-world scenario where a robot interacts with human actors. The proposed method uses max-margin learning to learn the task dependent reward functions in combination with a global reward function which is affected by the user's task independent preferences.

We evaluate our method within a prehabilitation scenario, in which a user may have multiple tasks recommended to them by a healthcare professional, such as a series of exercises to prevent injury. This is an excellent application domain for our method, as one of the roadblocks to successful health behavior change is patient adherence (e.g., continuing to do exercises even when losing interest). Our prior work (Riek (2017); Adamson et al. (2016); Riek (2015); Lee and Riek (2018)), and work by others, suggests a personalized approach to health technology will facilitate greater adherence (Ludden et al. (2015); Hermsen et al. (2016); Mohr et al. (2014)). Having a greater understanding of a user's preferences in this domain may enable us to provide better autonomous support.

In this domain, a robot knows at most the reward structures of the given tasks, but not their magnitude; it does not in general know the relative levels of motivation that the user will feel towards completing each task relative to other preferences. In addition, the multi-task inference method by Amin et al. (2017) has only been tested in simulation, with data generated by simulated experts who obey the modelling assumptions in the respective approaches. We are interested in empirical validation of multi-task preference inference with human users, in order to test modelling assumptions as well as algorithmic performance.

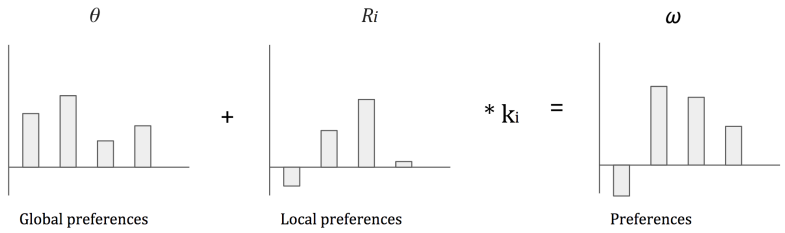

Figure 1: Decomposition of a task-specific preference. R_i is presumed to be known in advance, k_i and θ are not.

3. Observational Repeated Inverse Reinforcement Learning

Based on the classes of IRL methods described above, we introduce a new preference-inference formulation: Observational Repeated Inverse Reinforcement Learning, or ORIRL. We would like to highlight the difference between observed behavior and preferences in this context. We assume that the observable behavior that we record is partially determined by some hidden variables or a user's preferences. ORIRL aims to learn the preferences of a user which are not directly observable. We then propose a max-margin learning approach to infer a global reward function which combines the task-dependent reward functions with the user's task independent preferences. Some IRL approaches such as feature expectation matching attempt to infer enough about the user to predict their actions, without emphasis on learning the true underlying preference function. However, in many applications the true preference function itself may encode useful information. For instance, in rehabilitation applications, a clinician reward function may include relevant information about health and pain levels. In our case, we are trying to infer task-independent rewards to generalize the user's preferences to new tasks, which requires a model of the reward. Max-margin approaches are well-suited to our use-case, as they attempt to learn the underlying preference function directly (Abbeel and Ng (2011)).

3.1. ORIRL Formulation

We model tasks and environments in terms of Markov decision processes (MDPs). An MDP is defined by the state space S, action space A, transition dynamics $P : S \times A \to \Delta(S)$, initial state distribution $\psi \in \Delta(S)$, discount factor $\gamma \in [0, 1)$, and reward $Y : S \to \mathbb{R}$. In addition, an agent's strategy in an MDP is denoted by their policy $\pi : S \to \Delta(A)$, which determines the (possibly stochastic) action to take in any given state. The state inhabited at timestep t is denoted as s_t.

In many cases of interest, a full Markovian state-space formulation for the MDP is necessarily large or infinite, leading to problems with tabular reinforcement-learning methods (Sutton and Barto (1998)). This is commonly addressed through a mapping ϕ from states to low-dimensional state features. We can extend this to a function μ that maps policies to the expected exponentially-discounted sum of state features under that policy, known as feature expectations (Abbeel and Ng (2004)). That is, $\mu(\pi) = \mathbb{E}_\pi \sum_i \gamma^i \phi(s_i)$. The reward

function is commonly assumed to be approximately linear in these state features, so that the reward is parameterized by a vector ω. Thus, the expected discounted reward from a policy π is just the dot product between ω and $\mu(\pi)$.

In ORIRL, the true reward for a given task is a weighted sum between a task-independent reward term and a task-dependent reward term. If we use a linear approximation ω of the reward function, this means for task i that $\omega_i = \theta + k_i R_i$, where θ is the task-independent reward term, R_i is the reward for task i, and k_i is the weighting for task i. We define K as the vector of scalars $K = [k_1, k_2, \ldots, k_n]$, where n is the number of tasks.

3.2. Max-Margin ORIRL

The original IRL problem as defined by Abbeel and Ng (2004) was to infer a reward function that matched the observed behavior, assuming the expert behaves optimally. However, this leads to problems of ambiguity, as there are many reward functions that would explain any given observation. The problem is exacerbated if the expert's demonstration is suboptimal, as will be the case if the expert is a human acting in a complex environment, such as on a busy road or in an operating room.

In order to overcome the challenges inherent in the original IRL formulation, Ratliff et al. (2006) proposed the use of max-margin methods. In this formulation, we first turn the feasibility problem into one of optimization - namely we minimize the L2 norm of the weight vector, as a form of complexity penalty, subject to the optimality constraints. This means we will find a unique solution, but since we are minimizing the L2 norm, the reward function that always returns 0 will be selected for any given set of observations, despite the fact that humans rarely have exactly no preference over any possible state.

To deal with this class of problem, maximum-margin methods incorporate the intuition that the expert's policy is likely significantly better than alternatives. Structured-prediction maximum-margin methods require the expert's observed policy to not only match, but beat all other policies by an amount that scales with a measure of difference between the policies. This encodes the idea that "nearby" policies may be nearly as good as the expert's policy, but "faraway" policies are probably worse than the optimal policy by a larger amount.

This still leaves the unsatisfactory assumption that the expert's policy is exactly optimal with respect to the hidden reward. While it may be a good heuristic that the expert's policy is significantly better than others, this may not always be the case. Maximum-margin methods handle this concern by including the "slack" variable, ξ, that can allow policies to be close to or better than the expert's policy, at some cost in the optimization term. Thus, the full optimization problem becomes one of solving:

$$\min_{\omega,\xi} ||\omega||_2^2 + C\xi \tag{1}$$

$$\text{s.t.} \omega^\top \mu(\pi^*) \geq \omega^\top \mu(\pi) + m(\pi^*, \pi) - \xi \qquad \forall \pi \tag{2}$$

Where π^* is the expert's (near) optimal policy, ω represents the weights for the reward function, and $m(.)$ is a distance function which compares the optimal policy π^* and the alternative policy π.

We can additionally extend this to multiple MDPs that share the same reward function by using:

$$\min_{\omega,\xi_i} ||\omega||_2^2 + C\sum_i \xi_i \tag{3}$$

$$\text{s.t.}\omega^\top \mu(\pi_i^*) \geq \omega^\top \mu(\pi_i) + m(\pi_i^*, \pi_i) - \xi_i \qquad \forall i, \pi_i \tag{4}$$

The optimization term is quadratic and the constraints are linear, allowing efficient solutions for a given number of constraints using quadratic programming. However, there is a constraint for every possible policy, which may be large or infinite. Ratliff et al. (2006) address this by modifying the form of the constraints and using subgradient methods, but a simpler alternative is iterative constraint generation of the form used by Abbeel and Ng (2004).

With these existing formulations, modeling multiple MDPs can be done in one of two ways. We can assume all rewards are independent and solve distinct instances of equation 1, or conversely we can assume that all rewards are identical and solve the combined equation 3. RIRL proposes a third, hybrid approach where there is a shared task-independent reward term, as well as separate task-dependent reward terms. In ORIRL, the direction of the task-dependent reward terms are known, but their magnitudes are not. This encodes the situation in which we have some partial prior information about which aspects of a task are relevant and distinct from other tasks, but we do not know how strongly any given user will weight each task-dependent reward relative to their underlying task-independent preferences.

In order to extend equation 3 so that it solves ORIRL-style problems, we must make three changes. First, the single reward term ω in the constraints becomes the combination of task-dependent and task-independent rewards $\theta^* + K_i R_i$. Second, instead of minimizing over the full reward ω in the minimization term, we only minimize over θ^*, the task-independent portion of the reward. Finally, we must add a complexity penalty for K in the minimization term in order to re-establish the desirable properties of max-margin described above. The L2 norm fulfills this function, but unlike θ^* we normalize K towards a positive constant vector $\hat{d}$ instead of $\hat{0}$ to take into account a prior that users are more likely to have positive weights for task-dependent terms. We are left with the following equation:

$$\min_{\theta^*,K,\xi_i} ||\theta^*||_2^2 + B||K - \hat{d}||_2^2 + C\sum_i \xi_i \tag{5}$$

$$\text{s.t.}(\theta^* + K_i R_i)^\top \mu(\pi_i^*) \geq (\theta^* + K_i R_i)^\top \mu(\pi_i) + m(\pi_i^*, \pi_i) - \xi_i \qquad \forall i, \pi_i \tag{6}$$

With this formulation, the optimization remains quadratic and the constraints remain linear, allowing an efficient solution through quadratic programming methods.

4. Method Validation

4.1. Experimental Context

The motivating context for our work is healthcare. We are particularly interested in methods to infer preferences from users who have been given non-binding therapeutic advice from clinicians, and then must determine how to balance activity engagement given their own

Table 1: Prehabilitative activities for RSI prevention

Pathology	Forward Head Posture	Kyphotic Posture	Wrist Tension & Irritation	Pelvic Tilt	Prolonged Hip & Knee Flexion	Limited Hallux Dorsiflexion
Body Part	Neck	Pecs & Traps	Wrists	Back	Hips & Hamstrings	Toes
Non-Standing Variation	Lying	Lying	Lying	Lying	Lying (Hips) & Sitting (Hamstrings)	Sitting

preferences. This is a particularly interesting application space for two reasons. First, this is a common scenario in ambulatory care - clinicians provide proscriptive advice which may or may not be followed by users, or may be only followed for a short time, etc. However, if we can infer their preferences in these scenarios and build interactive, adaptive systems based on them, it can have a substantial practical impact – tailored, individualized treatment plans are far more likely to be adhered to (Ludden et al. (2015); Hermsen et al. (2016); Mohr et al. (2014)). Second, the clinician's advice can influence the choices a user might make, and hence the reward of corresponding tasks. Because of this, our system will not only depend upon the user's preferences, but also the expert's advice.

Thus, we evaluate our methods in a rehabilitation setting across multiple activities, in which participants receive written advice from a physical therapist, and then interact with a robot facilitator to choose a set of activities to perform. In our study, participants participated in a week-long, twice-a-day prehabilitation activity session with a robot (See Fig. 2(b)). The sessions lasted for 10 minutes, wherein the robot solely provided instruction and structure. In each session, participants would be greeted by the robot and presented with advice from a licensed physical therapist (randomized by day and participant). The advice was related to the different categories of activities to bias activity selection. The activity categories consisted of eight different body parts, each with standing or not standing versions, for a total of 16 unique prehabilitative activities (See Table 1).

Participants then navigated to a main screen where they could choose the activity they wanted to perform. They would then view an instruction screen which contained a video example of the activity and brief explanatory text. Once ready, participants would then begin the activity, at which point a session timer would start. Participants could perform an activity for as long as they liked. Once participants finished, they would navigate back to the main activity selection screen which would pause their session timer. The robot would offer a break if the time had reached five minutes; if it exceeded 10 minutes the robot would end the session.

4.2. Participants

We recruited four participants to participate in our study (three females and one male with ages ranging from 20-24). They were primarily undergraduate and graduate students who

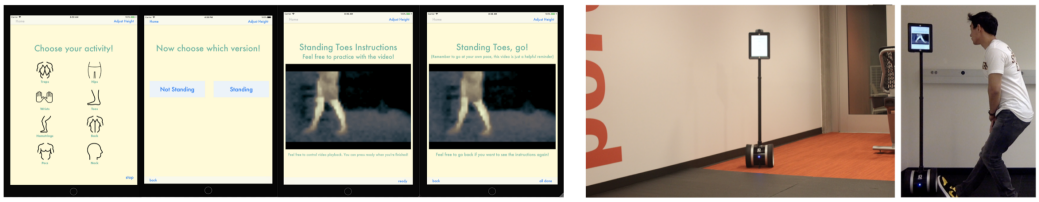

Figure 2: (a) - From left to right: The main activity selection menu, the standing/not standing version selection screen, a sample instructions screen, and a sample of the screen that plays while the user performs an activity. (b) - Left: the robot in a real-world setting. Right: A user interacting with the robot.

spend a majority of their time sitting, working at computers. No participants reported any pre-existing health conditions which would affect their participation.

4.3. Exercise Choices

Based on feedback from a licensed physical therapist, we focused on activities that would reduce susceptibility to repetitive stress injury (RSI) in office workers. The activities in our study focused on eight body parts, with a standing version and a non-standing version (either seated or lying down) for each, yielding a total of 16 exercises. Table 1 lists the six major conditions we focused on improving, and their corresponding eight body parts. For brevity we will not describe these conditions in detail; readers can consult physical medicine texts such as by DeLisa et al. (2005) for detail.

To emulate expert advice given for multiple tasks, there were six sets of advice provided by a therapy expert, targeting each condition. Each set of advice referred to a known expert-specified reward function over state features. We emphasize simple, easy to specify reward functions that give a positive reward when the task is completed and no reward everywhere else. For instance, a set of advice targeting neck posture would have a positive reward only when designated neck stretches and exercises are completed. Because we want to learn the preferences of the users, we do not assume to know ahead of time to what extent the users are influenced by the therapist's advice, and users are encouraged to perform whichever activities they want to perform regardless of the advice.

4.4. Equipment and Interaction Design

In our study, we utilized a Double robot, and two small exercise mats (See Fig. 2 (b)). The robot ran a custom application designed specifically for the study. In accordance with best practices in the health behavior change community (Hermsen et al. (2016); Notthoff and Carstensen (2014); Ybarra et al. (2014)), all messaging in the application was delivered in an positive manner, e.g., "Welcome [User]! It's so great to see you again!" and "You did great, I can't wait to see you again!".

A robot's morphological and behavioral features have been shown to have a significant impact on attentional capture (Li et al. (2015)) and user engagement, which can affect both the quality of data gathered and the algorithmic performance that depends on that data.

For this reason, the robot's height was dynamically adjusted at certain parts of the study to promote greater engagement. At the start of the study, the robot would be initially set to its minimum height (about three feet from the ground) to simulate that the robot was asleep/off. Once the user began a session, the robot would raise its display to its maximum height, which corresponded to eye level for our participants (about five feet from the ground). When choosing non-standing activities, the robot would lower itself back to its minimum height to help maintain a comfortable viewing angle. Once the participant completed this activity, the robot would then raise itself back to its maximum height. To ensure a comfortable interaction, the users had the option of adjusting the height manually at any point in the session.

Finally, to simulate cooperative behavior, the robot would provide the participant with an encouraging remark when the user completed an activity after the five minute mark of the study. The robot would then and inform the participant that it was itself motivated to take a stretching break. It would redisplay its smiling face, and turn 360 before prompting the user that it was ready to continue. This was the only time the robot moved, and its kickstands were deployed immediately after to ensure user safety.

4.5. Procedure

At the beginning of the study, participants participated in an orientation session. An experimenter taught them how to properly interface with the robot and performed demonstrations of the 16 different prehabilitation exercises of the study. Participants participated in approximately two sessions per day, five times per week (Monday - Friday), for one week.

Prehabilitation sessions consisted of participants entering the experimental area, where the interactive robotic facilitator (see Fig. 2(b)) would take them through the day's activities, as explained in Figure 2(a).

To mitigate possible ordering effects, exercises presented on the robot's display at the activity selection screen was randomly chosen for each session. Additionally, to minimize gender and/or cultural bias, all videos and pictures were presented as thermal images or silhouettes.

In our experiment we used the following measures: the group of the selected action, its duration, and whether the user performed the seated or standing variation of the activity. The action's group is linked to a unique identifier and it tells us which muscle group was activated during the exercise. The action's duration was measured in seconds and was also used to calculate the total duration of a session. Finally, the action's sit/stand variation was recorded as a boolean variable.

During each activity session, each participant read the session's prompt and chose a sequence of actions to perform. To emulate a home care setting, no human facilitators or other participants were present during the sessions; each participant's activities were classified by the interactive visual dialogue system on the robot's touch screen. Each session generated data consisting of the participant's ID, the session ID, and the sequence of actions the participant selected, which includes the three measures described above. Each action is then encoded as a vector that combines a one-hot encoding of its targeted muscle group, its discretized duration, and whether it was the sitting or standing variation. In addition, the session is linked to the advice given in that session.

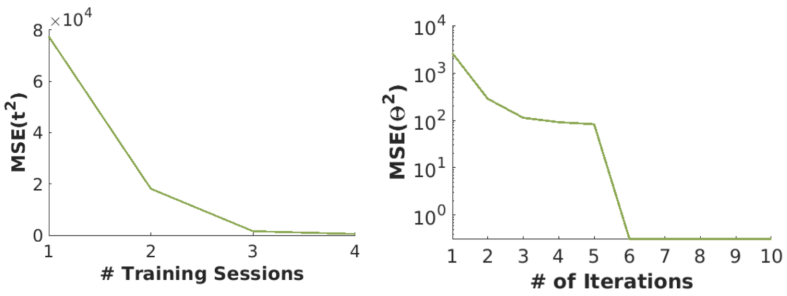

Figure 3: (a): Averaged mean squared error between max-margin ORIRL's prediction of the next session's duration and the empirical duration of that session, plotted against the number of sessions the system has already observed for a user.
(b): Convergence measured by the averaged MSE of θ at each iteration of ORIRL, relative to the final estimate.

This data is provided to ORIRL as a set of feature expectations generated by the participant's actions. That is, the algorithm models the set of sessions as an MDP, and observes the empirical discounted sum of state features resulting from the participant's choices.

The features provided to the algorithm approximate the relevant information about the actions and state. In addition to the action features described above, the state was mapped by concatenating the following vectors: $(V_i, {V_i}^2, V_i V_{i-1}, X_i, {X_i}^2, X_i X_{i-1}, T_i$, where V_i corresponds to an array that keeps count of the variations (standing or sitting) up to state i, X_i corresponds to an array that counts the muscles targeted up to state i and T_i corresponds to a one-hot encoding vector which discretizes the total duration of the session up to state i. These features were selected to give ORIRL sufficiently expressive representations with which to model preferences, without biasing the algorithm by providing extra prior data. For instance, the algorithm initially has no prior expectation that participants might tend to prefer to take varied actions in each session, and must learn such associations from observation.

5. Results

Ground truth data for user preference models is infeasible to obtain, so we measure the performance of the algorithm by its predictive power and its convergence properties.

To see how the predictive power of the model changes as it is given more sessions to train on per user, we gave the model the first n sessions for each user and measure how well it predicts aspects about session $n+1$. Namely, we measured the mean squared error between the time features of the user's actions with the time features of the policy optimal with respect to the inferred preferences. Unlike the other features, time is continuous, which allows a straightforward error analysis. The results in Figure 3(a) show how error decreases as more sessions are provided to train on, with diminishing returns.

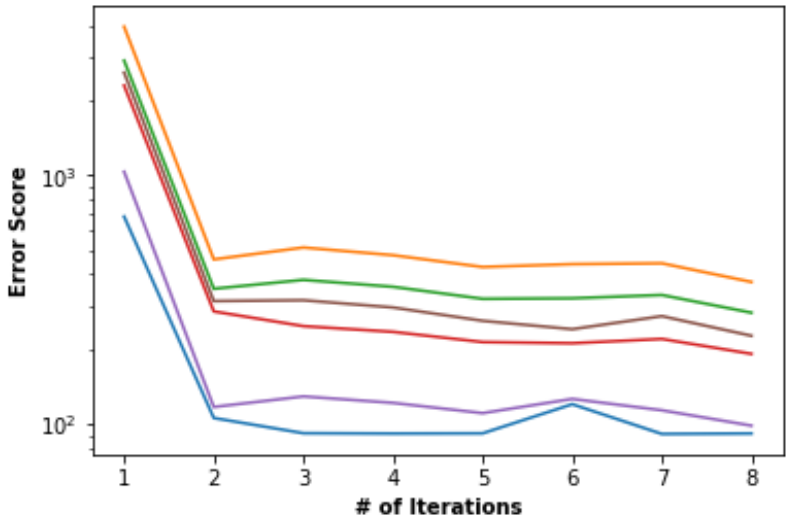

Figure 4: Averaged prediction error scores per task. Prediction error scores are based on the negative log-likelihood of observing the user-generated dataset under the inferred preferences at each iteration.

To track the convergence rate of the algorithm, we run the algorithm until it converges to an estimate θ of the user's task-independent preferences. We then observe the mean squared error of the estimate θ_i after iteration i with respect to the final converged θ. Figure 3(b) plots this convergence, in which we observe a rapid convergence in the first few iterations, until after iteration 6 the weight values for the inferred task-independent preferences only undergo small adjustments.

Finally, in order to get a more general estimate of predictive power beyond just predicting the time variable, we define an error metric based on the negative log likelihood of observing the user data under the inferred preferences. In the vein of Ramachandran and Amir (2007), we assume a generative model where a user selects actions based on an exponential distribution over the Q-values of that state-action pair. That is: $P(a|s,R) = \frac{1}{Z}e^{\alpha Q^*_R(s,a)}$, where α is a hyperparameter weighting how "noisy" the user's choices are, $Q^*_R(s,a)$ is the Q function that gives the expected discounted sum of rewards from following the optimal policy after taking action a in state s, and Z is a weighting term to make the probabilities for all actions sum to 1.

Given this generative model, we can compute the error score as the negative log likelihood of observing the user data under the total reward function inferred by the ORIRL model. That is, for each estimated total reward, we define a probability function over the set of actions selected in the observed states, and rate the model based on how much probability mass its inferred reward function assigned to the set of true user actions. Figure 4 plots this score per iteration for each task the model is inferring. The first iteration bases its predictions off a random policy, and we observe a very significant improvement in predictive

power after just one additional iteration. Afterwards, the error score continues to steadily improve (note the logarithmic scale on the Y-axis means improvement slows over time).

6. Discussion

The main contributions of this paper are the introduction of the ORIRL framework for realistic task-independent preference inference and the presentation of a max-margin algorithm that efficiently performs this inference. Unlike existing approaches, the ORIRL framework makes no assumptions about the ability of users to provide interpretable feedback to the robot in the form of corrective demonstrations. Instead, it relies completely on observational data gathered as the user completes tasks; the ease of gathering this kind of training data make it a much more suitable choice for an assistive robot working with users who are not robotics experts. Overall, the results we obtained with max-margin suggest that it is possible to learn task-independent preferences in this framework (See Figs. 3(a) and 4) in a small number of iterations (See Fig. 3(b)).

The algorithm achieves a sensible time inference accuracy on unseen sessions (see Figure 3(a)), suggesting that the inferred user preference model is able to successfully capture relevant information about the user's task-independent preferences. Moreover, max-margin ORIRL is able to successfully learn the global preferences of a user across different real-world tasks (see Fig. 4). For example, ORIRL learned that one user prefers to stand when performing an exercise, and also prefers activities involving the back. It can then transfer these general inferences to better predict the user's behavior even on tasks it has never seen the user perform before.

Inferring user preferences in observational settings, as ORIRL does, will greatly benefit robotics by promoting more personalized long term interactions between robots and humans. A robot with a stronger understanding of the user will be able to better handle uncertainty introduced by new situations. For example, they can rely upon previous knowledge in new situations, as was demonstrated when the algorithm predicted features of unseen tasks (See Fig. 3(a)).

While ORIRL's observational nature means it relies on weaker assumptions than prior work, it does assume that the robot can infer which task the user is performing. This is a sensible assumption, as it can often be done for distinctive tasks (Stauder et al. (2014); Arbab-Zavar et al. (2014)). However, even in cases where there is no easy way to identify tasks *a priori*, Babes et al. (2011) show that clustering approaches can be used to group demonstrations together based on the task that generated them.

In the future, we seek to improve the features used to map each action taken by the user. We could improve the current feature set by learning the features themselves, as in deep neural networks, but this requires much more data than we had the resources to collect. One possible way to tackle the lack of data could be the application of deep generative model learning on a pre-existing smaller dataset to generate new user data using a different Gaussian distribution. Another option is to craft stronger priors, so that the algorithm only has to learn how each user differs from some "average", rather than learning preferences from scratch. Construction of such priors would likely require significantly fewer data points than would be required for a fully-learned feature mapping.

We also plan to explore new applications of ORIRL. This work could be extended to content filtering in more interactive recommender systems, such as movie recommenders. Instead of modeling a user's movie preferences as a static mapping from movies to ratings, the system could learn preferences as task-independent rewards like RIRL and ORIRL, enabling such a system to better take context into account and different recommendations that adapt to the user's actions and environment, including interactions with the recommender itself.

The ORIRL framework described in this paper is a step forward towards better human understanding. This has impactful and far-reaching applications in robotics, and especially in the field of HRI. The ability to model a user's preferences across different situations over a long period of time will enable more personalization and improved collaboration between the user and robot. This is especially important now as robots are becoming more involved in our daily routines, and commonly operate inside our homes. It is our hope that future work continues to expand these ideals and ensure that personal robots are truly personal.

Acknowledgments

The authors would like to thank Michelle Coombs, PT, for her assistance in the development of the prehabilitation routine, and Monique F. Narboneta for designing the study robot's interface layout.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 1–8, 2004. ISSN 0028-0836. doi: 10.1145/1015330.1015430.

Pieter Abbeel and Andrew Y Ng. Inverse reinforcement learning. In *Encyclopedia of machine learning*, pages 554–558. Springer, 2011.

Abeer Ahmed Abdelhameed and Amr Almaz Abdel-aziem. Exercise training and postural correction improve upper extremity symptoms among touchscreen smartphone users. *Hong Kong Physiotherapy Journal*, 35:37–44, 2016.

M. Adamson, L.D. Riek, and D. Liston. Neurorehabilitation applications: Expanding the horizons for longitudinal assessments and non-invasive treatments in mild and moderate tbi. *American Academy of Physical Medicine and Rehabilitation*, 2016.

Kareem Amin, Nan Jiang, and Satinder Singh. Repeated inverse reinforcement learning. *Advances in Neural Information Processing Systems*, pages 1813–1822, 2017.

Banafshe Arbab-Zavar, John N. Carter, and Mark S. Nixon. On hierarchical modelling of motion for workflow analysis from overhead view. *Machine Vision and Applications*, 25 (2):345–359, 2014. ISSN 09328092. doi: 10.1007/s00138-013-0528-7.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Monica Babes, Vukosi Marivate, Kaushik Subramanian, and Michael L Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 897–904, 2011.

Crystal Chao, Maya Cakmak, and Andrea L Thomaz. Towards Grounding Concepts for Transfer in Goal Learning from Demonstration. *2011 IEEE International Conference on Development and Learning (ICDL)*, 2011.

Sonia Chernova and Andrea L Thomaz. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(3):1–121, 2014.

Jaedeug Choi and Kee-Eung Kim. Nonparametric bayesian inverse reinforcement learning for multiple reward functions. In *Advances in Neural Information Processing Systems*, pages 305–313, 2012.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 2017.

Joel A DeLisa, Bruce M Gans, and Nicholas E Walsh. *Physical medicine and rehabilitation: principles and practice*, volume 1. Lippincott Williams & Wilkins, 2005.

Zeynep Erkin, Matthew D Bailey, Lisa M Maillart, Andrew J Schaefer, and Mark S Roberts. Eliciting Patients' Revealed Preferences: An Inverse Markov Decision Process Approach. *Decision Analysis*, 7(4):358–365, 2010. doi: 10.1287/deca.1100.0185.

Owain Evans, Andreas Stuhlmüller, and Noah D. Goodman. Learning the Preferences of Ignorant, Inconsistent Agents. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 323–329, 2016.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative Inverse Reinforcement Learning. *Advances in neural information processing systems*, 2016.

Sander Hermsen, Jeana Frost, Reint Jan Renes, and Peter Kerkhof. Using feedback through digital technology to disrupt and change habitual behavior: a critical review of current literature. *Computers in Human Behavior*, 57:61–74, 2016.

Ming Jin, Andreas Damianou, Pieter Abbeel, and Costas Spanos. Inverse reinforcement learning via deep gaussian process. *arXiv preprint arXiv:1512.08065*, 2015.

Hee Rin Lee and Laurel D Riek. Reframing assistive robots to promote successful aging. *ACM Transactions on Human Robot Interaction*, 2018.

Alvin X Li, Maria Florendo, Luke E Miller, Hiroshi Ishiguro, and Ayse P Saygin. Robot form and motion influences social attention. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 43–50. ACM, 2015.

Geke DS Ludden, Thomas JL van Rompay, Saskia M Kelders, and Julia EWC van Gemert-Pijnen. How to increase reach and adherence of web-based interventions: a design research viewpoint. *Journal of medical Internet research*, 17(7), 2015.

D. Luxton and L.D. Riek. Ai and robotics in rehabilitative psychology. *Handbook of Rehabilitation Psychology*, 2018.

David C Mohr, Stephen M Schueller, Enid Montague, Michelle Nicole Burns, and Parisa Rashidi. The behavioral intervention technology model: an integrated conceptual and technological framework for ehealth and mhealth interventions. *Journal of medical Internet research*, 16(6), 2014.

Maryam Moosaei, Michael J Gonzales, and Laurel D Riek. Naturalistic pain synthesis for virtual patients. *International Conference on Intelligent Virtual Agents*, pages 295–309, 2014.

Maryam Moosaei, Sumit K Das, Dan O Popa, and Laurel D Riek. Using facially expressive robots to calibrate clinical pain perception. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 32–41, 2017.

Thibaut Munzer, Marc Toussaint, and Manuel Lopes. Preference learning on the execution of collaborative human-robot tasks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 879–885. IEEE, 2017.

AY Ng and SJ Russell. Algorithms for inverse reinforcement learning. *Icml*, 2000.

Stefanos Nikolaidis, Swaprava Nath, Ariel D. Procaccia, and Siddhartha Srinivasa. Game-Theoretic Modeling of Human Adaptation in Human-Robot Collaboration. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pages 323–331, New York, New York, USA, 2017. ACM Press. ISBN 9781450343367. doi: 10.1145/2909824.3020253.

Nanna Notthoff and Laura L Carstensen. Positive messaging promotes walking in older adults. *Psychology and aging*, 29(2):329, 2014.

Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. *Urbana*, 51(61801):1–4, 2007.

Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. Maximum margin planning. *International conference on Machine learning - ICML '06*, (23):729–736, 2006. ISSN 17458358. doi: 10.1145/1143844.1143936.

Laurel D Riek. Robotics technology in mental health care. In *Artificial Intelligence in Behavioral and Mental Health Care*, pages 185–203. Elsevier, 2015.

Laurel D Riek. Healthcare robotics. *Communications of the ACM*, 60(11):68–78, 2017.

Constantin A. Rothkopf and Christos Dimitrakakis. Preference elicitation and inverse reinforcement learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6913

LNAI, pages 34–48. Springer, Berlin, Heidelberg, 2011. ISBN 9783642238079. doi: 10.1007/978-3-642-23808-6_3.

Joe Saunders, Dag Sverre Syrdal, Kheng Lee Koay, Nathan Burke, and Kerstin Dautenhahn. 'Teach Me-Show Me'-End-User Personalization of a Smart Home and Companion Robot. *IEEE Transactions on Human-Machine Systems*, 46(1):27–40, feb 2016. ISSN 21682291. doi: 10.1109/THMS.2015.2445105.

J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

Aleksandrs Slivkins and Eli Upfal. Adapting to a Changing Environment: the Brownian Restless Bandits. *COLT*, 2008.

Ralf Stauder, Aslı Okur, Loïc Peter, Armin Schneider, Michael Kranzfelder, Hubertus Feussner, and Nassir Navab. Random forests for phase detection in surgical workflow analysis. *Lecture Notes in Computer Science*, pages 148–157, 2014. ISSN 16113349. doi: 10.1007/978-3-319-07521-1_16.

R S Sutton and a G Barto. Reinforcement learning: an introduction. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 9(5), 1998. ISSN 1045-9227. doi: 10.1109/TNN.1998.712192.

Christian Wirth and Johannes Fürnkranz. Preference-Based Reinforcement Learning : A preliminary survey. *ECML PKDD 2013 - Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards*, 2013.

Michele L Ybarra, Jodi Summers Holtrop, Tonya L Prescott, and David Strong. Process evaluation of a mhealth program: Lessons learned from stop my smoking usa, a text messaging-based smoking cessation program for young adults. *Patient education and counseling*, 97(2):239–243, 2014.