2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

# Leveraging Reinforcement Learning for Human Motor Skill Acquisition

Keya Ghonasgi[1], Reuth Mirsky[2], Bharath Masetty[1], Sanmit Narvekar[2],
Adrian Haith[3], Peter Stone[2,4] and Ashish D. Deshpande[1]

***Abstract*— Acquiring a motor ability is a complex process, whether for an athlete working toward peak performance or a post-stroke patient re-learning to control a limb. Curriculum selection is the process of choosing a sequence of sub-tasks, their training order, and their frequency in order to achieve a complex target task. Currently, in motor skill training, no systematic method exists for selecting curricula, and can result in long, costly and often unsuccessful training. At the same time, recent advances in artificial intelligence have introduced curriculum learning using Reinforcement Learning, which has enabled some impressive speed-ups in artificial agents' abilities to learn complex tasks. This paper delineates how the computational approaches used in curriculum learning for reinforcement learning can be modified, to represent the learning process of people in motor tasks. This paper also presents some preliminary results on a dynamic motor game designed to evaluate the process of motor task learning and the efficacy of different curricula.**

## I. INTRODUCTION

How best to train humans to perform new motor tasks is a challenging and open problem across diverse fields such as neuroscience, rehabilitation, medicine, and athletics [2], [11]. However, currently in human motor skill training, no systematic method exists for creating, organizing, and testing training programs (or curricula), and can result in long, costly, and often unsuccessful training. Curricula across these domains are developed largely based on tradition and intuition, centered around the ill-defined notion of 'practice' [13]. While there are rehabilitative robots that are designed to help people learn or regain skills, a physical therapist is still required to individually prescribe a curriculum for using the robot. At the same time, recent advances in AI enabled efficient curriculum learning using Reinforcement Learning (RL), which has enabled some impressive speed-ups in artificial agents' learning abilities [18]. Specifically, RL is a paradigm for learning sequential decision making tasks with delayed rewards that has been employed for training autonomous agents to perform sequential tasks [27]. Within RL, curriculum learning algorithms have been developed where the goal is to design a sequence of *source tasks* for an agent to train (practice) on, such that final performance or learning speed is improved compared to learning on the *target task* directly. This paper leverages the formulation of curriculum learning as an RL problem to enable automatic generation of such curricula for people.

The first contribution of this paper is a computational model for learning a motor task similar to a curriculum learning problem in RL. With the ability to formalize a motor learning task as an RL problem, many potential RL algorithms can become available to assist in learning a curriculum for the defined target task. The formulation involves defining the curriculum learning problem as a Markov Decision Process (MDP), where the actor is the teacher that needs to learn the best training policy for a student. The second contribution of this paper is a compilation of human skill acquisition and motor learning concepts, such as *challenge point* [12] and *flow channel* [34] into formal definitions within a MDP. The third contribution is a motor game specifically designed to evaluate learning performance using our new framework. We represent learning processes in this novel task as a MDP, and show preliminary results from a pilot study, where we investigate the potential applicability and the benefit of a curriculum for the newly introduced motor task.

## II. RELATED WORK

Various works have addressed the problem of modulating a sequence of tasks to improve learning. All of them share the common premise that there is value not only to the *content* of tasks given to the learner, but also in the *order* in which these motor tasks are introduced, so that the task is challenging enough to learn from, but not too challenging that the learner will not be able to learn from it. This section is ordered as follows: We start by presenting different approaches for choosing tasks and their difficulty, then we discuss works that identify the leading concepts behind the notion of a "good next task." We then detail the recent advancements in curriculum learning for RL agents. Finally, we describe how these different components can be combined together, and how they were combined in previous work.

### A. Task Choices in Motor and Cognitive Tasks

In rehabilitation, a curriculum usually amounts to sequentially selecting the difficulty (or challenge) of the task presented to the patient. This curriculum can be modified in various ways, including the amount of assistance provided by a robot during individual therapy, and the method of selecting it is often based on the idea of a "challenge point". [3], [15], [33]. Another approach to modulate challenge is to use a

[1]Department of Mechanical Engineering, The University of Texas at Austin. {keya.ghonasgi, bmasetty}@utexas.edu, ashish@austin.utexas.edu
[2]Department of Computer Science, The University of Texas at Austin. {reuth, sanmit, pstone}@cs.utexas.edu
[3]Department of Neuroscience, Johns Hopkins University School of Medicine. adrian.haith@jhu.edu
[4] SONY AI

social assistive robot that engages individuals to choose challenging tasks [9]. In all of these approaches, the challenge point was located manually by either the clinicians or the patients. Similar principles apply in cognitive tasks, though these have received a slightly more elaborate treatment, and many ways to automate the challenge level of tasks have been proposed. Baker et al. (2008) evaluate the knowledge level of students as a latent variable and learn to predict the next task to give to a student given past performances [8]. Segal et al. (2014) use collaborative filtering to match a student with a next task given past experiences of other students with similar performance level [22]. Other works did not focus on task sequencing, but rather on when to provide advice to students when they seem to be stuck [5], [28]. The recurring idea behind all of these works is that *the learner should always be challenged enough to avoid boredom and to promote learning, but not too much to cause frustration*. This idea has many different names in the literature – challenge point, flow channel, difficulty adjustment, zone of proximal development, and more [4], [12], [31], [34]. However, to the best of our knowledge, this idea was never explicitly formalized into a computational model for general motor learning tasks. Gentile [11] presented a taxonomy to categorize difficulty of different motor tasks, and proposed that this can be used, in a similar manner to the Challenge Point Theory, to identify a suitable progression of tasks. This taxonomy asserts that the amount of information the learner gets from the task depends on a combination of a learner's level and the task environment. As the learner's skill level cannot be controlled, reaching this optimal challenge point means that the task of the teacher is to find the optimal task environment given the learner's skill level. This work did not formalize or quantify the learner's skill level or the environment difficulty. In educational research, the Zone of Proximal Development [31] is one of the ideas behind educational scaffolding [30]: to present a student with examples and tests that are challenging enough to promote development, but are not so hard that the students are discouraged.

### B. Curriculum Learning in Reinforcement Learning

Several recent works have been proposed to learn a curriculum for reinforcement learning agents. Teacher-Student Curriculum Learning (TSCL) is a framework for automatic curriculum learning, where the student tries to learn a complex task, and the teacher automatically chooses subtasks from a given set for the student to train on [16]. This work was used to train a Long-Short Term Memory (LSTM) network, and RL agents for playing Minecraft. Narvekar et al. (2017) formulated the design of a curriculum as a Markov Decision Process, which directly models the accumulation of knowledge as an agent interacts with tasks, and proposed a method that approximates an execution of an optimal policy in this MDP to produce an agent-specific curriculum [19]. A later work demonstrated how several different representations can be used to learn a curriculum policy for multiple agents [20]. All of these works were used to train artificial agents on computational tasks. In this work, we leverage a formalization similar to the ones presented in these works to train a human agent on motor task learning.

### C. Motor Task Tuition as a Reinforcement Learning Problem

In this paper, we formalize the curriculum learning problem using an RL formulation that models how the various tasks that can be presented to the learner impact learning. Following this formulation, in order to maximize learning, the task of the teacher becomes choosing which of these tasks to present next such that it will be in the challenge point of the learner. The most relevant work to ours may be a formulation of a motor task as a multi-armed bandit problem, which is a simple version of an RL problem, where there is only one state. In this work, the decision on the next task to choose for the learner only depends on the current performance of the learner, rather than on some future goal [23]. As we are interested in constructing a complete curriculum in advance so that the learning process towards some target motor task is efficient, this work is less suitable. Other related work has applied optimization principles to derive curricula that attempt to maximize long-term retention of learning when practicing multiple different tasks [14].

## III. BACKGROUND

In this section, we provide background on reinforcement learning and curriculum selection, and discuss how these ideas can be used to improve human training for motor tasks.

### A. Reinforcement Learning (RL)

Reinforcement learning is a paradigm for learning sequential decision making tasks for an artificial agent acting in an environment. It models a *task* as a Markov Decision Process (MDP) [26]. A MDP $M$ is a 4-tuple $(\mathbb{S}, \mathbb{A}, p, r)$, where $\mathbb{S}$ is the set of states in the environment, $\mathbb{A}$ is the set of actions the agent can take, $p(s'|s,a)$ is a transition function that gives the probability of transitioning to state $s'$ after taking action $a$ in state $s$, and $r(s,a,s')$ is a reward function that gives the immediate reward for taking action $a$ in state $s$ and transitioning to state $s'$.

At each time step $t$, the agent observes its state and chooses an action according to its *policy* $\pi_\theta(a|s)$, which we assume is parameterized by $\theta$. The goal of the agent is to learn an *optimal policy* $\pi^*$, which maximizes the expected *return* (cumulative sum of rewards) until the episode ends. The optimal policy can be learned through value function methods (such as $Q$-learning [32]) or policy search methods (such as deterministic policy gradients [25]).

### B. Curriculum Learning in RL

In order to learn which actions to take, an RL agent must explore the environment and accumulate rewards. In some tasks, learning can be difficult due to sparse rewards or the presence of adversarial agents or elements. One way to accelerate learning in these settings is to first train the agent on an easier source task. This task might require fewer actions to reach the goal, or have fewer adversaries or elements in

the environment that the agent needs to learn about. The knowledge acquired in this simpler environment can then be *transferred* to improve learning on the challenging target task. Transfer learning [29] is an area of research dedicated to how an agent can transfer knowledge between tasks.

Instead of just training on one source task, an agent can train on a *sequence* of source tasks, where each subsequent task becomes progressively harder and builds upon skills learned in previous tasks. This choice of a sequence of tasks is called a *curriculum*. Curriculum learning [18] is a methodology to optimize the order in which tasks are presented to the agent, so as to improve learning speed or performance on a final target task.

In this paper, we draw inspiration from a model for curriculum design [20] that poses curriculum generation as an interaction between two MDPs. The first is an MDP for the *student* agent, which is the recipient of the curriculum. This agent interacts in the standard way with a given task. The second is a higher level curriculum MDP (CMDP) for the *teacher*, whose goal is to select tasks for the student to train on. Formally, a CMDP can be defined as follows:

*Definition 1:* A curriculum MDP (CMDP) $M^C$ is a tuple $(\mathbb{S}^C, \mathbb{A}^C, p^C, r^C)$ where:

**State Space** The set of states $\mathbb{S}^C$ represent the state of the student agent's knowledge or learning progress. For RL agents, the student agent's knowledge is represented by its policy $\pi$. Terminal states are those where the student's policy is able to surpass a specified performance threshold on the target task.

**Action Space** The set of actions $\mathbb{A}^C$ are the possible source tasks the student can train on.

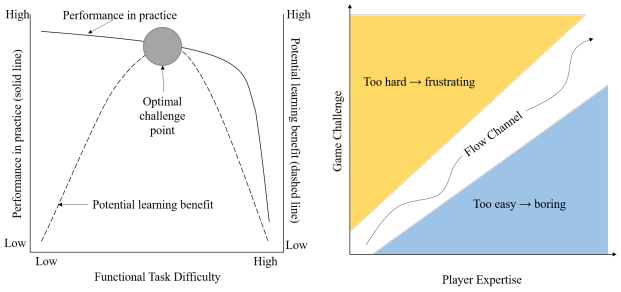
**Transition Function** The transition function $p^C(s^C, a^C, s'^C)$ gives the probability that $s'^C$ is the student agent's policy after training on task $a^C$ and starting with policy $s^C$.

**Reward Function** The reward function $r^C(s^C, a^C, s'^C)$ gives a scalar reward for the transition where the student starts with policy $s^C$, trains on $a^C$, and updates its policy to $s'^C$. We define a reward function to maximize asymptotic performance by rewarding transitions into terminal states by the final score achieved on the target task.

In this model of curriculum learning, the goal is to learn a policy over a CMDP, which specifies which task the student should train on next given it's current state of knowledge, so as to maximize performance on the target task. This model was inspired by human learning, and designed to improve training for autonomous reinforcement learning agents. In this work, we take the complementary view and aim to train human agents by adapting curriculum learning methods designed for RL agents to the human setting. In the curriculum MDP model, this effect is achieved by replacing the RL student agent with a human learner.

## IV. CURRICULUM SELECTION FOR HUMAN MOTOR LEARNING

Many frameworks and heuristics have been proposed to promote human motor and cognitive learning via task selection. The two most popular are the Challenge Point Framework [12] and the Flow Channel [7]. The challenge point framework (Figure 1a) suggests that the effect of practice conditions in learning a new task depends on two conditions: the skill level of the subject, and the task difficulty. The authors further hypothesize that for any subject learning a given task, an optimal challenge point exists where the potential learning benefit is maximized:




(a) Challenge point framework (b) Flow channel

Fig. 1: Theories on human skill and performance in literature: Fig. 1a depicts the relation between learning and performance curves according to the challenge point theory [12]; Fig. 1b depicts the flow channel concept [34].

*Proposition 1:* For any given skill level (state) of the learner, there exists a task difficulty level (source task) at which the learning can be maximized.

The Flow Channel theory (Fig. 1b), which is commonly used in game design, suggests that in order to maintain the game saliency towards the player and maximize engagement, the game must remain in a 'flow channel' by optimizing the difficultly level based on player's skill level:

*Proposition 2:* The choice of the optimal source task progresses from lower to higher difficulty level as the learner's skill level increases.

Using these two theories as motivation, we define a CMDP where the student is no longer an RL agent with its own MDP, but a human. However, the adaptation is not straightforward because of the challenges involved in formalizing the CMDP components for the human learner. For example, in RL agents, the state is usually defined as the agent's policy - which represents the agent's state of knowledge. With the agent's policy available, one can predict the agent's behavior in each possible state. However, for humans, there is no perfect way to fully capture the learner's state of knowledge. Similar issues arise when we try to define the transition dynamics and reward function for the CMDP for humans. As a first step, we propose to formalize a CMDP for human motor learning such that a person's performance on the target task will be used as a proxy to the state of knowledge, and thus it provides only partial information about the person's abilities:

*Definition 2:* The Human Curriculum Markov Decision Process (H-CMDP) $M^C$ is a tuple $(\mathbb{S}^C, \mathbb{A}^C, p^C, r^C)$ where:

**State Space** State $s^C$ is defined as the learner's score on the target task. The set of all possible scores in that task

defines the state space ($S^C$). For a given target task, we define a threshold score $\overline{s^C}$, such that all $s^C \geq \overline{s^C}$ are considered *terminal states*, meaning that the human was able to achieve the threshold score or higher.

**Action Space** The action space ($\mathbb{A}^C$) is defined by the possible source tasks the subject can train on.

**Transition Function** The transition function $p^C(s^C, a^C, s'^C)$ gives the probability that the subject will obtain a score of $s'^C$ on the diagnostic task (target task) after training on $a^C$ and starting with a score $s^C$.

**Reward Function** The reward $r^C(s^C, a^C, s'^C)$ for the teacher is defined for the specific threshold score that we wish the learner to achieve, $\overline{s^C}$. Then

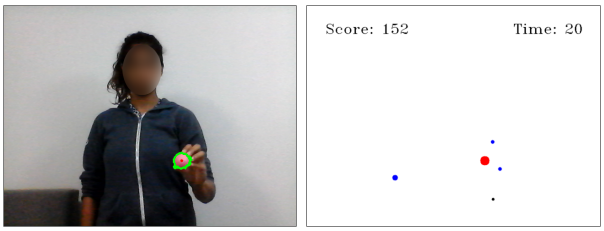
$$r^C(s^C, a^C, s'^C) = \begin{cases} s'^C, & s'^C \geq \overline{s^C} \\ 0, & otherwise \end{cases}$$

Notice that the target task is also part of $\mathbb{A}^C$ and can be chosen at any point as the next task to train on. According to Proposition 1, for every $s^C$ current score of the learner, there is an optimal source task $a^{C*}$ that will maximize the learning of the agent. This claim is a fundamental assumption used by many RL-based curriculum learning algorithms that aim to find such an optimal source task. Following Proposition 2, we choose the target task to be the most difficult task in $\mathbb{A}^C$. Every time the learner tries this task, it is used as a probe to estimate the skill level or state of knowledge of the learner and to get a new accurate value $s^C$. However, we cannot get the true $s^C$ after each task, as the learner continues to learn both while training on the target task and on other tasks. This challenge is beyond the scope of this paper, and we only evaluate the learner's performance in the beginning and in the end of a sequence of tasks, as we detail in the experimental design.

## V. THE REACH NINJA DOMAIN

Using our proposed formulation of a H-CMDP, we now present an example domain for evaluating different target tasks, learners, and curricula in a specific motor learning skill. We refer to this domain as *The Reach Ninja*, as it corresponds with both existing work on reaching as a motor skill [1], [21] and with the popular game "Fruit Ninja" [10]. Given the current global pandemic and keeping in mind the safety of the subjects, the task was designed to be conducted online without the need for in-person interaction.

*a) Task Setup:* The OpenCV Library [6] was used in conjunction with a webcam in order to track the subject's motion during experimentation. The method tracks an object held in the subject's hand (Fig. 2a), and is visualized on the screen in real time (red marker in Fig. 2b). The goal of this game is to gather points by reaching with the red marker to the blue markers on the screen, while avoiding black markers. These settings enable a variety of difficulty levels and potential target tasks. Next, we detail the design of a target task that we hypothesize to be challenging enough so it is non-trivial and can benefit from a curriculum, but not too challenging that subjects are unable to improve.




(a) Webcam view (b) On-screen feedback

Fig. 2: Gameplay setup: Fig. 2a shows the subject playing the game. The area marked by the green border tracks the subject's hand location and corresponds to the red marker in the on-screen feedback. Fig. 2b shows the feedback provided to the subject as they play the game. The red marker shows the tracked position of the hand, and the blue marker shows a virtual target.

*b) Task Selection:* Following Proposition 2, we define our target task as the hardest task out of a pool of potential tasks we can present to the subject. However, difficulty is not trivial to determine and likely varies among participants. For example, one participant might find that adding velocity to the blue markers makes the task more challenging than adding more negative markers, but the reverse may be true for another participant. To overcome this challenge, we designed a target task that combines several modifications of the task dynamics and held certain parameters constant while allowing others to vary.

*c) Task Dynamics:* A baseline task was created to emulate the static reaching task in earlier studies [1]. The subject would see a single static blue target marker on the screen and would try to reach it by moving their hand. Each successful encounter between the subject's hand (red marker) and the target (blue marker) resulted in an increase in the score. This score is depicted in the top left corner of the game screen (Fig. 2b). The remaining duration of the game is also shown on the top right corner of the game screen. We introduced a number of additional task components to increase the difficulty of the task beyond this simple point-to-point task:

**Target marker motion:** The blue virtual markers spawn at random locations at the bottom of the screen and are given a constant velocity. The orientation and magnitude of this velocity are randomly selected such that the markers travel across the screen. The subject thus has to move as fast as or faster than the markers in order to reach them.

**Target marker acceleration:** The markers are further acted upon with a constant downward acceleration imparting projectile motion.

**Negative marker:** 'Negative' markers, represented in black, are randomly created similar to the blue target markers. However, when the subject encounters these negative markers, they lose points instead of gaining them, and the game freezes for a duration of 2 seconds.

**Cursor Visibility:** We manipulated visibility of the cursor

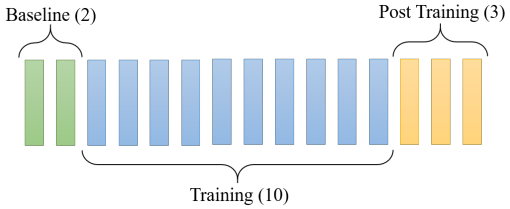



Fig. 3: Experimental protocol over 15 sessions: 2 baseline; 10 training; 3 post training of 60 second duration each.

in order to increase the difficulty of the task. The red marker is displayed on screen for a fixed period of time (we used 0.8 seconds in our experiments) and then disappears for another period (we used 0.2 seconds), without affecting game play. The cursor disappearance would require participants to be able to predict the cursor location based on their internal model of their actions [17].

These high level task components gave rise to several task parameters that could be changed to affect game play, such as marker size, marker velocity, downward acceleration, percentage of negative markers spawned, duration of partial observability, and more. The selected task presented in our empirical study used both target and negative markers that displayed projectile motion, while the participant's position was only partially available as feedback. The starting velocities and sizes of the markers were allowed to vary randomly, while the acceleration, percentage of negative markers and periods of observability for feedback were held constant. The game created based on the combination of these components will be referred to as the target task.

## VI. EXPERIMENTAL DESIGN

The first goal of our experiments is to identify a suitable target task that satisfies the condition of being difficult, but not too difficult. We hypothesize that participants will be able to improve their performance on the target task by practicing it repeatedly. This improvement would suggest that cognitive and motor adaptation and learning are taking place over the course of the training. If learning is in fact occurring, based on the background presented from existing literature, we expect a curriculum that is able to keep the learner within the flow channel (or to present the right challenge point) to further augment this learning. The second goal of our experiments is to demonstrate that a manually designed curriculum, following principles of curriculum design, improves the participants' learning. We hypothesize that a manually designed curriculum will result in greater eventual improvement on the target task compared to only ever practicing the target task.

*a) Experimental Protocol:* The experimental protocol shown in Fig. 3 was designed to inform both these goals. The subjects are asked to participate in 15 sessions, each session lasting for one minute. During the first two sessions, the subjects are exposed to the target task in order to collect baseline data on their performance. Following the baseline, the participants were asked to perform 10 training sessions depending on their experimental group: control or curriculum. Subjects in the control group practice with the same target task for 10 sessions. Subjects in the curriculum group are given a manually designed curriculum for the 10 training sessions as discussed in the following section. Subjects from both groups then play the target task for the final 3 sessions, referred to as the post-training sessions.

*b) Manual Curriculum Design:* The subjects in the curriculum group were asked to perform a set of training tasks that we manually selected based on the modifications used to create the final task. Five training tasks were administered, for two sessions each, in the following order:

1) The target and negative markers were held stationary and the subject received full feedback of the their hand position.
2) The target and negative markers were held stationary, but the subject received only partial feedback of the their hand position (0.5 seconds on, 0.5 seconds off).
3) The target and negative markers were imparted random velocity magnitudes in the vertically upward direction, while being acted upon by a downward acceleration. The subject received full feedback of their hand position.
4) The target and negative markers were imparted random velocity magnitudes in the vertically upward direction, while being acted upon by a downward acceleration. The subject received partial feedback of their hand position (0.5 sec on, 0.5 sec off).
5) The target and negative markers were imparted random velocities both in magnitude and direction, while being acted upon by a downward acceleration. The subject received full feedback of their hand position.

## VII. PILOT STUDY RESULTS

In this section, we detail the results of running this protocol with two participants, from among the authors of this paper, as a pilot study towards a larger experiment. These participants were randomly assigned to either the control or the curriculum group. Fig. 4 shows the performance of the subject from the control and the subject from the curriculum group. The X axis displays the name of the session shown on the plot (BL1 and BL2 refer to Baseline attempt 1 and 2 respectively, and PT1, PT2, PT3 refer to Post-Training 1, 2, and 3 respectively). The solid line shows the change in average end of session score from baseline to post training. The blue lines refer to the control subject and red refer to the curriculum subject.

Both subjects showed an improvement in average performance from baseline to post training (Fig. 4). If this result holds in larger studies, we may conclude that training does result in improvement, suggesting that practice does indeed enable participants to improve their performance through practice, regardless of the curriculum followed. Further, we see that the improvement seen in the curriculum subject (of 53 points) is higher than that seen by the control subject (39 points). This difference across these two pilot participants

is consistent with our hypothesis that an appropriate curriculum can improve learning. In future work, we will more rigorously test this hypothesis using a much larger cohort of participants. It is also worth noting that improvement is easier to achieve if a player begins at a lower baseline performance, rather than at a higher baseline performance. This effect makes it challenging to compare two participants where one started with a score of 40 and improved to 50, while the other started at 50 and improved to 55. In our pilot experiment, the curriculum subject started with a higher score and showed higher average performance during the baseline sessions, so in this case it is clear that there was a larger absolute improvement in the curriculum subject's performance over the baseline subject's performance.

In order to delineate the effect of the curriculum, we compared the technique used by both subjects before and after training. By averaging the total score with respect to the number of times the subject reached a marker (whether a target or a negative), we were able to infer the subjects' strategies. For example, if a participant had hit a marker a total of 10 times over a session, and had a total score of 80, the participant's score increase per hit is 8 (Fig. 5). Again, the blue line depicts the control subject's performance and red depicts the curriculum subject's performance. The high variability in the control subject's average increase per hit across the two baseline and three post-training sessions is interesting. The variability suggests that the subject did not learn a strategy to consistently hit markers that would result in a better score [24]. Instead, the subject's better performance in the post-training sessions may be attributed to simply hitting more markers without regard for the differences in value due to size, speed, or type of marker. On the other hand, the curriculum subject's data shows that while the strategy the subject used was similar in the two baseline sessions, resulting in an almost constant score increase per hit, there was an improvement during the post-training sessions. This result suggests that the curriculum subject learned to both hit more targets as well as better targets, i.e. to get a higher score per successful hit.

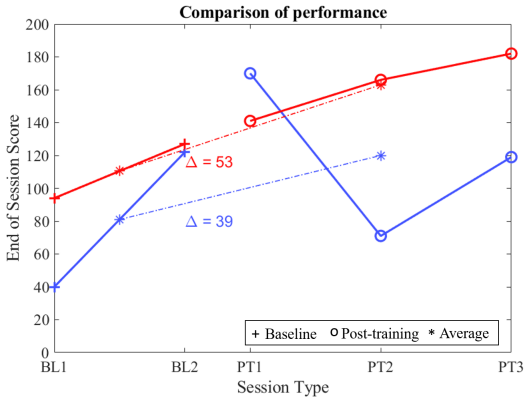



Fig. 4: Results from the pilot experiments (blue: control subject; red: curriculum subject) show that the scores of the subjects at the end of the baseline (BL) and post training (PT) sessions as well as the average performances

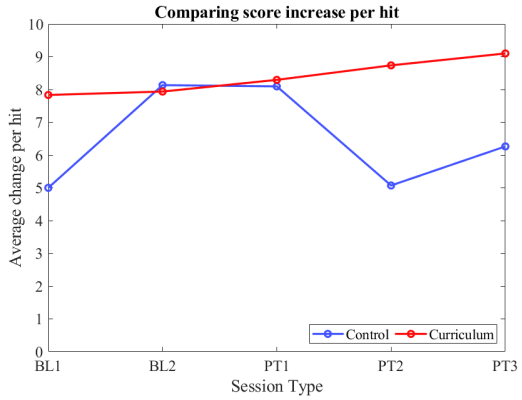



Fig. 5: Results from the pilot experiments (blue: control subject; red: curriculum subject) show that the average change in the subject's score for every marker hit in each of the 5 sessions.

## VIII. DISCUSSION

There are many similarities between task sequencing in human motor learning and curriculum learning using RL: they both try to maximize similar objectives (proficiency level, accuracy, time to target, etc.); they both use similar notions (incremental difficulty adjustment); and they both consist of a set of source tasks and a target task. In this paper, we presented a formal description of a human motor task curriculum learning using a MDP, namely H-CMDP. We discussed the key differences from CMDP learning for RL agents, such as the learner's knowledge level that is not directly observable and needs to be evaluated by proxies. We then showed how common motor learning concepts, such as an optimal challenge point, can be compiled and represented in the H-CMDP model. Lastly, we presented a rich test-bed environment, the "reach ninja" game, that is suitable for evaluating different curricula without the need for in-person interaction with subjects, given the safety restrictions of the current pandemic.

Our H-CMDP formulation is a first step to bridge the gap between task sequencing for motor learning and curriculum learning using RL. We are currently investigating how existing curriculum learning approaches are used, and how they can be modified to handle people instead of RL agents as learners, which is not a trivial shift. For example, these works make assumptions that are unrealizable in practice for motor learning, such as an ability to execute each possible action infinitely often and having a complete, accurate representation of the learner's knowledge. In addition, due to attention span, fatigue, and other constraints, people cannot train for extensive amounts of time. Thus, curriculum learning that allows for subjects to meet their potential will have to use relatively small amount of data and to provide a curriculum that is appropriate to human abilities.

## ACKNOWLEDGMENT

This work has taken place in the ReNeu Robotics Lab and Learning Agents Research Group (LARG) at UT Austin. Effort in the ReNeu Lab is supported, in part, by Facebook Reality Lab. LARG research is supported in part by NSF (CPS-1739964, IIS-1724157, NRI-1925082), ONR (N00014-18-2243), FLI (RFP2-000), ARO (W911NF-19-2-0333), DARPA, Lockheed Martin, GM, and Bosch. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

## REFERENCES

[1] Farnaz Abdollahi, Emily D Case Lazarro, Molly Listenberger, Robert V Kenyon, Mark Kovic, Ross A Bogey, Donald Hedeker, Borko D Jovanovic, and James L Patton. Error augmentation enhancing arm recovery in individuals with chronic stroke: a randomized crossover design. *Neurorehabilitation and Neural Repair*, 28(2):120–128, 2014.

[2] Deborah L Adams. Develop better motor skill progressions with gentile's taxonomy of tasks. *Journal of Physical Education, Recreation & Dance*, 70(8):35–38, 1999.

[3] Priyanshu Agarwal and Ashish D Deshpande. Subject-specific assist-as-needed controllers for a hand exoskeleton for rehabilitation. *IEEE Robotics and Automation Letters*, 3(1):508–515, 2017.

[4] Priyanshu Agarwal and Ashish D Deshpande. A framework for adaptation of training task, assistance and feedback for optimizing motor (re)-learning with a robotic exoskeleton. *IEEE Robotics and Automation Letters*, 4(2):808–815, 2019.

[5] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara J Grosz. Interactive teaching strategies for agent training. In *Proceedings of the Twenty-Fifth international joint conference on Artificial Intelligence (IJCAI'16)*. International Joint Conferences on Artificial Intelligence, 2016.

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[7] Ben Cowley, Darryl Charles, Michaela Black, and Ray Hickey. Toward an understanding of flow in video games. *Computers in Entertainment (CIE)*, 6(2):1–27, 2008.

[8] Ryan SJ d Baker, Albert T Corbett, and Vincent Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *International conference on intelligent tutoring systems*, pages 406–415. Springer, 2008.

[9] Ronit Feingold Polak and Shelly Levy Tzedek. Social robot for rehabilitation: Expert clinicians and post-stroke patients' evaluation following a long-term intervention. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 151–160, 2020.

[10] fruit ninja. *Fruit Ninja*, Accessed July 2020. `https://fruitninja.com/`.

[11] AM Gentile. Skill acquisition: Action, movement, and the neuromotor processes. *Movement Science: Foundations for Physical Therapy in Rehabilitation*, 1987.

[12] Mark A Guadagnoli and Timothy D Lee. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior*, 36(2):212–224, 2004.

[13] John W Krakauer, Alkis M Hadjiosif, Jing Xu, Aaron L Wong, and Adrian M Haith. Motor learning. *Comprehensive Physiology*, 9:613–663, 2019.

[14] Jeong Yoon Lee, Youngmin Oh, Sung Shin Kim, Robert A Scheidt, and Nicolas Schweighofer. Optimal schedules in multitask motor learning. *Neural computation*, 28(4):667–685, 2016.

[15] Laura Marchal-Crespo and David J Reinkensmeyer. Review of control strategies for robotic movement training after neurologic injury. *Journal of Neuroengineering and Rehabilitation*, 6(1):20, 2009.

[16] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning. *IEEE transactions on neural networks and learning systems*, 2019.

[17] Daniel McNamee and Daniel M Wolpert. Internal models in biological control. *Annual review of control, robotics, and autonomous systems*, 2:339–364, 2019.

[18] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *arXiv preprint arXiv:2003.04960*, 2020.

[19] Sanmit Narvekar, Jivko Sinapov, and Peter Stone. Autonomous task sequencing for customized curriculum design in reinforcement learning. In *International Joint Conference on Artificial Intelligence*, pages 2536–2542, 2017.

[20] Sanmit Narvekar and Peter Stone. Learning curriculum policies for reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 25–33. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[21] James L Patton, Yejun John Wei, Preeti Bajaj, and Robert A Scheidt. Visuomotor learning enhanced by augmenting instantaneous trajectory error feedback during reaching. *Public Library of Science one*, 8(1):e46466, 2013.

[22] Avi Segal, Ziv Katzir, Kobi Gal, Guy Shani, and Bracha Shapira. Edurank: A collaborative filtering approach to personalization in e-learning. In *Educational Data Mining 2014*, 2014.

[23] Yoones A Sekhavat. Mprl: Multiple-periodic reinforcement learning for difficulty adjustment in rehabilitation games. In *2017 IEEE 5th international conference on serious games and applications for health (SeGAH)*, pages 1–7. IEEE, 2017.

[24] Lior Shmuelof, John W Krakauer, and Pietro Mazzoni. How is a motor skill learned? change and invariance at the levels of task success and trajectory control. *Journal of neurophysiology*, 108(2):578–594, 2012.

[25] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, 2014.

[26] Richard S Sutton and Andrew G Barto. Reinforcement learning: an introduction MIT Press. *Cambridge, MA*, 1998.

[27] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[28] Matthew E Taylor, Nicholas Carboni, Anestis Fachantidis, Ioannis Vlahavas, and Lisa Torrey. Reinforcement learning agents providing advice in complex video games. *Connection Science*, 26(1):45–63, 2014.

[29] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

[30] Irina Verenikina. Understanding scaffolding and the zpd in educational research. 2003.

[31] Lev Vygotsky. Zone of proximal development. *Mind in society: The development of higher psychological processes*, 5291:157, 1987.

[32] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[33] Eric T Wolbrecht, Vicky Chan, David J Reinkensmeyer, and James E Bobrow. Optimizing compliant, model-based robotic assistance to promote neurorehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(3):286–297, 2008.

[34] Mohammad Zohaib. Dynamic difficulty adjustment (DDA) in computer games: A review. *Advances in Human-Computer Interaction*, 2018, 2018.