An Algorithm and User Study for Teaching Bilateral Manipulation via Iterated Best Response Demonstrations

Carolyn Chen,[1] Sanjay Krishnan,[1] Michael Laskey,[1] Roy Fox,[1] Ken Goldberg[1,2]

***Abstract*— Human demonstrations can be valuable for teaching robots to perform manipulation and coordination tasks. However, it can be difficult for human supervisors to provide demonstrations for multilateral (multi-arm) tasks, which require divided attention. In this paper, we propose a new algorithm called Bilateral Iterated Best Response (BIBR), which builds on the game-theoretic concept of Iterated Best Response. This algorithm allows a supervisor to train each manipulator iteratively, thereby reducing supervisor burden and improving the quality of demonstrations. We present a web-based user study of 51 participants controlling two agents in a GridWorld environment with a keyboard interface. We confirm prior work that bilateral demonstrations are noisier and longer than demonstrations provided separately for either manipulator when the task is asymmetric. As unilateral demonstrations lack coordination, this paper proposes learning coordinated bilateral policies from unilateral demonstrations by rolling out an estimated robot policy for one arm while the human demonstrates for the other, iteratively updating the estimated policy. Compared to a bilateral demonstration baseline, BIBR improves the success rate of the learned policy from 29.17% to 55.55% in the asymmetric task in the first full round of demonstrations. Furthermore, these policies learn trajectories that have 8.63% fewer steps and smoother trajectories, i.e., have 44.15% fewer changes in direction.**

I. Introduction

Bilateral manipulation of two independent end-effectors is common in many automation tasks in manufacturing, surgery, and personal robotics [22]. One approach for designing bilateral control policies is Learning from Demonstration (LfD) [1], where the robot learns to perform some task by observing demonstration trajectories generated by a human supervisor. Existing work in bilateral LfD assumes that the human can jointly control both arms — either by simultaneously providing demonstrations for both arms [17, 21, 28], or by making an assumption that both arms are coupled by a known function such as a distance field [6, 14, 25]. While the former approach applies to a wider set of bilateral scenarios, results in psychology and biomechanics suggest that humans are significantly less reliable in bilateral motion compared to moving a single arm [24]. Studies have found that humans tend to exhibit spatial and temporal synchronization in left and right motor control, which degrades performance in asymmetric bilateral tasks, i.e., tasks that require different motions of the two arms [2, 4, 5, 23, 27]. Intuitively, human demonstrators have limited attention, and their performance suffers when this attention is further divided into multiple simultaneous control tasks. Since LfD algorithms often assume an optimal supervisor, reduced demonstration quality can degrade robot performance [3].

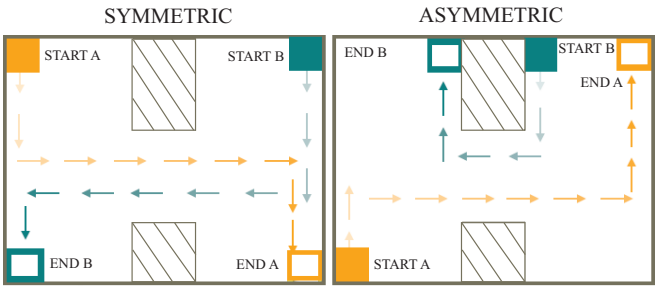


Fig. 1: This paper presents the Bilateral Iterated Best Response (BIBR) algorithm and a comparative user study of 51 participants performing symmetric and asymmetric GridWorld block moving tasks with keyboard input. We explore learning a policy from demonstrations moving the blocks from the start locations to the goal locations. Users provide input simultaneously for the two blocks using either both hands or train the policies iteratively by providing single-handed demonstrations with BIBR. Optimal trajectories are shown with light to dark shading.

We investigate the feasibility of allowing the human supervisor to train a single manipulator at a time for bilateral LfD tasks. Naively, however, these unilateral demonstrations could overlook any opportunity for coordination between the manipulators. To account for coordination, this paper proposes learning coordinated bilateral policies from unilateral demonstrations by rolling out an estimated robot policy for one manipulator while the human demonstrates for the other. The human alternates between manipulators in batches, iteratively teaching the robot an updated approximate policy. We call this algorithm Bilateral Iterated Best Response (BIBR), as it is inspired by the game-theoretic concept of a best response, with each manipulator modeled as an independent agent in a cooperative game.

We present an initial web-based user study of 51 participants in a GridWorld environment where participants controlled two agents with a keyboard interface (Figure 1). The agents (represented by solid blocks) start in fixed locations and need to cross between the two rooms to the outlined goals over a narrow doorway. Crossing this doorway requires coordination between the two agents. We consider both symmetric tasks, where the two agents have to perform the same actions up to reflection, and asymmetric tasks, where the agents take markedly different motions to get to their respective targets.

All authors are with the AUTOLAB at UC Berkeley (automation.berkeley.edu). [1] EECS, [2] IEOR, UC, Berkeley, CA USA; {carolyn.chen, sanjaykrishnan, laskeymd, royf, goldberg}@berkeley.edu

Summary of Contributions

1) We propose an algorithm, BIBR, for teaching bilateral manipulation using Iterated Best Response demonstrations.
2) We evaluate this algorithm in a 51 participant web-based user study, where results suggest that, under the constraints of the GridWorld environment, the learned policies from demonstrations provided via BIBR induce trajectories that are shorter and smoother. Successful policies are achieved at a rate of 56% as compared to 29% for the bilateral baseline within the first round of full demonstrations, and user survey responses suggest that BIBR makes bilateral demonstrations significantly easier to provide.

II. Related Work

Learning from Demonstrations (LfD) is a widely studied field [1] with applications that include self-driving cars [16], grasping in clutter [10], and surgical suturing [25]. In principle, the theory and algorithms of LfD can be applied to demonstrations for multiple manipulators in multilateral tasks by modeling the joint state and action spaces of all manipulators. However, it is common in practice to only consider unilateral tasks, or tasks with a single robotic manipulator. Bilateral manipulation tasks have been shown to be difficult for novice human demonstrators. For instance, the comparative study in [26] suggests that dividing endoscopic endonasal surgical bilateral tasks between two surgeons instead of a single surgeon positively affects training and performance of the procedure. The difficulty associated with bilateral tasks can be attributed to a phenomenon known as bilateral deficit, where the sum of unilateral performances is greater than the simultaneous performance of two limbs [27] as well as the fact that perceptual and visuomotor tasks compete for the same cognitive resources [23].

Furthermore, the challenges of bilateral manipulation are amplified as the task becomes asymmetric [2]. An example of asymmetry in robotic manipulation is cloth cutting, which requires one manipulator to perform precise cutting movements while the other tensions the cloth appropriately [12]. Results from a case study on human–robot collaboration for cloth cutting addressed this difficulty, showing that participants were able to complete the task faster when they were only required to operate the cutting arm while tensioning was automated, instead of jointly teleoperating both manipulators [20].

Within the existing literature for bilateral LfD, manipulators are simultaneously controlled and frequently considered as coupled. For instance, in [7], the authors kinesthetically demonstrate trajectories on both manipulators to infer a relative position function between the two end-effectors for automated coordination. More recently, the authors of [14] demonstrate online trajectory planning in dynamic environments for two surgical subtasks by maintaining an explicit coupling of the two manipulators when learning bilateral motions. In a similar sense, [21] encodes the demonstrations and determines the coordination patterns between the end-effectors. The authors of [15] present learning from demonstrations with constraints (C-LEARN), which includes constraints relating multiple end-effectors.

However, to the best of our knowledge, no work has yet considered performing demonstrations for bilateral coordination tasks iteratively rather than simultaneously. The authors of [11] provide insight as to how one might approach this problem of cooperative control using game theory. They demonstrate how cooperative control problems such as consensus or dynamic sensor coverage can be posed as a potential game. Using this model, an equilibrium mutual-best-response control profile can be found iteratively. In [13], the authors improve human-robot collaboration by allowing the robot to model human partial adaptation as a best-response game. Here, we consider how human demonstrations can be applied iteratively as best responses on bilateral systems to teach locally optimal policies in collaborative control problems.

III. Problem Statement

We propose a new training protocol that follows the game-theoretic concept of Iterated Best Response, allowing a single supervisor to train each manipulator iteratively. While the ideas generalize to more complex scenarios, this paper considers an initial user study on a GridWorld game where the state and action spaces are discrete and finite.

A. Imitation Learning

As a starting point for our problem statement, we consider imitation learning (IL), which studies the problem of *imitating* a supervisor's policy by observing demonstration trajectories sampled using this policy. For a state set $\mathcal{X}$, an action set $\mathcal{U}$, a fixed initial state $x_0 \in \mathcal{X}$, a state transition distribution $P(x_{t+1}|x_t,u_t)$, and some parametrization $\theta \in \Theta$ of allowed policies $\pi_\theta : \mathcal{X} \mapsto \mathcal{U}$, we consider the distribution $p(\xi|\theta)$ of the trajectory $\xi = (x_0,u_0,x_1,u_1,\ldots,x_T)$ induced by the policy π_θ

$$p(\xi|\theta) = \prod_{t=0}^{T-1} \mathbb{1}_{\{u_t=\pi_\theta(x_t)\}} P(x_{t+1}|x_t,u_t).$$

The objective is to minimize the expected disagreement with the supervisor on states visited by the robot while applying π_θ:

$$\arg\min_\theta \mathbf{E}_{\xi\sim p(\xi|\theta)}[\ell(\pi_\theta(x),\pi_{\theta^*}(x))],$$

for a loss function ℓ and unknown supervisor policy π_{θ^*}. Note that we do not assume θ^* is a member of the class Θ. Imitation learning seeks to mimic the supervisor's behavior as a proxy for some unknown global reward function R: $\mathcal{X}\times\mathcal{U}\mapsto\mathbb{R}$:

$$\mathbf{E}_{\xi\sim p(\xi|\theta^*)}\left[\sum_{t=0}^{T-1} R(x_t,u_t)\right]$$

By matching the supervisor's policy well enough, a high total expected reward can also be achieved by the robot.

Suppose that the N demonstration trajectories ξ_i of length T are sampled from the distribution $p(\xi_i|\theta^*)$ induced by the

supervisor policy. A common approach to determining θ is via Behavioral Cloning, where the empirical disagreement is minimized over the observed states and actions:

$$\arg\min_{\theta} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \ell(\pi_\theta(x_{i,t}), \pi_{\theta^*}(x_{i,t})).$$

This approach is justified when the dynamics are Markovian and fully observed. Behavior Cloning has been shown to better facilitate human supervision, compared to adaptive methods [9].

However, a weakness of IL is that it aims to match the performance of the demonstrator. If the demonstrator is inherently deficient at a task, the robot will learn this deficiency. This is particularly a limitation in asymmetric bilateral manipulation, where the quality of the demonstrations can be poor due to the bilateral deficit of bounded human cognition.

Therefore, we propose to apply IL over partitioned demonstrations, where the human only demonstrates for a single manipulator at a time. The challenge of this approach is in modeling and learning a coordinated bilateral policy, despite only observing unilateral demonstrations.

B. Multi-Agent Formulation

We consider the bilateral manipulation problem as a two-player game. We can define a finite-time Markov game to describe this problem, where each manipulator is an independent agent:

$$\langle \mathcal{X}, \{\mathcal{U}^1, \mathcal{U}^2\}, P, R, T \rangle. \tag{1}$$

Here $\mathcal{X}$ defines the state of the environment (which jointly describes both agents) and $\mathcal{U}^i$ defines the action sets of each individual agent. In each of the T rounds of the game, the two agents choose an action and play them simultaneously, and the world state transition distribution is $P(x_{t+1}|x_t, u_t^1, u_t^2)$. Both agents (manipulators) jointly receive the same reward $R(x_t, u_t^1, u_t^2)$ for their actions.

The partitioning of the action space into two agents means that we have to learn two policies:

$$\pi^1 : \mathcal{X} \mapsto \mathcal{U}^1 \quad \pi^2 : \mathcal{X} \mapsto \mathcal{U}^2,$$

which can similarly be denoted by their respective parameters θ^1 and θ^2. These policies are coupled through the dynamics P. If one of these policies were known, e.g. π^2, we would exactly recover single-agent IL by taking

$$P^1(x_{t+1}|x_t, u_t^1) = P(x_{t+1}|x_t, u_t^1, \pi^2(x_t))$$
$$R^1(x_t, u_t^1) = R(x_t, u_t^1, \pi^2(x_t)).$$

We could roll out policy π^2 simultaneously while the human demonstrates for arm 1, and apply the Behavioral Cloning algorithm to estimate π^1.

Instead of assuming that π^2 is known, we assume that it is learned alternatingly with π^1. Both policies are initialized arbitrarily. Then, we first demonstrate for manipulator 1 and update π^1 with an estimate. Following this step, we switch to demonstrate for manipulator 2 using the previous estimate of π^1 to roll-out a trajectory for arm 1. This process alternates until a fixed-point is reached or the human supervisor deems the task complete.

We assume that the human can provide *best responses*; that is, given any fixed policy for the automated manipulator, we assume that the supervisor can provide a demonstration trajectory by controlling the demonstration manipulator using the policy that maximizes the total expected reward. Under this assumption, the expected reward never decreases in any iteration, and it monotonically improves towards a limit value. By iteratively applying imitation learning, we approximate this equilibrium solution of the Markov game. We formalize this intuition in the next section.

IV. Bilateral Iterated Best Response

A. Iterated Best Response

A *best response* is defined as the optimal policy for a single agent, given the other agent's policy is fixed. For the Markov game in Equation (1), the best response for agent 1 given agent 2's policy is defined as:

$$\beta^1[\theta^2] = \arg\max_{\theta^1} \mathbf{E}_{\xi \sim p(\xi|\theta^1,\theta^2)} \left[\sum_{t=0}^{T-1} R(x_t, u_t^1, u_t^2) \right],$$

with $\xi = (x_0, u_0^1, u_0^2, x_1, u_1^1, u_1^2, \ldots, x_T)$ a game trajectory, and similarly for $\beta^2[\theta^1]$.

An *equilibrium solution* is any control profile (π^1, π^2) in which each policy is simultaneously a best response to the other policy. Iterated Best Response (IBR) is an optimization technique to approximate such a solution. In IBR, one iteratively chooses one policy to optimize given the other policy, which is fixed. IBR will converge to an equilibrium under the conditions described in [11, 19].

B. Bilateral Iterated Best Response

Inspired by IBR, we propose Bilateral Iterated Best Response (BIBR) to learn bilateral manipulation tasks iteratively by alternating demonstrations for different manipulators. The manipulators' policies are first initialized, the supervisor provides a set of demonstrations for the first manipulator, and the robot learns an approximate policy for this manipulator. The supervisor then demonstrates for the second manipulator while the robot rolls out the learned policy for the first manipulator. This process is iterated by alternating between rolling out the policy for one manipulator and demonstrating the best response for the other.

We assume that in every iteration, the supervisor for manipulator i demonstrates the best response for the learned policy for the other manipulator, denoted $-i$. In iteration k, BIBR collects a batch of N demonstrations and uses IL to imitate the best response policy for one manipulator given the other manipulator's fixed policy θ_{k-1}^{-i}:

$$\theta_k^i = \arg\min_{\theta^i} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \ell(\pi_{\theta^i}(x_{i,t}), \beta^i[\theta_{k-1}^{-i}]). \tag{2}$$

This procedure is outlined in Algorithm 1.

Algorithm 1 Bilateral Iterated Best Response

Hyperparameters: batch size of demonstration set N, number of rounds K
Initialize: θ_0^1 and θ_0^2, $i \leftarrow 1$
for $k = 1, \ldots, K$ **do**
  Use θ_{k-1}^{-i} to control other manipulator $-i$
  Collect N demonstrations for $\beta^i[\theta_{k-1}^{-i}]$
  Update θ_k^i by imitation (2)
  Switch manipulators $i \leftarrow -i$

V. Experiments

We now present an initial investigation of BIBR in symmetric and asymmetric bilateral manipulation tasks in a simulated GridWorld environment. As a baseline, we consider *joint learning*, where data are collected on both agents simultaneously and a policy is learned with imitation learning. We collect data in rounds and measure the value of the policy learned with the data in the current round. In this constructed environment, the number of agents to reach a goal state, time-to-goal, and smoothness of trajectory are examples of variables that are used to evaluate the policies at each round.

The experiments are organized as follows: (1) in a 7-subject pilot study, we present results that suggest that effects of symmetry vs. asymmetry observed in prior work [4] are also observed in the GridWorld game; and (2) in a 51-participant study, our results indicate that BIBR has a higher probability of converging to a successful policy with fewer demonstrations than joint learning. The study lasted 15 minutes per participant, and all were informed that participation was fully voluntary and that they could leave the experiment at any time[1]. Out of the 90 total participants to initialize an instance of the experiment, 51 completed the entire experiment. We use the 51 participants who completed the study as our subject pool.

A. Game Parameters

We evaluate BIBR using a custom two-handed web game (see Figure 1). The experimental environment consists of a simple GridWorld containing two moveable blocks (the agents), two goals, and impenetrable walls. The objective of the game is to navigate the agents around the walls *and* each other while moving towards their respective goals. The state $x_t \in \mathbb{R}^2$ is composed of the grid location of the two moveable agents at time t, and we restrict $\mathcal{U}$ to only contain four possible actions: left, up, right, and down. The dimensions of the GridWorld are 8×10, scaled by 30, so each agent can move to approximately 72000 distinct locations. Because there are two agents, the size of the state space $\mathcal{X}$ is therefore on the scale of 72000×72000.

The keyboard serves as the interface between the participant and the game. The box that begins on the left side of the GridWorld is controlled by the keys (A,W,D,S), and the box that begins on the right side of the world is controlled by the keys ($\leftarrow$, $\uparrow$, $\rightarrow$, $\downarrow$), respectively for the four actions listed above. Once the participant begins the game, the two agents move at a constant speed of 2 pixels per 20 milliseconds. This feature was implemented to ensure simultaneous movement of the two agents, i.e. that a participant could not trivialize the task by having the agents move one at a time. The game ends when both agents reach their respective goals. We count a *step* for each agent every 20 milliseconds before it reaches its goal and a *turn* for each agent when it changes its direction (or action). For the duration of the experiments, the GridWorld game was hosted on an Amazon EC2 instance.

[1]IRB Number 2017-01-9520

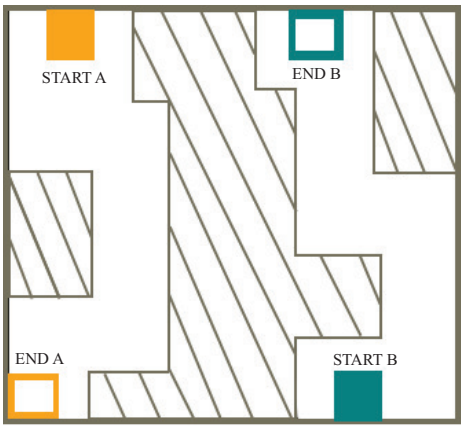


Fig. 2: First experimental environment. The user has to control the blue and the yellow agents through the disjoint mazes. Each user was asked to perform this task twice — once, with both hands, and once single-handed using BIBR. The task is successful when both agents arrive at their goal states.

B. Experiment 1: Demonstrations of Asymmetric Tasks

In our first experiment, we constructed a simplified GridWorld map to understand whether the results in prior work e.g., [4], applied to our task and scenario. Figure 2 illustrates the first scenario. A yellow agent and a blue agent start on opposite sides of the map and have to traverse disjoint mazes. This scenario is simplified as it does not involve coordination between the agents, in contrast with the following experiments.

7 subjects participated in this pilot human study on providing asymmetric demonstrations. Each participant was prompted to play two full games (excluding a short practice round). One game tested BIBR demonstrations and the other tested joint demonstrations. The ordering by which the games were played was randomly selected for each participant. For each *game*, there are 6 *rounds*, each of which prompts the participant to provide 5 *demonstrations*. A single *demonstration* consists of all the state-action pairs along the trajectory until the goal is reached. The five demonstration trajectories are used to train a Random Forest Classifier, composed of 10 trees and splitting nodes on the state ($x \in \mathbb{R}^2$), for a single agent per round. This classifier takes in the current state of the system and outputs the corresponding action for the agent. The classifier is then used to produce a *rollout* automatic trajectory. Note that, by the nature of the BIBR method, within the 6 rounds, each agent is only trained 3 times, whereas for the joint method, each agent is trained

	Method	Trials	Successes %	Success Rate Mean	Time Steps Mean	Turns Mean	Bumps Mean
Demonstrations	Joint	$7\times6\times5$	Not applicable	Not applicable	238.40 ± 5.42	52.67 ± 4.27	23.73 ± 2.26
	BIBR	$7\times6\times5$			$\mathbf{178.83\pm1.87}$	$\mathbf{6.86\pm0.95}$	$\mathbf{1.06\pm0.45}$
Learned	Joint	7×6	42.86 ± 14.97	0.44 ± 0.15	215.08 ± 7.34	31.67 ± 6.12	7.26 ± 0.81
	BIBR	7×6	$\mathbf{66.67\pm14.26}$	$\mathbf{0.80\pm0.09}$	$\mathbf{180.55\pm2.82}$	$\mathbf{7.38\pm1.03}$	$\mathbf{1.33\pm0.46}$

TABLE I: We constructed an asymmetric GridWorld, depicted in Figure 2, and performed a pilot study comparing BIBR and joint demonstrations for 7 participants. Each subject played two full games, which consist of 6 rounds $\times$ 5 demonstrations each. One game asked the participant to provide joint demonstrations while the other asked for BIBR demonstrations. We report both the evaluation metrics for the demonstrations as well as the rollout trajectories induced by the learned policies. There are a total of $7\times6\times5$ demonstrations provided and 7×6 learned trajectories. From the learned trajectories, we determine success as the percentage of automated trajectories where both agents arrive at the goal. We define success rate to be 0, 0.5, or 1 if no agents, one agent, or both agents navigate to their goals, respectively. We report the average time steps, turns, and bumps (when the agents hit the wall) for both the demonstrations and the learned trajectories.

6 times in the total span of the game. By construction, this task does not require coordination, so it does not require a joint policy to command both agents.

We compare collecting demonstrations on each single agent (BIBR) and collecting them jointly. Table I presents the evaluation metrics of both the demonstrations and the learned policies from these demonstrations. There were not enough participants for this pilot study to make significant claims, but aggregated across all their rounds of demonstrations, the experiment illustrates the difficulty of demonstrating trajectories for the two agents simultaneously in asymmetric tasks. As shown in Table I, the provided two-handed demonstration trajectories are longer, less smooth (i.e. more turns), and bump into the walls more, on average, than the provided single-handed demonstrations using BIBR. These comparisons hold true when evaluating the rollout trajectories with the same metrics. We include two additional metrics for the rollout trajectories. The percentage of success is simply the percentage of rounds aggregated across the participants that were able to automatically navigate successfully to the goal. We define success rate to be 0, 0.5, or 1 if no agents, one agent, or both agents succeed in arriving at their respective goals. From a sample size of 42 rounds per each method, we see that for BIBR, the expected success rate is, on average, greater than 0.5, i.e. more than one agent makes it to the goal on average, as compared to 0.44 for the joint method.

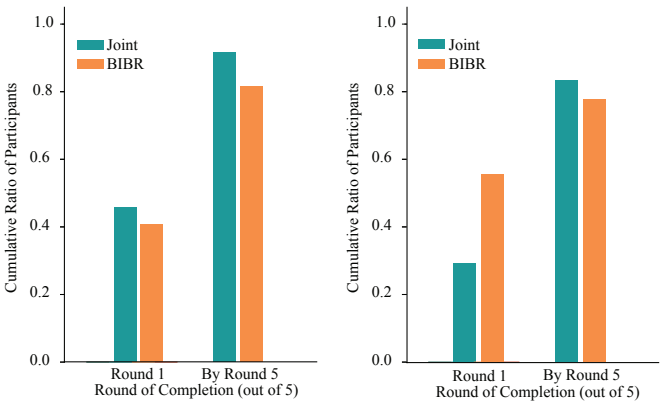
Analysis of the provided joint demonstrations shows that they have a greater number of disallowed actions (bumps against the walls) and turns, and take a greater number of time steps to complete. This can lead to inconsistent data for learning and potentially less successful policies. On the other hand, by demonstrating a single agent at a time, BIBR has a more consistent training dataset. This difference in inconsistency is addressed in a following section and illustrated in Figure 7.

C. Experiment 2: Factorial Study

In the next experiment, we consider a factorial study of the following conditions: symmetric task with joint, asymmetric task with joint, symmetric task with BIBR, and asymmetric task with BIBR. For each participant, we assign a random ordering of a symmetric and asymmetric task, illustrated in Figure 1. We randomly assign each participant (with equal probability) either to play the game with BIBR or joint demonstrations. In addition to the symmetric and asymmetric game, we include a practice round of the symmetric map at the very beginning to familiarize the participant with the system and interface. We omit data from the practice round in our analysis.

The participant demonstrates a trajectory by guiding one of the agents (if BIBR) or both of the agents (if playing two-handed) to the respective goals, as in Experiment 1, and five demonstrations are required of each round. After each round, there is a *rollout*, where the joint policies are rolled out and evaluated by whether or not the automated agents arrive at the goal, and how smooth and long the induced trajectory is. Differing from Experiment 1, if after the rollout both agents make it to their respective goals, then we assume that a local optimum has been reached and the *game* is complete, and the participant is prompted with the next game. If not, then the game proceeds to another round. The participant will then need to provide five more demonstrations for this round. If the participant is assigned to provide demonstrations in the manner of BIBR, then the participant switches to navigate the other agent for the new round while the agent in the prior round moves autonomously using its corresponding Random Forest Classifier. Otherwise, the participant is asked to demonstrate on both agents again for the new round. We set the maximum number of rounds to 5 to ensure that the experiment has a finite horizon, at which case the subject is prompted with the next game.

1) Successful Policies: Of the 51 subjects that completed the game, 27 were randomly assigned to use BIBR, and 24 were assigned to the joint demonstrations. After the first round of data, the learned policy with BIBR is more likely to complete the task successfully (Figure 3). Note that we consider the first round for BIBR to really be the second round, when both arms have been provided five demonstrations. While the joint method has a 29.17% probability of success by the end of the first round, the BIBR method has higher probability of success of 55.55%. Eventually, as more data is collected, this advantage narrows. By the end of the 5th round, 77.78% of the subjects assigned BIBR were able to complete the symmetric game, compared to 81.48% who were able to complete the asymmetric game. Probability of success for the joint demonstration subjects was higher, with 91.67% of the subjects completing the symmetric game by the end of the 5th round and 87.5% completing the asymmetric game. We



(a) Symmetric Task (b) Asymmetric Task

Fig. 3: Ratio of subjects who completed the game in the first round and cumulatively within the allotted five rounds out of the total number of subjects in each category (i.e. joint vs. BIBR) that completed the entire experiment. We define completion as success of the learned policy to navigate both agents to their respective goals. The total number of subjects to successfully finish both the symmetric and asymmetric tasks is 24 under the joint demonstrations and 27 under BIBR. In Figure 3a, a symmetric map was used to evaluate the joint method versus BIBR. 22 subjects are represented under both the joint and BIBR methods. In Figure 3b, an asymmetric map was used to evaluate joint versus BIBR. 21 subjects in each category represent the complete set of people to complete the asymmetric task successfully. For the asymmetric task, 55.55% of the participants who used BIBR were successful after the first round (where we take the first round to be when both arms have been demonstrated on five times), while 29.17% of the joint demonstrations subjects completed the task after one round. By the end of the fifth round, 87.5% of the joint demonstrations subjects completed the asymmetric game versus 77.5% of BIBR subjects. However, it can be noted that by the fifth round, one agent was trained 3 times and another was trained 2 times for BIBR as opposed to the joint method, where both agents were trained 5 times each.

do notice that joint has a lower ultimate failure rate (never completing the game by the fifth round) of 16.67% instead of 22.22% for BIBR.

2) Length and Smoothness of the Trajectories: Next, we look at the length and smoothness of the trajectories collected using the BIBR and joint training procedures. The objective is to understand why BIBR improves early training performance. Our hypothesis is that BIBR collects more consistent demonstrations.

We omit data in each condition from the subjects of the pool of 51 participants who never successfully trained the agents for the given task within the alloted five rounds, and we only consider the success pool (22 and 21 subjects complete the symmetric and asymmetric games, respectively, for both BIBR and joint demonstrations). Figure 4 and Figure 5 illustrate the average number of steps and turns of the final successful learned trajectories, respectively, for each of the categories within the success pool. As shown in Figure 4, there is no significant difference in the length of the learned trajectory within the symmetric games between demonstrations provided jointly versus via BIBR, but there is a small significant difference within the asymmetric games, with joint demonstrations averaging at 278 ± 12.22 steps compared to 254 ± 10.52 steps for the BIBR demonstrations. The same observations hold for the number of "turns" in the learned trajectory — there is no significant difference in performance for the symmetric game, but within the asymmetric games, BIBR collects significantly smoother trajectories at 17.10 ± 4.51 turns compared to 30.62 ± 5.10 under the joint demonstrations. Results thus suggest that, in the context of the GridWorld game and the constraint that the participants successfully complete the game, the average trajectory induced by the rollout policy is smoother and shorter under demonstrations collected using BIBR than the average trajectory induced by demonstrations collected jointly for asymmetric tasks. However, when the task is symmetric, there is no observable difference in the performance of the learned policy, which supports the experimental results of [5] and [8]. These results, along with the consideration that BIBR inherently requires twice the amount of time to provide the same number of total demonstrations between two agents, presents a tradeoff between the two evaluated methods.

3) User Survey of Difficulty of Demonstrations: To evaluate these methods from a human supervisor's perspective, we asked all participants who completed the entire experiment to rate each task in an exit survey. For the asymmetric and symmetric tasks, participants were allowed to choose a discrete rating from 1 to 5, with 5 being the most difficult. As shown in Figure 6, the subjects of the experiment found that the BIBR approach was significantly easier than the joint approach for the asymmetric task. The joint asymmetric games were rated 3.42 ± 0.52 while the BIBR asymmetric games were rated 2.11 ± 0.43. While there is not a significant difference between the joint and BIBR methods for the symmetric task, Figure 6 illustrates that, under the constraints of this task, BIBR is not perceived to be more difficult than the joint method. When asked in the exit survey for an explanation of their rating, some subjects remarked that *the two-handed task was much more exhausting*, which again underlines the ability of BIBR to address the core motivations of this paper.

4) Coverage of Demonstrations: Finally, we wish to understand why, despite being less smooth than BIBR demonstrations, joint demonstrations could train a successful policy, sometimes with greater success than BIBR (see Figure 3). Our hypothesis was that the joint demonstrations were naturally noisier, as they were more difficult to provide than the single-handed demonstrations. By the nature of the experiment, since the joint demonstration trajectories were longer and covered a greater portion of the state space of the system, the Random Forest Classifier was provided more data from the joint method over BIBR. In comparison, BIBR encourages precise demonstrations by eliminating the challenging cognitive load of a bilateral task, thereby effectively providing the Random Forest Classifier with a reduced training set. We illustrate the difference in coverage for the factorial experiment in Figure 7. As predicted, BIBR demonstrations cover a smaller portion of the state space (e.g. 52.27% versus 58.70% for joint method in the asymmetric task). We plan to address this artifact in future work.

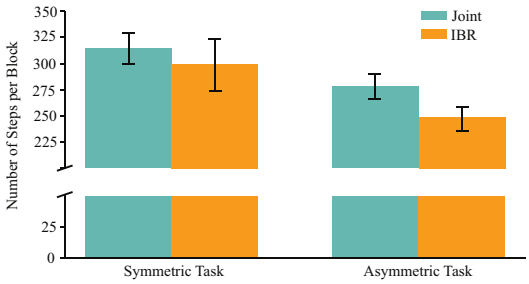


Fig. 4: Average trajectory length among subjects who completed the game within the five allotted rounds. For both BIBR and joint, 21 and 22 participants, respectively, completed the asymmetric and symmetric game. The average number of steps per agent for the final successful rollout trajectory is summarized here, along with 1.96× standard error of the mean. There is a small but statistically significant difference in the number of steps taken between the joint (278.08 ± 12.22) and BIBR (254.00 ± 10.52) methods for the asymmetric game but a minimal difference for the symmetric game.

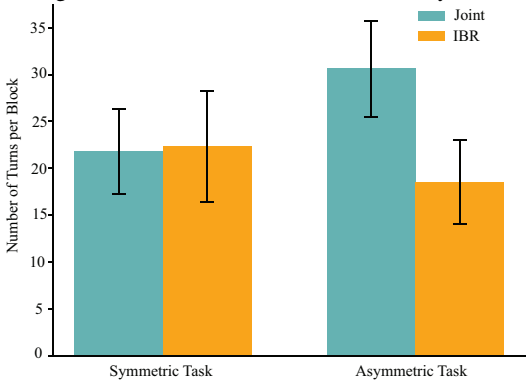


Fig. 5: Average trajectory smoothness among subjects who completed the game within the five allotted rounds. The average number of turns per agent for the final successful rollout trajectory is summarized here as a measure of smoothness of the induced trajectory. There is a statistically significant difference in the number of turns taken between the joint (30.62 ± 5.1) and BIBR (17.1 ± 4.5) methods for the asymmetric game, but there is again no significant difference for the symmetric game.

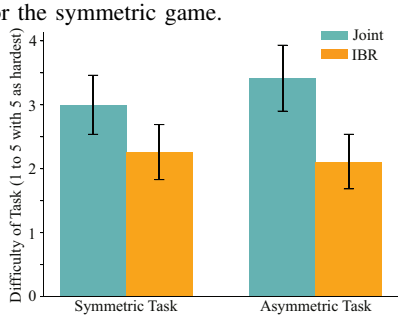


Fig. 6: We asked all 51 participants to rate the difficulty of the symmetric and asymmetric games from 1 to 5, with 5 being the most difficult. We summarize the average results from this survey here and present the 95% confidence interval about the mean. Mirroring Figure 4 and Figure 5, we see no significant difference between BIBR and joint demonstrations under the symmetric task. However, the subjects rated joint demonstrations as significantly more difficult than those who used BIBR for the asymmetric task. Subjects under the condition of joint demonstrations rated, on average, that the asymmetric task was a difficulty of 3.42 ± 0.52 on a scale of 1 to 5, which is significantly higher than the difficulty of 2.11 ± 0.43 reported for the BIBR demonstrations.

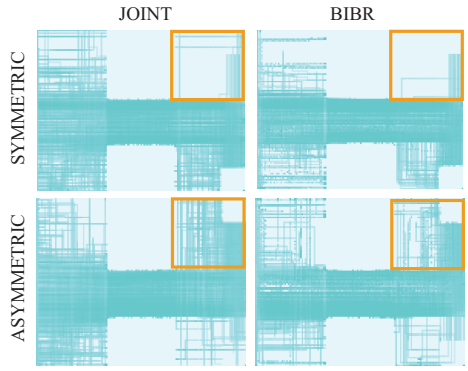


Fig. 7: We visualize the aggregated coverage of all demonstrations provided by users for each factorial of the experiment. This visualization shows a slightly sparser coverage of the state space from the BIBR demonstrations than the joint demonstrations. For the symmetric task, the joint demonstrations cover 48.72% of the allowable state space, as compared to 44.93% by BIBR. For the asymmetric task, the joint and BIBR demonstrations cover 58.70% and 52.27% of the allowable state space, respectively. We highlight a region in orange for visual comparison.

VI. Future Work and Discussion

This paper presents an initial exploration of learning coordinated bilateral policies from unilateral demonstrations by rolling out an estimated policy for one manipulator while the human demonstrates for the other. Bilateral Iterated Best Response (BIBR) in an initial user study of 51 participants suggests: (1) a confirmation of prior work that bilateral (both manipulators at once) demonstrations are noisier and longer than demonstrations for either single manipulator when the task is asymmetric, (2) in such settings, BIBR can learn viable policies with significantly fewer demonstrations than jointly providing demonstrations for both arms. These results are promising as they suggest a number of important directions for future work:

Cost-sensitivity: One limitation of our current study is that we do not consider rewards when learning policies through imitation. This leads to potentially confounding demonstrations that are of poor quality, skewing the learned results. We believe that this can be addressed with techniques similar to off-policy Reinforcement Learning by approximating a Q-function instead of trying to learn the policy. This would use the cost-to-go (i.e., the Q function) to weight predictions so that mispredicting strategic state-action pairs is penalized more.

Aggregating Rounds: Another consideration is aggregating data from multiple rounds of demonstrations. Prior work suggests that this problem is subtle and a number of challenging issues can arise [9, 18]. Importance sampling can be used to aggregate data collected under different policies.

Physical User Studies: We will also use BIBR in physical studies in robotic manipulation tasks. We are particularly interested in learning from demonstrations policies for deformable manipulation, such as cutting thin tissue, knot tying, and buttoning/unbuttoning. These user studies will also illustrate additional challenges in kinematics and configuration in

addition to multilateral coordination, e.g., when a human is controlling two manipulators with different kinematic properties. A current challenge with rolling out polices on physical robots is ensuring movement of realistic smoothness and speed. This is necessary in order to obtain an accurate *best response* from the supervisor at each iteration.

Three or more arms: We will also consider generalizations to more than two arms. For example, the Intuitive Surgical da Vinci has four arms in the clinical version, but a human teleoperator can only control two at one time. We hope to show that in such setups an algorithm like BIBR can be used to facilitate coordinated multilateral operation.

ACKNOWLEDGMENT

This research was performed at the AUTOLAB at UC Berkeley in affiliation with the RISELab, BAIR, and the CITRIS "People and Robots" (CPAR) Initiative: http://robotics.citris-uc.org in affiliation with UC Berkeley's Center for Automation and Learning for Medical Robotics (Cal-MR). The authors were supported in part by the U.S. National Science Foundation under NRI Award IIS-1227536: Multilateral Manipulation by Human-Robot Collaborative Systems, and by Google, UC Berkeley's Algorithms, Machines, and People Lab, and by a major equipment grant from Intuitive Surgical and by generous donations from Andy Chou and Susan and Deepak Lim. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Sponsors. We thank Anca Dragan, Matthew Matl, Jacky Liang, Molly Nicholas, Aimee Goncalves, Daniel Seita, Sequoia Beckman, and Richard Liaw for their extensive advice and feedback on this manuscript.

REFERENCES

[1] 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 (cit. on pp. 1, 2).

[2] J. Blinch, B. D. Cameron, E. K. Cressman, I. M. Franks, M. G. Carpenter, and R. Chua, "Comparing movement preparation of unimanual, bimanual symmetric, and bimanual asymmetric movements", *Experimental brain research*, vol. 232, no. 3, pp. 947–955, 2014 (cit. on pp. 1, 2).

[3] C. Chuck, M. Laskey, S. Krishnan, R. Joshi, and K. Goldberg, "Statistical data cleaning for deep learning of automation tasks from demonstrations", in *ICRA, 2017 IEEE*, IEEE, 2017 (cit. on p. 1).

[4] B. Fowler, T. Duck, M. Mosher, and B. Mathieson, "The coordination of bimanual aiming movements: Evidence for progressive desynchronization", *The Quarterly journal of experimental psychology*, vol. 43, no. 2, pp. 205–221, 1991 (cit. on pp. 1, 4).

[5] E. A. Franz, "Spatial coupling in the coordination of complex actions", *The Quarterly Journal of Experimental Psychology: Section A*, vol. 50, no. 3, pp. 684–704, 1997 (cit. on pp. 1, 6).

[6] A. Gams, B. Nemec, A. J. Ijspeert, and A. Ude, "Coupling movement primitives: Interaction with the environment and bimanual tasks", *IEEE Transactions on Robotics*, vol. 30, no. 4, pp. 816–830, 2014 (cit. on p. 1).

[7] E. Gribovskaya and A. Billard, "Combining dynamical systems control and programming by demonstration for teaching discrete bimanual coordination tasks to a humanoid robot", in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, IEEE, pp. 33–40 (cit. on p. 2).

[8] D. Kourtis, L. De Saedeleer, and G. Vingerhoets, "Handedness consistency influences bimanual coordination: A behavioural and electrophysiological investigation", *Neuropsychologia*, vol. 58, pp. 81–87, 2014 (cit. on p. 6).

[9] M. Laskey, C. Chuck, J. Lee, J. Mahler, S. Krishnan, K. Jamieson, A. Dragan, and K. Goldberg, "Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations", *ArXiv preprint arXiv:1610.00850*, 2016 (cit. on pp. 3, 7).

[10] M. Laskey, J. Lee, C. Chuck, D. Gealy, W. Hsieh, F. T. Pokorny, A. D. Dragan, and K. Goldberg, "Robot grasping in clutter: Using a hierarchy of supervisors for learning from demonstrations", *Automation Science and Engineering (CASE), 2016 IEEE*, pp. 827–834, 2016 (cit. on p. 2).

[11] J. R. Marden, G. Arslan, and J. S. Shamma, "Cooperative control and potential games", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 6, pp. 1393–1407, 2009 (cit. on pp. 2, 3).

[12] A. Murali, S. Sen, B. Kehoe, A. Garg, S. McFarland, S. Patil, W. D. Boyd, S. Lim, P. Abbeel, and K. Goldberg, "Learning by observation for surgical subtasks: Multilateral cutting of 3d viscoelastic and 2d orthotropic tissue phantoms", in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, IEEE, 2015, pp. 1202–1209 (cit. on p. 2).

[13] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa, "Game-theoretic modeling of human adaptation in human-robot collaboration", in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2017, pp. 323–331 (cit. on p. 2).

[14] T. Osa, N. Sugita, and M. Mitsuishi, "Online trajectory planning in dynamic environments for surgical task automation.", in *Robotics: Science and Systems*, 2014 (cit. on pp. 1, 2).

[15] C. Pérez-D'Arpino and J. A. Shah, "C-learn: Learning geometric constraints from demonstrations for multi-step manipulation in shared autonomy", in *IEEE ICRA*, 2017 (cit. on p. 2).

[16] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network", Carnegie-Mellon University, Tech. Rep., 1989 (cit. on p. 2).

[17] C. E. Reiley, E. Plaku, and G. D. Hager, "Motion generation of robotic surgical tasks: Learning from expert demonstrations", in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, IEEE, 2010, pp. 967–970 (cit. on p. 1).

[18] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning", *AISTATS. Vol. 1. No. 2*, 2011 (cit. on p. 7).

[19] T. Roughgarden, *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, 2016 (cit. on p. 3).

[20] K. Shamaei, Y. Che, A. Murali, S. Sen, S. Patil, K. Goldberg, and A. M. Okamura, "A paced shared-control teleoperated architecture for supervised automation of multilateral surgical tasks", in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, IEEE, 2015, pp. 1434–1439 (cit. on p. 2).

[21] J. Silvério, L. Rozo, S. Calinon, and D. G. Caldwell, "Learning bimanual end-effector poses from demonstrations using task-parameterized dynamical systems", in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, IEEE, 2015, pp. 464–470 (cit. on pp. 1, 2).

[22] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic, "Dual arm manipulation—a survey", *Robotics and Autonomous systems*, vol. 60, no. 10, pp. 1340–1353, 2012 (cit. on p. 1).

[23] I. Stanciu, S. C. Biehl, and C. Hesse, "Increased cognitive demands boost the spatial interference effect in bimanual pointing", *Psychological research*, pp. 1–14, 2016 (cit. on pp. 1, 2).

[24] S. P. Swinnen and N. Wenderoth, "Two hands, one brain: Cognitive neuroscience of bimanual skill", *Trends in cognitive sciences*, vol. 8, no. 1, pp. 18–25, 2004 (cit. on p. 1).

[25] J. Van Den Berg, S. Miller, D. Duckworth, H. Hu, A. Wan, X.-Y. Fu, K. Goldberg, and P. Abbeel, "Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations", in *ICRA, 2010 IEEE*, IEEE, 2010, pp. 2074–2081 (cit. on pp. 1, 2).

[26] F. Vaz-Guimaraes, M. M. Rastelli, J. C. Fernandez-Miranda, E. W. Wang, P. A. Gardner, and C. H. Snyderman, "Impact of dynamic endoscopy and bimanual-binarial dissection in endoscopic endonasal surgery training: A laboratory investigation", *Journal of Neurological Surgery Part B: Skull Base*, vol. 76, no. 05, pp. 365–371, 2015 (cit. on p. 2).

[27] S. Vieluf, G. Aschersleben, and S. Panzer, "Lifespan development of the bilateral deficit in a simple reaction time task", *Experimental Brain Research*, pp. 1–8, 2016 (cit. on pp. 1, 2).

[28] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable folding task by humanoid robot worker using deep learning", *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 397–403, 2017 (cit. on p. 1).