Learning Human Objectives from Sequences of Physical Corrections

Mengxi Li[1], Alper Canberk[1], Dylan P. Losey[2], Dorsa Sadigh[1]

***Abstract*— When personal, assistive, and interactive robots make mistakes, humans naturally and intuitively correct those mistakes through physical interaction. In simple situations, one correction is sufficient to convey what the human wants. But when humans are working with multiple robots or the robot is performing an intricate task often the human must make *several* corrections to fix the robot's behavior. Prior research assumes each of these physical corrections are *independent* events, and learns from them one-at-a-time. However, this misses out on crucial information: each of these interactions are *interconnected*, and may only make sense if viewed together. Alternatively, other work reasons over the *final* trajectory produced by all of the human's corrections. But this method must wait until the end of the task to learn from corrections, as opposed to inferring from the corrections in an online fashion. In this paper we formalize an approach for learning from sequences of physical corrections during the current task. To do this we introduce an auxiliary reward that captures the human's trade-off between making corrections which improve the robot's immediate reward and long-term performance. We evaluate the resulting algorithm in remote and in-person human-robot experiments, and compare to both *independent* and *final* baselines. Our results indicate that users are best able to convey their objective when the robot reasons over their sequence of corrections.**

I. Introduction

Imagine that you just got back from the store, and a two-armed personal robot is helping you unpack a bag of groceries (see Fig. 1). You don't want this robot to bump the bag into any cabinets or the hat on the left, but you also don't want the robot to stretch and rip the bag, or squeeze it and crush your groceries. Out of the corner of your eye, you notice that the robot is heading towards a side of the table that is wet (the blue region in the figure). So you *physically correct* it — pushing, pulling, or twisting the robot's arms to guide it away from that region and move it toward the green region on the left. Importantly, you can only physically interact with one arm at a time, since it's difficult to pay attention to how to correct both arms simultaneously: in the process of guiding this arm away from the obstacle, you might inadvertently move both arms closer together, squeezing the bag. Teaching this robot about all your preferences requires more than just one physical correction. You must provide a *sequence*: alternating between fixing the position of the arm closest to the obstacle, and adjusting the other arm so that the bag is held correctly.

State-of-the-art methods *learn* from physical corrections by treating each interaction as an *independent* event [1]–[4]. These works assume that the human makes corrections based only on their objective, without considering the other corrections they have already made or are planning to provide. But if we view human corrections as isolated events, we can misinterpret what they convey: for instance, when the human moves one arm away from the hat and closer to the other arm, this robot will mistakenly learn that the human wants to squeeze the bag.

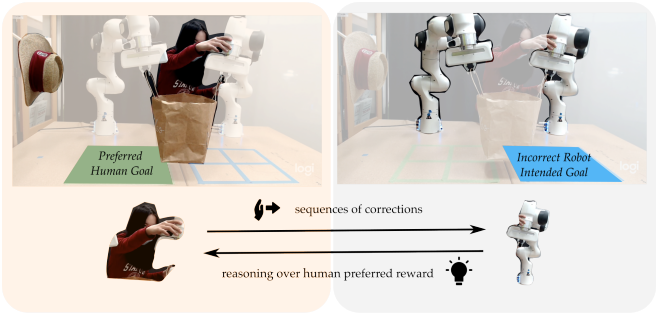

Fig. 1: Learning from physical sequences. Two robot arms are carrying a grocery bag toward an undesirable wet region, blue region on the right. The human provides a sequence of physical corrections to guide the robot toward their preferred objective, i.e., placing the bag on the green region while also avoiding the obstacles on the left, and holding the bag upright without squeezing or stretching it.

At the other end of the spectrum, robots can learn by looking at the *final* trajectory collectively produced by all of the human's corrections [5]–[9]. There are two issues with this: i) the robot does not learn or update its behavior until after the entire task is over, and ii) even the final trajectory may not capture what the human really wants. Returning to our example: because the user can only interact with one arm at a time, the final trajectory has some parts where the distance to the obstacle is right, and other sections where the bag is held correctly — but the final trajectory fails to capture both throughout.

At their core, prior works miss out on part of the process that humans use to correct the robot's behavior. Not everything can be fixed at once, or even fixed perfectly:

Humans corrections are not independent events — we often use multiple correlated interactions to correct the robot.

We leverage this insight to learn the human's reward function online from sequences of physical corrections, *without assuming* that human corrections are conditionally independent. Let's jump back to our example: a robot that reasons over the sequence recognizes that *collectively* the human corrections keep the robot away from the obstacle while holding the bag, even though the corrections *individually* fail to convey this objective, and can even be counter-productive.

Overall, we make the following contributions:

[1]Intelligent and Interactive Autonomous Systems Group (ILIAD), Dept of Computer Science, Stanford University, Stanford, CA 94305.
[2] Collaborative Robotics Lab (Collab), Virginia Tech.
(e-mail: mengxili@stanford.edu)

Capturing Conditional Dependence. We enable robots to learn from sequences of corrections by introducing an auxiliary reward function. This reward captures the human's trade-off between making corrections that increase the short-term reward (i.e., avoiding the obstacle) and reaching their long-term objective (i.e., carrying groceries).

Learning from Sequences. We introduce a tractable method to learn from a sequence of physical corrections by i) using the Laplace approximation to estimate the partition function and ii) solving a mixed-integer optimization problem.

Conducting Online and In-Person User Studies. Participants interacted with robots in both single- and multi-agent environments. We recorded user's corrections, and compared our approach to both *independent* and *final* baselines. Our proposed method outperforms both baselines, demonstrating the effectiveness of reasoning over sequences.

II. Related Work

Physical Human-Robot Interaction. When humans and robots share a workspace, physical interaction is *inevitable*. Prior work studies how robots can safely respond to physical interactions [10]–[12]. This includes impedance control [13] and other reactive strategies [14]. Most relevant to our setting is *shared control* [15]–[18], where the human and robot arbitrate between leader and follower roles. Although shared control enables the user to temporarily correct the robot's motion, it does not alter the robot's long term behavior: the robot does not *learn* from physical corrections.

Learning from Corrections (Online). Recent research recognizes that physical human corrections are often *intentional*, and therefore informative [1]–[4]. These works learn about the human's underlying objective in real-time by comparing the current correction to the robot's previous behavior. Importantly, each correction is treated as an *independent* event. Outside of physical human-robot interaction, shared autonomy follows a similar learning scheme — the robot uses human inputs to update its understanding of the human's goal online, but does not reason about the connections between multiple interactions [19]–[22]. We build upon these prior works by learning from sequences of corrections.

Learning from Corrections (Offline). By contrast, other works learn from the *final* trajectory produced after all the corrections are complete. This research is closely related to learning from demonstrations [6]. For example, in [5] the human corrects keyframes along the robot's trajectory, so that the next time the robot encounters the same task it moves through the corrected keyframes. Most similar to our setting are [9] and [8], where the robot iteratively updates its understanding of the human's objective *after* the human corrects the robot's entire trajectory. Although these works take multiple corrections into account, they do so offline, and are not helpful during the current task.

III. Formalizing Sequences of Physical Corrections

In this section we formalize a physical human-robot interaction setting where one or more robots are performing a task incorrectly. The human expert knows how these robots should behave, and physically corrects the robots to convey the true objective. But the human doesn't interact just once: the human may need to interact *multiple times* in order to correct the robots. Our goal is for these robots to learn the human's objective from this sequence of physical corrections.

A. Task Formulation

We formulate our problem as a discrete-time Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \rho_0)$. Here $\mathcal{S} \subseteq \mathbb{R}^n$ is the robot state space, $\mathcal{A} \subseteq \mathbb{R}^m$ is the robot action space, $\mathcal{T}(s, a)$ is the transition probability function, r is the reward, γ is the discount factor, and ρ_0 is the initial distribution.

Reward. Let the robot start from a state s^0 at time $t = 0$. As the robot completes the task it follows a trajectory of states: $\xi = \{s^0, s^1, \ldots, s^T\} \in \Xi$. The human has in mind a trajectory that they prefer for the robot to follow. Recall our motivating example — here the human wants the robot arms to follow a trajectory that avoids the cabinets without squashing the bag. Similar to prior work [7], [23]–[25], we capture this objective through our reward function: $R(\xi; \theta) = \theta \cdot \Phi(\xi)$. Here Φ denotes the feature counts over trajectory ξ, and θ captures how important each feature is to the human. We let ξ_R^t denote the robot's trajectory at timestep t, and we let θ^t denote the robot's current reward weights.

Suboptimal Initial Trajectory. The system of one or more robots starts off with an initial reward function $R(\xi; \theta^0)$, and optimizes this reward function to produce its initial trajectory.

$$\xi_R^0 = \arg\min_{\xi \in \Xi} \theta^0 \cdot \Phi(\xi)$$

But this initial trajectory ξ_R^0 misses out on what the human really wants — going back to our example, the robot does not realize that the blue region is wet and it needs to place the bag on the green region. More formally, the robot's estimated reward function (which is parameterized by θ^0) does not match the human's preferred reward function (parameterized by the true weights θ^*).

Human Corrections. The robot learns about the human's reward — i.e., the true reward weights — from physical corrections. Intuitively, these corrections are applied forces and torques which push, twist, and guide the robots. To formulate these interactions we must revise our problem definition: let a_R be the robot's action and let a_H be the human's *correction*. In practice, both a_R and a_H could be applied joint torques [1], [2]. Now the overall system transitions based on both human and robot actions: $s^{t+1} = \mathcal{T}(s^t, a_R + a_H)$. We use $A_H = \{(t_i, a_H^i), i = 1, \ldots, K\}$ to denote a *sequence* of K ordered human corrections a_H^i at time step t_i, where i keeps track of order of the corrections. Our goal is to learn the human's true reward weights from the sequence of corrections A_H.

B. Physical Corrections as Observations

When robots are performing a task suboptimally, the human expert intervenes to correct those robots towards the

right behavior. Going back to our example from Fig. 1. The user sees that the robot is making a mistake (moving towards the wet blue region), and physically intervenes. In the process of fixing this first issue, the human is forced to create another problem: by moving the first robot arm away from the blue region, they also move both arms closer together, and start to squash the bag. We note two important characteristics of these corrections: i) each human correction is intentional, and conveys information about the human's objective, but ii) the corrections viewed together may provide more information than isolating each interaction.

Leveraging these corrections, our goal is to find a better estimate of the reward parameters $P(\theta \mid A_H, \xi_R^0)$. We start by applying Bayes' rule:

$$P(\theta \mid A_H, \xi_R^0) \propto P(\theta) P(A_H \mid \xi_R^0, \theta) \tag{1}$$

In line with prior work [1]–[4], we will model $P(A_H|\xi_R^0, \theta)$ by mapping each human correction to a *preferred trajectory*. Given the human's correction (t_i, a_H^i), we deform the robot's trajectory to reach ξ_H^i. One simple example of this is to let the robot execute $a_R^{t_i} + a_H^{t_i}$ at this time step t_i, and stick to its original action plan a_R^t for future time steps $t > t_i$. More generally, we propagate the human's applied correction along the robot's current trajectory [26]:

$$\begin{aligned} \xi_H^1 &= \xi_R^0 + \mu A^{-1} a_H^1 \\ \xi_H^i &= \xi_H^{i-1} + \mu A^{-1} a_H^i, \quad i \in \{2, \dots, K\} \end{aligned} \tag{2}$$

Consistent with [26] and [27], μ and A are hyperparameters that determine the deformation shape and size. We emphasize that here the robot is not yet learning — instead, it is locally modifying its trajectory in the direction of the applied correction. Within our motivating example, let the human apply a force pushing the first robot arm away from the blue region. Equation (2) maps this correction to ξ_H, a trajectory that moves the robot arm farther from the blue region than ξ_R. In Fig. 2, we demonstrate how a sequence of corrections lead to a sequence of trajectories that enable the robot to correct its path and reach the preferred goals.

Now that we have this tool for mapping corrections to preferred trajectories, we can rewrite Equation (1):

$$\begin{aligned} P(\theta \mid A_H, \xi_R^0) &\propto P(\theta) P(A_H \mid \xi_R^0, \theta) \\ &= P(\theta) P\Big((t_1, a_H^1), \dots, (t_K, a_H^K) \mid \xi_R^0, \theta\Big) \\ &\approx P(\theta) P(\xi_H^1, \dots, \xi_H^K \mid \xi_R^0, \theta) \end{aligned} \tag{3}$$

Here $P(\theta)$ is the robot's prior over the human's objective, and $P(\xi_H^1, \dots, \xi_H^K \mid \xi_R^0, \theta)$ is the likelihood that the human provides a specific *sequence* of preferred trajectories given the robot's initial behavior ξ_R^0 and the reward weights θ.

IV. Learning from Sequences of Physical Corrections

In the previous section we outlined how robots can learn from physical corrections using Equation (3). However, we still do not know how to evaluate $P(\xi_H^1, \dots, \xi_H^K \mid \xi_R^0, \theta)$, which captures the relationship between a sequence of human corrections and the human's underlying objective. Prior work [1]–[4] has avoided this problem by assuming that the human's corrections are *conditionally independent*:

$$P(\xi_H^1, \dots, \xi_H^K \mid \xi_R^0, \theta) = \prod_{t=1}^{K} P(\xi_H^t \mid \xi_R^t, \theta) \tag{4}$$

Intuitively, this assumption means that there is no relationship between the human's previous corrections and their current correction. But we know this is not always the case — think about our motivating example, where the human's corrections are intricately coupled! Accordingly, here we propose a method for evaluating $P(\xi_H^1, \dots, \xi_H^K \mid \xi_R^0, \theta)$ *without* assuming conditional independence.

A. Reasoning over Sequences of Physical Corrections

To learn from sequences of human corrections online, we introduce an auxiliary reward function. In this section we also describe the modeling assumptions made by this auxiliary reward function, as well as an algorithm for leveraging this function for real-time inference.

Accumulated Evidence. We start by introducing the auxiliary reward function: $D(\xi_H^1, \dots, \xi_H^K, \theta)$. Let's refer to D as the *accumulated evidence* of a sequence of preferred trajectories $\xi_H^1, \dots, \xi_H^K$ under reward parameter θ. We hypothesize that the accumulated evidence should **(a)** reward not only the behavior of the final trajectory after all corrections, but also the intermediate trajectories the robot follows during the sequence of corrections. This recognizes that — when users make corrections — they don't sacrifice long-term reward for short-term failure. Consider our motivating example: the human is not willing to correct the robot into crushing the bag, even if that will reduce the overall number of corrections needed to avoid the obstacle. Of course, **(b)** humans also try to minimize their overall effort when making corrections — and any rewards for which the human's corrections are redundant or unnecessary are therefore not likely to be the human's true reward. Combining these two terms:

$$\begin{aligned} D\big(\xi_H^1, \dots, \xi_H^k, \theta\big) = \\ \sum_{t=1}^{K} \alpha^{K-t} R(\xi_H^t, \theta) - \gamma\Big(\sum_{t=1}^{K} \|a_H^t\|^2\Big) \end{aligned} \tag{5}$$

α and γ are two hyperparameters decaying the importance of previous corrections, and determining the relative trade-off between intermediate reward and human effort respectively.

Learning Rule. Similar to prior work in inverse reinforcement learning [7], [23], [25], [28], we model humans as noisily rational agents whose corrections maximize accumulated evidence for their preferred θ:

$$P(\xi_H^1, \dots, \xi_H^K \mid \xi_R^0, \theta) \propto \exp\Big(D\big(\xi_H^1, \dots, \xi_H^k, \theta\big)\Big) \tag{6}$$

Equation (6) expresses our likelihood function for learning from human corrections. Using Laplace's method (as applied

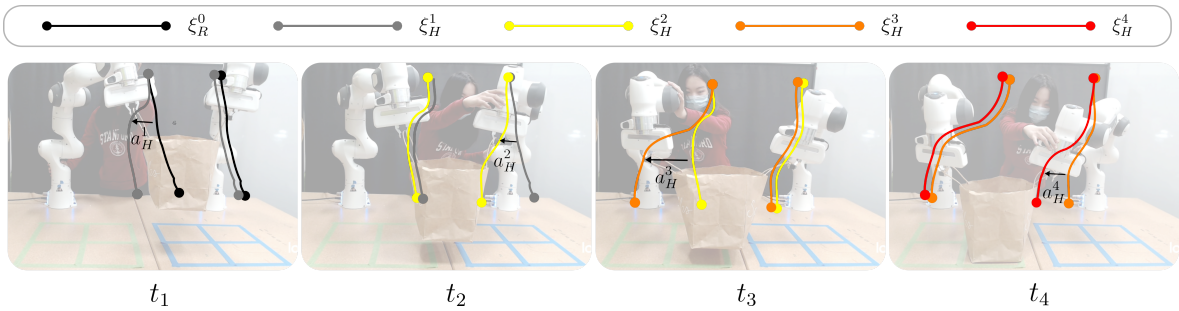

Fig. 2: An example of a sequence of human corrections along with her corresponding correction trajectories $\xi_H^1, \xi_H^2, \xi_H^3, \xi_H^4$ to guide the robot to place the grocery bag on the green region while avoiding any stretching or squeezing of the bag.

in [20]), we can approximate the normalization factor:

$$\begin{aligned} & P(\xi_H^1, \dots, \xi_H^K \mid \theta) \\ &= \frac{\exp\Big(D\big(\xi_H^1, \dots, \xi_H^k, \theta\big)\Big)}{\int_{\xi_H^1, \dots, \xi_H^K} \exp\Big(D\big(\xi_H^1, \dots, \xi_H^k, \theta\big)\Big) d\xi_H^1 \dots d\xi_H^K} \\ &\approx \frac{\exp\Big(D\big(\xi_H^1, \dots, \xi_H^k, \theta\big)\Big)}{\exp\Big(\max_{\xi_H^1, \dots, \xi_H^K} D\big(\xi_H^1, \dots, \xi_H^k, \theta\big)\Big)}. \end{aligned} \tag{7}$$

Monte-Carlo Mixed-Integer Optimization. Inspecting the denominator of Equation (7), we see that — in order to evaluate the likelihood of a sequence of corrections — we need to find the *highest possible* accumulated evidence the human *could* have achieved given that their objective is θ. Put another way, we need to search for the sequence of K corrections that maximize Equation (5) under θ:

$$\begin{aligned} D_K^*(\theta) &= \max_{\xi_H^1, \dots, \xi_H^K} D\big(\xi_H^1, \dots, \xi_H^k, \theta\big) \\ &= \max_{(t_1, a_H^1), \dots, (t_K, a_H^K)} D\big(\xi_H^1, \dots, \xi_H^k, \theta\big) \end{aligned} \tag{8}$$

Unfortunately, solving Equation (8) is hard because we need to figure out both *when* to make each correction $(t_1, \dots, t_K)$, which is a discrete decision, as well as *what* to correct during each interaction $(a_H^1, \dots, a_H^K)$, which is a continuous action.

To tackle this mixed-integer optimization problem, we develop a Monte-Carlo mixed-integer optimization method, where we first randomly sample discrete $(t_1, \dots, t_K)$, and then solve $(a_H^1, \dots, a_H^K)$ with gradient-based continuous optimization methods. The full pipeline for the optimization is summarized in Algorithm 1.

Importantly, the inputs to this optimization are only the robot's initial trajectory ξ_R^0, the reward parameter θ, and our model hyperparameters. Hence, we can conduct *offline* optimization to solve Equation (8) for several sampled values of θ. We then use the resulting library of stored D^* values to learn online, during physical interaction.

Online Inference. We have a way for solving for the denominator in Equation (7) offline. What remains to be evaluated is the numerator $\exp\Big(D\big(\xi_H^1, \dots, \xi_H^k, \theta\big)\Big)$. This can be easily computed online when the human is physically correcting the robot. Thus, we can now evaluate $P(\xi_H^1, \dots, \xi_H^K \mid \xi_R^0, \theta)$ while accounting for the relationships between corrections.

Algorithm 1: Monte-Carlo Mixed-Integer Program

Output: Maximum accumulated evidence $D_K^*(\theta)$, and K optimized human corrections $(t_1, a_H^1), \dots, (t_K, a_H^K)$

Input: The suboptimal initial robot plan ξ_R^0, reward parameter θ, and hyperparmeters in Eq. (5). $G(t_1, \dots, t_K)$ is a continuous optimizer.

```
Initialize maximum accumulated evidence and
  correction times: D*_K = −∞, T = ∅.
for i ← 0 to T_max do
    while (t_1, ..., t_K) ∈ T do
        Randomly sample (t_1, ..., t_K)
    end
    T = T ∪ {(t_1, ..., t_K)}
    D_i , (a_1, ..., a_K) = G(t_1, ..., t_K, ξ^0_R, θ)
    if D_i > D*_K then
        D*_K = D_i
        T*_H = (t_1, ..., t_K)
        A*_H = (a_1, ..., a_K)
    end
end
return D*_K, T*_H, A*_H
```

Our overall pipeline is as follows. Offline, we compute $D_K^*(\theta)$ with Alg. 1. Online, the robot starts by following the optimal trajectory ξ_R^0 with respect to its prior over θ. However, this initial reward might not capture the human reward and thus the human physically corrects the robot. The robot would then perform inference and update its belief over the reward when it receives new human corrections. At time t, the human has provided a total of K corrections $(t_1, a_H^1), \dots, (t_K, a_H^K)$, where $t_1 < t_2 < \dots < t_K \leq t$. We propagate these human corrections to get the deformed trajectories $\xi_H^1, \dots, \xi_H^K$ with Equation (2). We then perform Bayesian inference by solving Equation (7) and plugging the likelihood function into Equation (3). Finally, the robot uses its new understanding of θ to solve for an updated optimal trajectory ξ_R^t.

B. Relation to Prior Works: Independent & Final Baselines

To evaluate the effectiveness of our proposed method that reasons over correction sequences, we compare against two baselines: *Independent* and *Final*.

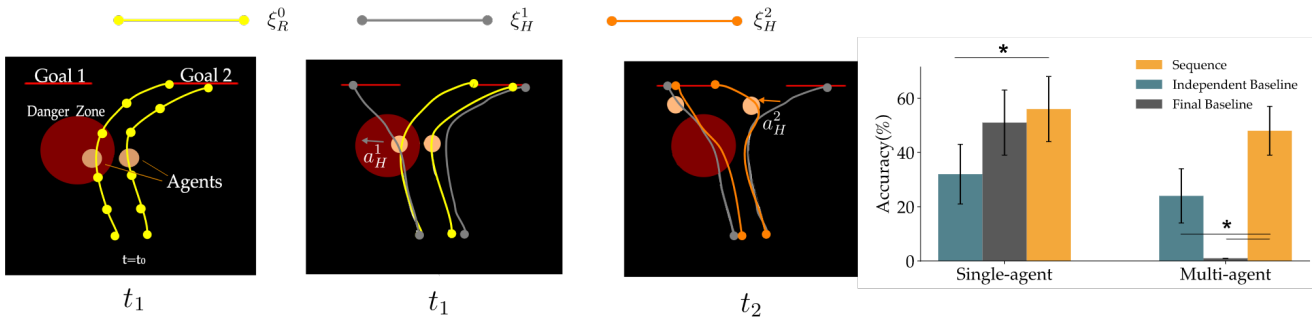

Fig. 3: Navigation Simulation. We show the sequence of corrections in the multi-agent task, and the accuracy results of the single- and multi-agent tasks.

Independent Baseline (Online). This baseline follows the same formalism as our approach, but assumes each human correction is conditionally independent. Hence, here we use Equation (4) to learn from each correction separately. The likelihood of observing an individual correction is related to the reward and effort associated with that correction [1]:

$$P(\xi_H^i \mid \xi_R^i, \theta) \propto \exp\left(R(\xi_H^i, \theta) - \gamma \|\xi_H^i - \xi_R^i\|^2\right) \quad (9)$$

Final Baseline (Offline). At the other end of the spectrum, we can always look at the final trajectory when all corrections are finished [9]. In this case, the robot learns by comparing the final trajectory, ξ_H^K, to the initial trajectory, ξ_R^0:

$$P(\xi_H^K \mid \xi_R^0, \theta) \propto \exp\left(R(\xi_H^K, \theta) - \gamma \|\xi_H^K - \xi_R^0\|^2\right) \quad (10)$$

To summarize, both our method and these baselines use a Bayesian inference approach. The difference is the likelihood function: *Independent* assumes conditional independence, while *Final* only considers the initial and final trajectory.

V. Experiments

To test our proposed algorithm, we conduct experiments with human users in a simulated navigation task and a robot manipulation task. We will discuss details that are consistent in both tasks and then elaborate on each one respectively.

Tasks. Our two tasks are: a web-based online simulated navigation task and an in-person robot manipulation task.

1) Navigation Simulation. In this task, a team of robots are navigating together through a specified region as in Fig. 3. The robot's objective function considered four features related to: reaching the goal, keeping formation, avoiding the danger zone, and minimum travel distance. We test our algorithm and baselines in different scenarios by varying robot team size, robot initial policy, and specifying different human preference reward values.

2) Robot Manipulation. In this task, two robot arms are carrying a full grocery bag to place it on the table. There are four features concerned in the reward: reaching the goal basket (blue or green regions as shown in Fig. 1), keeping the groceries inside of the bag while avoiding squeezing or stretching the bag, avoid touching nearby housewares such as cabinets or the hat shown in Fig. 1, and minimum trajectory length for efficiency. The robot starts off with the assumption of reaching an incorrect goal while also not realizing the bag is full, and should not be stretched or squeezed. Users apply forces to guide the two robot arms toward the correct goal region, while trying to keep the groceries inside of the bag.

Independent Variables. We compared three different inference models: reasoning over corrections independently (*Independent*), performing inference only based on the final correction (*Final*), and our model that reasons over the sequence of corrections (*Sequence*). *Independent* and our *Sequence* model can perform online inference, while *Final* will conduct offline inference after all corrections are provided.

Dependent Measures. We conduct experiments with human users and evaluate the effectiveness of the models by measuring the inference accuracy. Since we are unable to measure users' internal reward, we specify users' preferred reward function out of a predefined finite set of candidate reward parameters. We convey the preferred reward by explaining the priorities of the features to the user and demonstrating a desired robot trajectory using the preferred reward. Users are instructed to correct the robot to behave as optimally as possible, while minimizing their physical correction effort.

Hypotheses:

H1. *Compared to the online* Independent *baseline, reasoning over* Sequences *of corrections leads to higher accuracy and faster convergence to the preferred reward.*

H2. *Compared to the offline* Final *baseline, in addition to the advantages of online inference, our* Sequence *model achieves higher accuracy in challenging tasks – specifically tasks where fully correcting the robots is infeasible.*

A. Navigation Simulation

Experimental Details. We recruited 15 participants for two simulated navigation tasks. Participants interact with point-mass robots using a web browser, where they can observe the robots' trajectories, select a robot to correct, and provide corrections using the arrow keys. We collected data from humans for 5 episodes in two different scenarios: a simple single-agent scenario with only one robot, and a more complex scenario containing a two-robot team. In both settings, participants are only able to correct one robot at a time.

In both scenarios, robots' initial trajectory goes to an incorrect goal region as shown in Fig. 3. In the human's preferred reward function, not only is going to the correct goal encouraged, but other features are also encoded, includ-

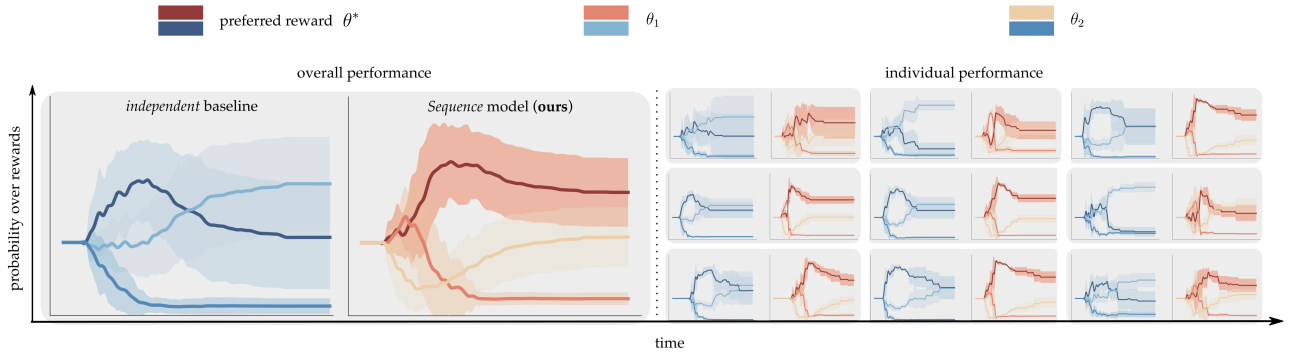

Fig. 4: Likelihood of three different reward parameters $(\theta^*, \theta_1, \theta_2)$ as more corrections are received over time. The preferred reward θ^* is shown with the darker red or blue. Each pair of plots demonstrate the *Sequence* model compared with the *Independent* model. We demonstrate the aggregate results over all users on the left, and the individual performance on the right. As shown *Sequence* outperforms *Independent* in identifying the preferred reward θ^*.

ing avoiding a danger zone and keeping the formation (equal distance between the robots throughout the trajectories).

Results & Analysis. We compute the average inference accuracy for the two baseline methods and our method across all users in both scenarios. The results are shown in Fig. 3.

Across both scenarios, our *Sequence* model demonstrates leading performance compared to both the *Independent* and the *Final* baselines. One thing to note is that in the single-agent scenario, the *Final* method has a comparable accuracy to our method. This indicates that in this simple task, the user can correct the robot to behave almost optimally so that simply looking at the final trajectory conveys sufficient information about the preferred reward. While in a more complex multi-agent scenario, the performance of the *Final* method drops significantly . This is because in this complex scenario, even if the user acts optimally, the final trajectory will still not have sufficient information to identify the preferred reward. In this example reasoning over sequences is dominant over only considering the final correction.

B. Robot Manipulation

Experimental Details. We recruited 9 participants for the in-person robot manipulation task (Fig. 1). Here, the two Franka Emika Panda robot arms are carrying a grocery bag to the table. The participants are asked to physically push or pull the robot to correct its behavior. Similar to the navigation task, the user can only correct one robot at a time, and we collected 5 episodes of corrections from each participant.

As shown in Fig. 1, there are two containers on the table for the grocery bag (the blue region and the green region). The robots' initial plan is to carry the bag to the right toward the blue region. However, the human wants to put the grocery to the left container. Meanwhile, since the grocery bag is almost full, in order to keep the groceries from falling out, the participants are also instructed not to squeeze or stretch the bag. In this setting there are three possible reward parameter θ, and the robot tries to infer the correct reward parameter θ^* from the human corrections.

Results & Analysis. We calculate the inference accuracy for all the three models (*Independent*, *Final*, and *Sequence*), and summarize the results in Table I. Our method demonstrates superior performance compared to the baselines.

TABLE I: Inference accuracy over 9 participants for robot manipulation.

	Sequence (**ours**)	*Independent*	*Final*
accuracy (%)	**82.22 ± 21.99**	31.11 ± 26.99	53.33 ± 13.30

In addition, we illustrate the probability distribution over θ with time in Fig. 4. Since the *Final* baseline performs the inference offline, we only visualize our *Sequence* model and the *Independent* baseline. As can be seen, across all of the participants, the probability for the preferred reward consistently dominates the other candidate rewards. However, for the *Independent* baseline, even if the probability for the preferred reward sometimes starts off high, it ends up not being the most likely reward as we receive more corrections. This is because in this two-robot task, redirecting the system to the correct goal while simultaneously maintaining the shape of the bag (formation) is not possible. The best possible corrections are: push one arm towards the goal, while stretching or squeezing the bag by a small amount, and then push the other arm in the same direction so that the bag shape is recovered. However, if we reason over these corrections independently, the robot will think that the first push indicates that preserving the bag shape is not important, and this leads to an incorrect inference.

C. Summary

Our results empirically support both of our hypotheses **H1** and **H2**. Our *Sequence* model not only conducts an online inference, but also demonstrates superior performance specially in complex multi-agent tasks.

VI. Conclusion

We developed a framework for learning from sequences of user corrections during physical human-robot interaction. We introduced an auxiliary reward that models the connections between corrections, and leveraged mixed-integer programming to solve for the best possible sequence of corrections. Our results from online and in-person users demonstrate that our approach outperforms methods that take each correction independently, or wait for the final trajectory.

ACKNOWLEDGEMENT

We acknowledge funding from the NSF Award #1941722 and #2006388, Future of Life Institute, and DARPA Hicon-Learn project.

REFERENCES

[1] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning robot objectives from physical human interaction," in *Conference on Robot Learning (CoRL)*, 2017, pp. 217–226.

[2] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning from physical human corrections, one feature at a time," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 141–149.

[3] D. P. Losey and M. K. O'Malley, "Including uncertainty when learning from human corrections," in *Conference on Robot Learning*, 2018.

[4] A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan, "Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 835–854, 2020.

[5] B. Akgun, M. Cakmak, K. Jiang, and A. L. Thomaz, "Keyframe-based learning from demonstration," *International Journal of Social Robotics*, vol. 4, no. 4, pp. 343–355, 2012.

[6] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[7] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Foundations and Trends in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[8] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 729–736.

[9] A. Jain, S. Sharma, T. Joachims, and A. Saxena, "Learning preferences for manipulation tasks from online coactive feedback," *The International Journal of Robotics Research*, vol. 34, no. 10, pp. 1296–1313, 2015.

[10] S. Haddadin and E. Croft, "Physical human–robot interaction," in *Springer handbook of Robotics*. Springer, 2016, pp. 1835–1874.

[11] A. De Santis, B. Siciliano, A. De Luca, and A. Bicchi, "An atlas of physical human–robot interaction," *Mechanism and Machine Theory*, vol. 43, no. 3, pp. 253–270, 2008.

[12] S. Ikemoto, H. B. Amor, T. Minato, B. Jung, and H. Ishiguro, "Physical human-robot interaction: Mutual learning and adaptation," *IEEE robotics & automation magazine*, vol. 19, no. 4, pp. 24–35, 2012.

[13] N. Hogan, "Impedance control: An approach to manipulation: Part ii—implementation," 1985.

[14] S. Haddadin, A. Albu-Schaffer, A. De Luca, and G. Hirzinger, "Collision detection and reaction: A contribution to safe physical human-robot interaction," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 3356–3363.

[15] Y. Li, K. P. Tee, W. L. Chan, R. Yan, Y. Chua, and D. K. Limbu, "Continuous role adaptation for human–robot shared control," *IEEE Transactions on Robotics*, vol. 31, no. 3, pp. 672–681, 2015.

[16] D. P. Losey, C. G. McDonald, E. Battaglia, and M. K. O'Malley, "A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction," *Applied Mechanics Reviews*, vol. 70, no. 1, 2018.

[17] D. A. Abbink, T. Carlson, M. Mulder, J. C. de Winter, F. Aminravan, T. L. Gibo, and E. R. Boer, "A topology of shared control systems – finding common ground in diversity," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 509–525, 2018.

[18] A. Mörtl, M. Lawitzky, A. Kucukyilmaz, M. Sezgin, C. Basdogan, and S. Hirche, "The role of roles: Physical cooperation between humans and robots," *The International Journal of Robotics Research*, vol. 31, no. 13, pp. 1656–1674, 2012.

[19] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 717–742, 2018.

[20] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.

[21] S. Jain and B. Argall, "Probabilistic human intent recognition for shared autonomy in assistive robotics," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–23, 2019.

[22] H. J. Jeon, D. Losey, and D. Sadigh, "Shared autonomy with learned latent actions," in *Proceedings of Robotics: Science and Systems (RSS)*, July 2020.

[23] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning." in *AAAI*, vol. 8, 2008, pp. 1433–1438.

[24] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *International Conference on Machine Learning*, 2004.

[25] H. J. Jeon, S. Milli, and A. D. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," *arXiv preprint arXiv:2002.04833*, 2020.

[26] D. P. Losey and M. K. O'Malley, "Trajectory deformations from physical human–robot interaction," *IEEE Transactions on Robotics*, vol. 34, no. 1, pp. 126–138, 2017.

[27] A. D. Dragan, K. Muelling, J. A. Bagnell, and S. S. Srinivasa, "Movement primitives via optimization," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2339–2346.

[28] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning." in *IJCAI*, vol. 7, 2007, pp. 2586–2591.