Real-time Prediction Learning for the Simultaneous Actuation of Multiple Prosthetic Joints

Patrick M. Pilarski, Travis B. Dick, and Richard S. Sutton

Abstract— **Integrating learned predictions into a prosthetic control system promises to enhance multi-joint prosthesis use by amputees. In this article, we present a preliminary study of different cases where it may be beneficial to use a set of temporally extended predictions—learned and maintained in real time—within an engineered or learned prosthesis controller. Our study demonstrates the first successful combination of actor-critic reinforcement learning with real-time prediction learning. We evaluate this new approach to control learning during the myoelectric operation of a robot limb. Our results suggest that the integration of real-time prediction and control learning may speed control policy acquisition, allow unsupervised adaptation in myoelectric controllers, and facilitate synergies in highly actuated limbs. These experiments also show that temporally extended prediction learning enables anticipatory actuation, opening the way for coordinated motion in assistive robotic devices. Our work therefore provides initial evidence that real-time prediction learning is a practical way to support intuitive joint control in increasingly complex prosthetic systems.**

I. INTRODUCTION

The natural movement of human limbs relies on the coordination and scheduling of the multiple actuators that drive their motion. Synergies (both learned and innate) are thought to govern the timing and fluid motion of joints, ensuring that each joint is properly aligned throughout complex sequences of human movement [1], [2]. Elegant coordination of this kind is currently missing from most if not all prosthetic devices, and, in particular, myoelectric prostheses—electromechanical devices that map surface electromyographic (EMG) signals to control commands for a robotic appendage. The problem of naturally coordinating multiple actuators becomes more critical as prosthetic technology improves to meet the functionality, control, and feedback needs of amputee users [3]–[6]. New experimental prosthetic systems have actuation capacities that approach those of a biological limb [7], [8]. However, even with surgical advances toward new human-machine interfaces [9], [10], it remains challenging for amputees to provide a rich set of signals as control commands for a highly actuated robotic limb. As amputation becomes more severe, the number of signals an amputee can provide decreases, further limiting the functionality and controllability of their myoelectric devices; the simultaneous proportional control of multiple actuators is still an open and challenging problem [3], [6].

As suggested by contemporary literature on motor planning in humans, an important part of coordinating and planning actions may be anticipatory movements enabled by motor predictions in the brain [2], [11], [12]. Predictions are also thought to be learned as a precursor to motor skill acquisition in humans, with prediction learning occurring at a faster timescale than control-related learning [11]. A similar integration of learned anticipatory predictions into the control of prosthetic actuators could alleviate some of the barriers to multi-joint prosthesis use by amputees, especially amputees with limited signal recording sites on their amputated limbs.

Prediction is not restricted to use by autonomous biological systems—prediction and anticipation play a major role in recent advances in automation research, robotics, and machine intelligence. There is growing international interest in predictive approaches to robotic control (e.g., [13]–[15]), including ideas extended from classical model-predictive control [16]. Typically, methods for predictive robot control involve predictions that span only a single timestep into the future, and most require that predictive models be computed offline prior to use [17]. In contrast, recent work by our group has demonstrated computationally efficient methods for learning and maintaining a large set of temporally extended predictions in real time [17]–[21]. There are a number of compelling reasons for seeking to integrate online or offline predictions into a robotic control system. Of note, predictive representations of state have been demonstrated to improve the performance and learning speed of reinforcement learning systems, successfully compressing (generalizing) large state spaces and enabling data-driven learning methods with both high representational capacity and compactness [22]–[24]. Prediction learning is also a key part of much of the present body of work on real-valued and classifier-based myoelectric decoding, wherein predictions of instantaneous motor intent are distilled from the complex signal space being received from an amputee's body and their robotic device [4], [6], [13], [25].

While offline prediction learning and engineered predictive control has received much attention in the literature (as noted above), the integration of real-time prediction learning into myoelectric control is an area with a number of remaining open questions. In particular, no studies have been conducted to evaluate how real-time prediction learning methods can be deployed *in a general sense* to enable intuitive multi-joint prosthetic control. Studies are also needed to objectively assess the costs and benefits of adding a secondary learning system to engineered or learning-based myoelectric controllers—online prediction learning entails a new subsystem with related parameters to optimize and balance.

In this article, we therefore seek to provide a first set

P. M. Pilarski, T. B. Dick and R. S. Sutton are with the Reinforcement Learning and Artificial Intelligence Laboratory, Department of Computing Science, University of Alberta, Edmonton, AB, T6E 2E8, Canada.

Please direct correspondence to: pilarski@ualberta.ca

Proceedings of the 2013 IEEE International Conference on Rehabilitation Robotics (ICORR), Seattle, USA, June 24–26, 2013, to appear.

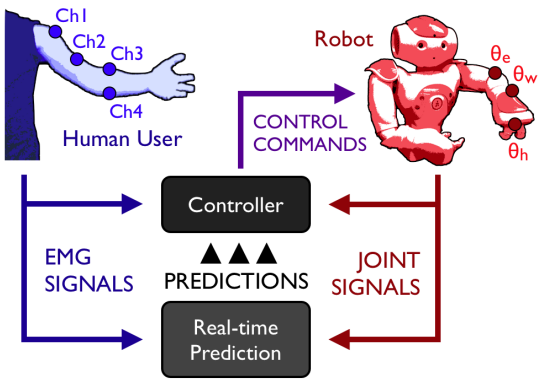

Fig. 1. Schematic showing the flow of information through the experimental setup. A controller forms new robot joint velocity commands using myoelectric information from the human user, sensor readings from the robot limb, and predictions from a real-time machine learning system.

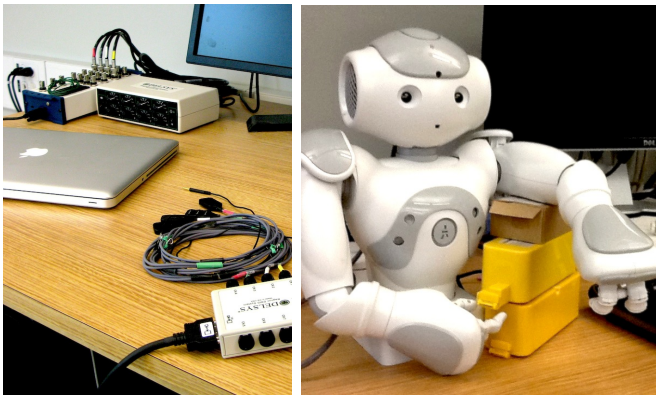

Fig. 2. Experimental setup for these experiments, which includes an Aldebaran Nao T14 Robot Platform and a DelSys Bagnoli 8 EMG system.

of insights and preliminary answers to the following open questions regarding the integration of real-time prediction learning into multi-joint myoelectric control:

- Can real-time prediction learning improve the quality of learned or engineered prosthesis control policies?
- Can real-time prediction learning allow a prosthetic system to adapt during online operation without the need to re-learn or re-calibrate its controller?
- Can real-time prediction learning facilitate the control of multiple actuators that exceed the control dimensions available to a human user (or effect synergies of the kind suggested by d'Avella et al. [1])?
- Can the use of temporal extension in real-time prediction learning help a controller plan anticipatory movements for more natural, coordinated multi-joint motion?
- Can the use of real-time prediction learning allow the compression of prosthetic input signal spaces into fewer content-rich signals to speed controller learning (as in low-dimensional embeddings [1], [25])?
- Do the benefits of adding a second learning subsystem promise to outweigh the increased complexity and optimization challenges it entails?

We explore these questions within the context of human-driven myoelectric control, using control-learning techniques from continuous-action Actor-Critic Reinforcement Learning (ACRL). Our intent is not to present the reader with a perfect control approach for simultaneous joint actuation; rather we hope to shed new light on the different ways real-time prediction learning methods can be used to support intuitive joint control in increasingly complex prosthetic systems.

II. METHODS

As shown in Fig. 1, the context for our experiments was the real-time human myoelectric control of a multi-joint robot arm. In this setting, the number of control channel pairings available to the human user was less than the number of control dimensions enabled on the robot system. Able-bodied subjects were asked to freely actuate two joints of the robot arm using classical myoelectric control. The control system was tasked with learning and maintaining a control policy for one additional actuator, such that the actuator's instantaneous position was appropriate given the user's action choices. Suitability of the achieved configuration was evaluated in terms of a set of real-valued angular targets for the third joint—desired angles were assigned for all human-controlled effector positions in advance of the task, as described below.

The control system received joint angle inputs $\langle \theta_e, \theta_h, \theta_w \rangle$ from the robot platform (Nao T14, Aldebaran Robotics, France) and EMG signals $\langle$Ch1,...,Ch4$\rangle$ from a recording system attached to the human subject (Bagnoli 8 EMG System with DE3.1 electrodes, DelSys, USA). Output control commands were actuator velocity signals $\langle \dot{\theta}_e, \dot{\theta}_h, \dot{\theta}_w \rangle$. This experimental setup is shown in Fig. 2. As described in further detail below, for a subset of the explored control approaches the controller also received a set of one or more predictions $\langle P_e, P_h, P_w, ... \rangle$ learned in real-time during each trial.

Myoelectric control of the robot's elbow and hand actuators was performed by each subject via conventional linear-proportional control. Using the specific approach described previously by Pilarski et al. [20], [26], mean-absolute-value signals recorded from the biceps, deltoid, wrist flexor and wrist extensor muscle groups were mapped to joint velocity commands using the muscle pairings bicep/deltoid and wrist flexor/extensor, shown as Ch1/2 and Ch3/4 respectively in Fig. 1. The subjects were free to move the two joints under their control, but had no direct myoelectric control over the robot's wrist actuator.

Wrist targets, denoted θ_w^*, were specified as a function of θ_e and θ_h to emulate the configurations achieved by a natural limb in simple grasping-and-placing functional tasks. When the user-controlled joints were both moved to within 0.1rad of one of their endpoints, the new values of θ_e and θ_h we used to deterministically select a contingent wrist target according to a pre-defined function. For example, when the user effected an open hand and fully extended the robot's elbow, the desired wrist pose was set to be palm down (as if to grab an object resting on a flat surface), while when the user effected an open hand and bent elbow the desired wrist configuration was palm up (as if passing the grasped object to another person).

A. Learning a Control Policy in Real Time using ACRL

ACRL is an approach to control learning that is both computationally suitable for real-time deployment (incremental and linear in both processing and memory) and also allows the learning of continuous, real-valued control policies that are appropriate for robot actuator control. One contribution of the present work is to assess the feasibility of ACRL for intuitive multi-actuator control during direct patient-robot interactions, extending our previous demonstrations with pre-recorded myoelectric data and offline training [26]. For our present experiments, the system's action policy was learned online using previously described methods for deploying standard ACRL within the context of myoelectric control learning [26], optimized as per observations by Degris et al. [27]. These methods are summarized as follows.

Control actions proposed by the learning system, denoted a, were sent to the robot as joint velocity signals, in this case real-valued commands indicating the desired wrist velocity $\dot{\theta}_w$. At each time step, actions were drawn from a normal distribution; parameters of the normal distribution were functions of the system's learned weight vectors $\mathbf{w}_\mu$ and $\mathbf{w}_\sigma$ as given by $\dot{\theta}_w \approx a \leftarrow \mathcal{N}\{\mu = \mathbf{w}_\mu^T \mathbf{x}(\mathbf{s}), \sigma = \exp[\mathbf{w}_\sigma^T \mathbf{x}(\mathbf{s})]\}$, where $\mathbf{x}(\mathbf{s})$ was a linear tile-coded function approximation of the input signal space s [26]. To allow the system to better explore the task space during learning, selected actions persisted for three time steps ($\sim$100ms). The learning system was presented with a reward signal r, the state approximation for the current time step $\mathbf{x}(\mathbf{s})$, and a state approximation for the subsequent time step $\mathbf{x}(\mathbf{s}')$. Standard temporal-difference learning was used to update the weight vector $\mathbf{v}$ in the critic. Actor weight vectors $\mathbf{w}_\mu$ and $\mathbf{w}_\sigma$ were updated according to compatible features for the policy parameterization (in this case, the normal distribution). Replacing eligibility traces ($\mathbf{e}_\mathbf{v}$ in the critic) and accumulating eligibility traces ($\mathbf{e}_\mu, \mathbf{e}_\sigma$ in the actor) were used to accelerate learning [27]. Trace decay rates $\lambda_\mathbf{v}, \lambda_\mathbf{w}$ and step sizes $\alpha_\mathbf{v}, \alpha_\mu, \alpha_\sigma$ (determined empirically) governed the magnitude of the weight vector updates. A single step of this incremental learning procedure was implemented as follows:

$\delta \leftarrow r + \gamma \mathbf{v}^T \mathbf{x}(\mathbf{s}') - \mathbf{v}^T \mathbf{x}(\mathbf{s})$
$\mathbf{e}_\mathbf{v} \leftarrow \mathsf{min}[1.0, \lambda_\mathbf{v} \mathbf{e}_\mathbf{v} + \mathbf{x}(\mathbf{s})]$
$\mathbf{v} \leftarrow \mathbf{v} + \alpha_\mathbf{v} \delta \mathbf{e}_\mathbf{v}$
$\mathbf{e}_\mu \leftarrow \lambda_\mathbf{w} \mathbf{e}_\mu + (a - \mu)\mathbf{x}(\mathbf{s})$
$\mathbf{w}_\mu \leftarrow \mathbf{w}_\mu + \alpha_\mu \delta \mathbf{e}_\mu$
$\mathbf{e}_\sigma \leftarrow \lambda_\mathbf{w} \mathbf{e}_\sigma + [(a - \mu)^2 - \sigma^2]\mathbf{x}(\mathbf{s})$
$\mathbf{w}_\sigma \leftarrow \mathbf{w}_\sigma + \alpha_\sigma \delta \mathbf{e}_\sigma$

B. Real-time Prediction Learning

Another contribution of the present work is to explore the use of real-time temporally extended prediction learning to supplement ACRL and conventional joint control approaches for increased actuation capacity. In recent work, we described a generalized prediction learning approach based on reinforcement learning [19], and demonstrated how learned temporally extended predictions can accurately forecast signals during robot control by both amputees and able-bodied subjects [20], [21]. Predictions were phrased as a linear combination of a learned weight vector, here denoted $\mathbf{v}_p$, and a state approximation $\mathbf{x}_p(\mathbf{s})$ (n.b., this need not be the same approximation function used in ACRL learning). Predictions P for a given signal r_p were then computed using $P = \mathbf{v}_p^T \mathbf{x}_p(\mathbf{s}_p)$, where $\mathbf{v}_p$ was updated on each time step using the following incremental procedure:

$\delta_p \leftarrow r_p + \gamma_p \mathbf{v}_p^T \mathbf{x}_p(\mathbf{s}'_p) - \mathbf{v}_p^T \mathbf{x}_p(\mathbf{s}_p)$
$\mathbf{e}_p \leftarrow \mathsf{min}[1.0, \lambda \mathbf{e}_p + \mathbf{x}_p(\mathbf{s}_p)]$
$\mathbf{v}_p \leftarrow \mathbf{v}_p + \alpha_p \delta_p \mathbf{e}_p$

As in the ACRL procedure above, temporal-difference learning was used to update the weight vector $\mathbf{v}_p$. Replacing eligibility traces ($\mathbf{e}_p$) were again used to accelerate learning [27], where the trace decay rate λ_p and step size α_p (determined empirically) governed the magnitude of the weight vector updates; γ_p was the time scale or degree of temporal extension for each prediction. As such, each learned prediction approximated the exponentially discounted expectation of a signal with a time scale of $1/(1 - \gamma_p)$ time steps [20]. As discussed in our prior work, predictions of this kind can be specified for arbitrary signal types, and with respect to different temporal extension levels and control policies [19]. Using the above ACRL and predictive methods as a basis, we now examine how a learned set of sensorimotor predictions can be integrated into the space of input signals for a myoelectric joint controller.

C. Experimental Comparisons

A principal goal of our study was to suggest and compare methods for actuating one or more supplementary joints in response to human actions and intent. We divide the span of possible methods into *direct control policies*, wherein the control actions are a direct pre-scripted mapping from input signals to joint velocity commands, and *learned control policies*, acquired over the course of time (in this case via ACRL). We further divide the span of methods into *reactive* and *predictive* methods. In the reactive case, only instantaneous measurements from the system are used in a control policy. In the predictive case, temporally abstracted expectations of one or more system signals are given as inputs to a controller. As evidenced by the breadth of approaches noted in the literature, the number of possible combinations of predictive and reactive information into a control policy is vast. As such, we have limited our preliminary exploration to the five specific cases described below. For clarity, we begin with simple cases and incrementally add complexity in an effort to evaluate the effect of predictive knowledge and understand its suitability for different scenarios.

1) Direct W-Reactive Control: In our experimental setting of interest, the actuator target θ_w^* is inferred from user control actions or defined by the user on a moment-by-moment basis. As the new target is not known until it is reached, one simple control policy is to observe θ_w^* at every point and output an action a that moves θ_w as quickly as possible toward the

target angle. This approach assumes that the target angle can be known and observed by the system at all times.

2) Direct W-Predictive Control: While it is reasonable to assume that θ_w^* can be known during an initial training or calibration phase, in many cases a measure of the target angle will be unavailable during deployed operation. One approach is therefore to learn a prediction of the target signal, denoted P_{w*}, during a training phase and then directly link this prediction to the system's action choices during use. P_{w*} was used in place of θ_w^* within the controller above, moving θ_w as quickly as possible towards the anticipated value of the target angle. The use of a temporally extended prediction as a control target was also expected to help reduce or remove control delays, as compared to the purely reactive case above.

3) Full-Reactive ACRL Control: In cases where the exact value of θ_w^* is unknown or unspecified, it may become necessary to learn the correct control policy from less specific feedback signals like a scalar reward signal or human-delivered good-bad feedback [26], [29]. A first natural approach to this case is an ACRL controller that is given as its input a full selection of salient signals from within the human-robot system: $\mathbf{s} = \langle \theta_e, \theta_h, \theta_w, \dot{\theta}_e, \dot{\theta}_h$,Ch1–Ch2, Ch3–Ch4$\rangle$; this control learner received negative rewards proportional to the difference between the current and target wrist angles: $r = -|\theta_w^* - \theta_w|$, in radians.

4) EH-Predictive ACRL Control: As a predictive extension of the previous case, we explored the use of predictive signals regarding the two user controlled joints as input to the ACRL learner, in addition to the robot's current wrist angle: $\mathbf{s} = \langle P_e, P_h, \theta_w \rangle$; the control learner received reward as described above. This configuration was also expected to show greater anticipatory movement than a learner based on immediate measurements from the user-controlled joints.

5) W-Predictive ACRL Control: Finally, should θ_w^* be available during training, there may be advantages to learning a control policy that takes into account anticipatory knowledge about the target angle θ_w^*, but does not directly follow P_{w*} as a target. To examine this idea, we created an ACRL control learner with an input space $\mathbf{s} = \langle P_{w*}, \theta_w \rangle$; a prediction of the target wrist angle was available as an input signal to the learning system. Reward was given as described above. Faster learning was expected for this approach via its reduced input state space and predictive compression.

D. Testing Procedures

Multiple trials were performed using each of the five control cases. Each trial began with a 1.7min pre-learning phase, wherein subjects were allowed to familiarize themselves with the control of the robot. To provide a baseline error value, the wrist actuator was controlled using the Direct W-Reactive control policy during pre-learning. 21min of machine learning followed pre-learning, during which the ACRL control learners and/or prediction learners freely updated their internal weights and policies (n.b., to provide stable starting values for ACRL cases with predictive state inputs, prediction learners began their updates during pre-learning). Throughout learning and pre-learning subjects remained in full control of the robot's elbow and hand actuators; however, they were cued by the system to switch their pose every 5–15s in order to maintain a regular distribution of poses throughout the trial and allow the fair comparison of achieved wrist control policies. All subjects gave informed consent in accordance with the study's authorization by the University of Alberta Health Research Ethics Board.

All ACRL systems used the same function approximation approach; the learners' input signal space $\mathbf{s}$ was mapped into a binary feature vector using tile coding as per Pilarski et al. [26]. The real-time prediction learners used to approximate P_e, P_h and P_{w*} all were presented with a signal space comprised of $\mathbf{s} = \langle \theta_e, \theta_h, \dot{\theta}_e, \dot{\theta}_h$,Ch1–Ch2, Ch3–Ch4$\rangle$. As detailed in previous work [20], signals were normalized to their maximum ranges prior to use in function approximation; tile coding approximation involved two levels of discretization on each signal axis (3 and 5 for ACRL learners and 4 and 9 for prediction learners, with 5 and 8 overlapping tilings respectively) and a single bias unit. ACRL learning parameters were set as follows and remained unchanged for different tests and subjects: $\alpha_{\mathbf{v}} = 0.1/m, \alpha_\mu = 0.005/m, \alpha_\sigma = 0.5\alpha_\mu$ ($\alpha_\sigma = 0.25\alpha_\mu$ for offline learning), $\gamma = 0.96, \lambda_{\mathbf{w}} = 0.3, \lambda_{\mathbf{v}} = 0.7$; m denotes the number of non-zero features in the binary feature vector $\mathbf{x}(\mathbf{s})$. Parameters used in real-time prediction were $\lambda_p = 0.999, \gamma_p = 0.97$, and $\alpha_p = 0.3/m$. All weight vectors were initialized to 0, and and σ was bounded by $\sigma \geq 0.01$. Learning updates occurred at 30Hz (33ms time steps); all signal acquisition, control, and computation was done on a single MacBook Pro 2.53 GHz Intel Core 2 Duo laptop.

III. RESULTS

The five control approaches presented above were compared in terms of the accuracy of their achieved wrist control policies over the learning phase of each trial, with accuracy measured in terms of the system's reward received at each time step (a reflection of joint control error). Approaches were also evaluated in terms of the qualitative behaviour of their final policies (for example, the potential for anticipatory or preemptive natural movements). In order to assess the convergence and expected asymptotic performance of the learning methods during longer trials, we also compared the performance of the different approaches during iterative offline learning passes through each of the recorded online datasets using a simulated wrist actuator.

A. Quantitative Assessments

Figure 3a presents a learning curve comparison for the predictive and reactive control approaches over ∼20min of learning. Traces show the reward received by each method, averaged over four independent trials performed by two subjects; plotted data were binned into 100 time step segments. As shown, all three ACRL approaches undergo a period of exploration before slowly converging toward policies that produce lower error (greater reward). For this number of runs, no statistically significant quantitative differences in received reward were observed between the ACRL methods

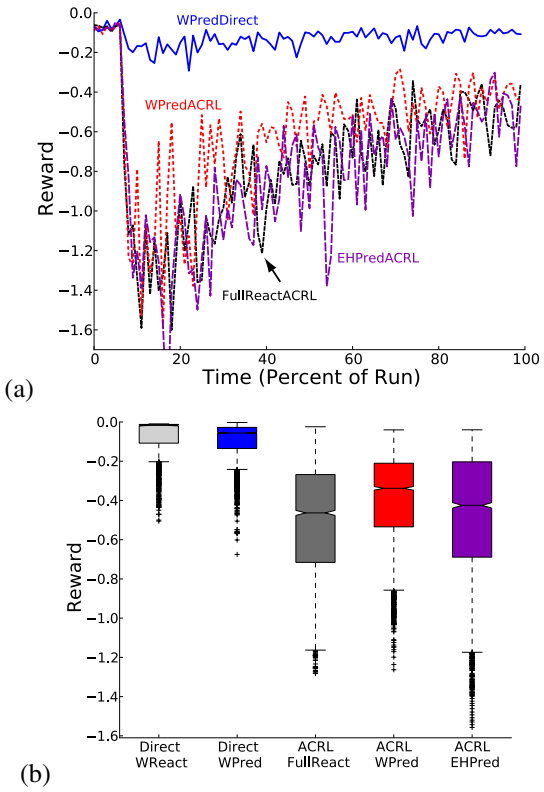

Fig. 3. Comparison of predictive and reactive control learning approaches (n=4) over the course of ∼20min of online learning, following a 1.7min pre-learning phase: (a) binned per-time-step reward over time, and (b) quartile analysis of median values shown over the last 1.7min of learning, as compared to 1.7min of the direct reactive policy during pre-learning.

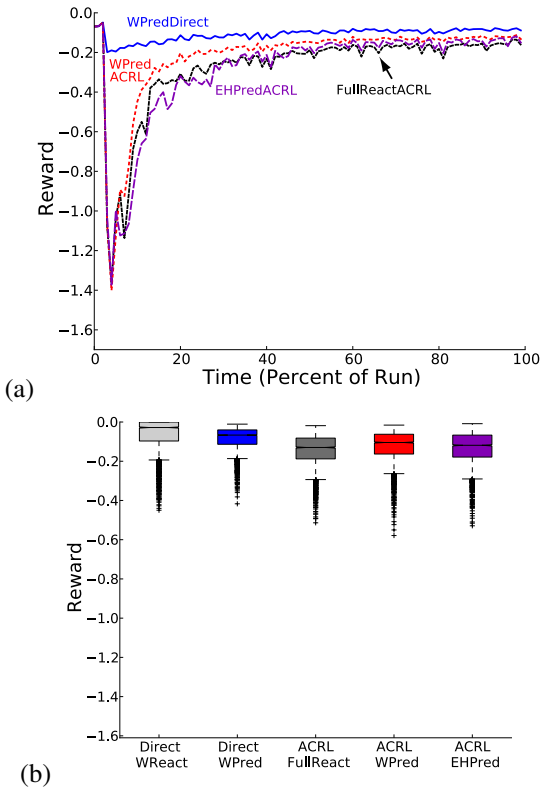

Fig. 4. Comparison of predictive and reactive control learning approaches (n=16) over the course of ∼50min of offline learning (2.5 passes through 21min of logged online learning data, following 1.7min of pre-learning): (a) binned per-time-step reward over time, and (b) quartile analysis of median values shown over the last 1.7min of learning.

after 20min. As seen in Fig. 3a, top trace, Direct W-Predictive control received more reward in all cases than ACRL, learned a viable policy in ∼2.5min, and showed continued performance increases over the course of learning as the accuracy of its learned wrist target predictions increased.

Visible differences between control approaches were observed through an analysis of reward values over the last 3k (1.7min) timesteps of learning. Figure 3b displays a quartile analysis of the received reward over this period. As shown by the median values (notch centres) and their 95% confidence bounds (width of notches), Direct-W Predictive achieved a better final control policy than all three ACRL methods, approaching the reward received by the Direct W-Reactive Controller. Position of the first and third quartile lines (blue box tops and bottoms, respectively) convey observed relationships between the different ACRL approaches and their end-of-run performance.

As suggested by the performance observed during offline iterative learning (Fig. 4), continued online learning is expected to improve the tracking performance of the ACRL learners. Traces in Fig. 4a show the reward received by each method, averaged over the 16 online data files recorded from the two subjects; plotted data were binned into 100 time step segments. For this number of runs, salient differences were observed between ACRL learners during the first 20min of offline learning; W-Predictive ACRL demonstrated a statistically significant increase in learning speed over the other two ACRL approaches. Direct W-Predictive control maintained performance superiority over the ACRL methods during offline learning.

B. Quality of the Learned Policies

Figure 5 presents examples of wrist control trajectories achieved at the end of the learning phase by the Direct W-Predictive controller and the three ACRL approaches after both online learning and offline iterative learning. After 17min of learning both predictive and reactive ACRL learners achieved the desired wrist trajectory, though visible oscillation around the target profile were still visible. Following 2.5 iterations of offline learning the profiles had converged to tightly track the target trajectory. Stochasticity during early learning was partly due to the rate of convergence of σ in the actors' policies—at 17min, the σ values computed from $\mathbf{w}_\sigma$ at each step were still observed to be ≥ 0.3, leading to continued exploratory actions around the learned policy mean μ. As shown in Fig. 5ae, Direct W-Predictive was able to closely track the wrist target profile, giving much better qualitative performance than ACRL learners trained for the same amount of time. However, we did observe small but noticeable spikes in joint angle near some transition points; these spikes corresponded to residual errors in the prediction learner's forecasts (as P_{w*} was directly mapped to θ_w).

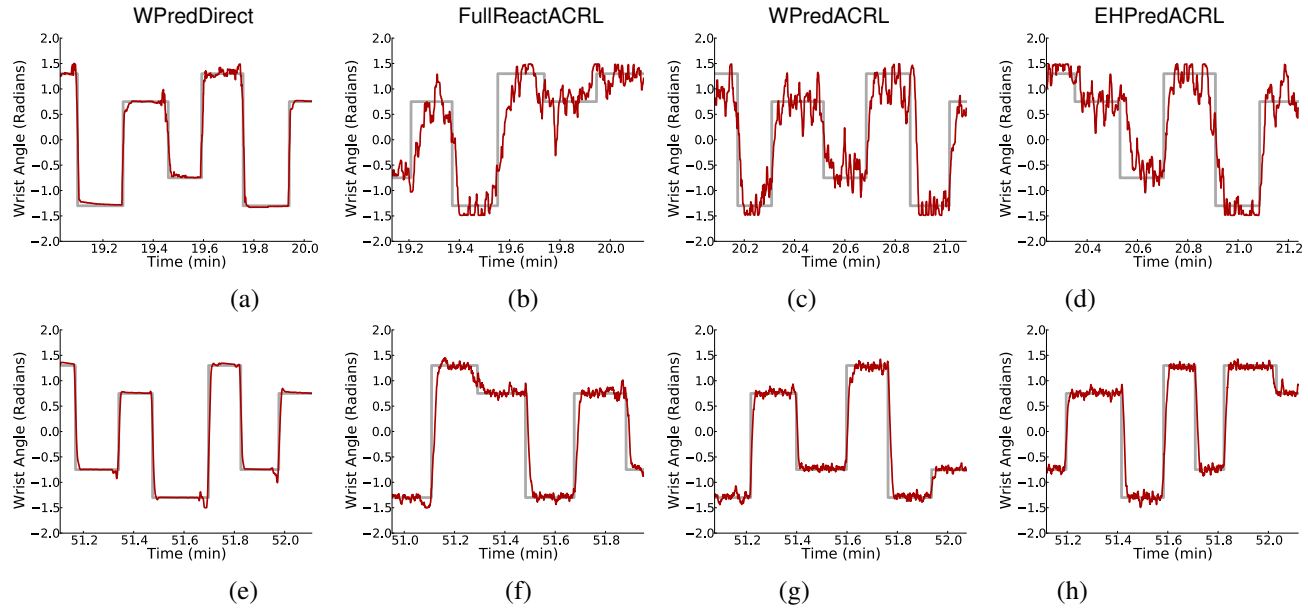

Fig. 5. Comparison of target (grey line) and achieved (red line) wrist trajectories after (a–d) ∼20min of online learning and (e–h) ∼50min of offline learning. Shown for (a/e) Direct W-Predictive control, (b/f) Full-Reactive ACRL, (c/g) W-Predictive ACRL, and (d/h) EH-Predictive ACRL.

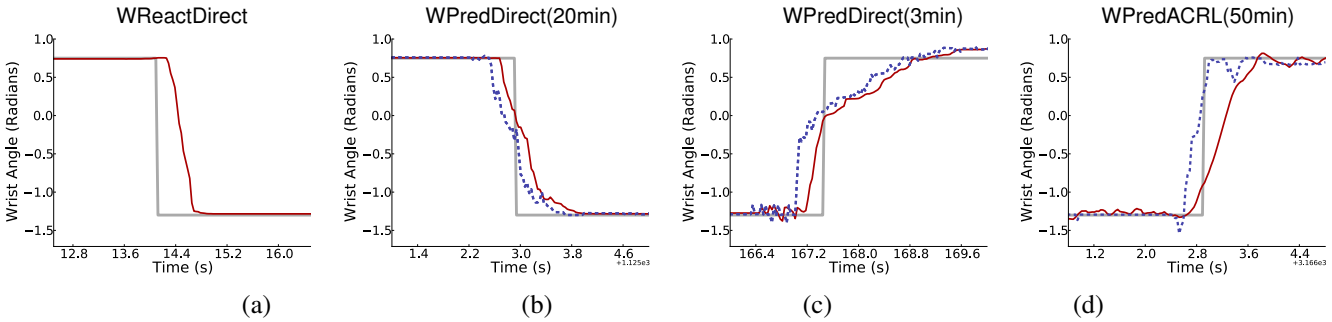

Fig. 6. Examples of target (grey), achieved (red), and predicted (dashed blue) wrist trajectories over a 4s period for (a) Direct W-Reactive control, (b–c) Direct W-Predictive control at the end and start of learning, and (d) W-Predictive ACRL control after 50min of offline learning.

C. Capacity for Preemptive Movement

The degree of motor anticipation displayed by each method provides another important area for comparison. Figure 6 shows representative results indicating the ability of each method to provide preemptive actuator motion. As shown in Fig. 6a, the Direct W-Reactive policy suffers from a minor but consistent delay inherent to the robot control setting. Conversely, the anticipatory information used in the Direct W-Predictive controller, Fig. 6bc, provided proactive motion that occurred in advance of the change in the target wrist trajectory; the predictive controller consistently started wrist motion before the target policy changed, arriving at the desired target shortly after the human-controlled elbow and hand joints ceased moving. In Fig. 6bc, predictions of the target trajectory (dotted blue trace) precede both the effected robot actuator position (red trace), and the target trajectory (grey trace). Predictive ACRL approaches also showed potential for preemptive motion, as depicted in Fig. 6d. More work is needed to quantitatively compare the amount of anticipation achievable by predictive and reactive ACRL methods following asymptotic convergence.

IV. DISCUSSION

Can real-time prediction learning improve the quality of learned or engineered prosthesis control policies? Our results suggest that direct predictive control learning may provide a way to quickly develop control policies for supplementary prosthetic joints. Direct predictive control was able to learn a reasonable wrist-actuation policy in only 2.5min, and consistently improved this policy to the point that it compared favourably to a direct reactive control policy with full persistent knowledge of the target trajectory. However, a direct predictive approach is only viable in cases where a joint target function is explicitly available to a system during calibration or training. For cases where only a surrogate or simplified representation of the target trajectory is available (in the form of either a reward/success signal or a compressed target signal), both reactive and predictive ACRL controllers were found to be able to learn a viable target policy from a scalar reward signal and low-complexity combinations of different instantaneous and anticipatory input signals. A learned approach like ACRL also appears suitable for cases where the specifics of the human-machine system are unknown, or where it is challenging to distill the system's dynamics

into a concrete mapping between target configurations and individual actuator commands. Our three ACRL approaches learned in a timeframe amenable to human use, suggesting they may be suitable for use in patient-based trials. We also expect that with additional training and/or experience re-use, ACRL learning could result in policies that are superior to those of direct predictive controllers. While a direct mapping from prediction to motor commands will transmit any errors made by the real-time prediction learner (Fig. 5ae), ACRL with an objective function should be able to learn a policy that takes into account spurious predictions (Fig. 6d); this expectation needs to be extensively tested in future work.

Can real-time prediction learning allow a system to adapt during online operation without relearning its controller? Based on our results, we expect the use of two independent learning systems may allow a controller to adapt to real-time changes without a time consuming re-calibration process. Prediction learning progressed much faster than control learning in our study, as in studies of human motor learning [11]. Our ACRL controllers learn a policy that is contingent on predictions; when changes occur in the user or their device, these predictions can be automatically updated and maintained by the independent prediction learner without the need to change the policy learned by the ACRL system. This approach also promises to benefit existing prosthetic pattern recognition approaches, where temporally extended predictions can serve as adaptive inputs to a standard classifier. Detailed analysis of learning rates and the capacity for real-time adaptation of predictive and reactive ACRL control learning approaches is an important area for further study.

Can real-time prediction learning facilitate the control of multiple prosthetic actuators and effect synergies? While wrist targets in our study were specified as a simple function of two user-controlled joints, it is easy to imagine settings where there is no clear way to map user-achieved poses to high-dimensional multi-actuator configurations, or where target configurations are subjective and need to be defined by the user through methods such as human-delivered reward [26], [29], [30] and kinesthetic teaching [31]. Our presented methods are amenable to reward functions provided in terms of binary success/failure feedback, and cases where human users provide scalar reward indicating their approval/disapproval so as to train abstract, user-defined control policies (e.g., using the techniques of Knox and Stone [29], as supplemented for ACRL by Vien and Ertel [30]). It is also reasonable for a device's user to desire target poses that take into account movement history, poses that vary or change over time (i.e., have a progression or desired non-static trajectory), and poses that depend on additional state information such as grip force or explicit user cues (e.g., vocal commands). These extensions seem reasonable with more varied input signal spaces and function approximation strategies, and are open areas for future work.

Can the use of temporal extension in real-time prediction learning help a prosthetic controller plan anticipatory movements for coordinated multi-joint limb motion? We showed that the use of real-time predictions allowed both direct and learned controllers to initiate anticipatory motions that help remove control and mechanical delays, moving joints preemptively in advance of changes to the target joint configuration. Interestingly, the reactive ACRL controller was also found to exhibit motor anticipation on some transitions. Our observations suggest that this behaviour was enabled by information contained in the movement of the user-controlled joints; as some joints move more slowly than others (in this case, our robot's hand actuator), the controller learned to pick up on sight motor cues in the environment that signalled an impending transition. As the level of task richness increases (e.g., toward the complexity of daily functional tasks) we expect that both reactive and predictive controllers should be able to increase their ability to forecast complex motor trajectories. Additional state information, for example abstracted historical features of motor activity, may enable improved long-term forecasting during control. As suggested in our previous work [20], signal-based predictive planning of this kind is expected to help enable natural or synergistic multi-joint motion in high-dimensional actuator spaces when an amputee's control channels are precious and limited; the capacity for computationally efficient ongoing prediction learning also facilitates multi-joint control polices that respond to and compensate for ongoing variations that occur in a patient's signals and their use of an assistive device's multiple functions [21].

Can real-time prediction learning allow the compression of a large input state space into content-rich signals to speed controller learning? The use of prediction is known to potentially abstract a large and possibly more complex state space into one or more clear and information-rich signals that can be used by a control learning system [22]–[24]. As one example from the present study, the W-Predictive ACRL learner compressed the complete user-control state space into a scalar approximation of the target function that increased the learning speed of the ACRL system (as shown in Fig. 4). Our preliminary comparison of W-Predictive ACRL with the other ACRL methods indicates that there may be learning advantages to adding additional input signals into a prediction learner's state space instead of adding them to that of a predictively coupled ACRL or other control learner. Whether similar levels of compression can be achieved by the careful manual design of a function approximation system for a single learner remains to be tested.

Do the benefits of adding a second learning subsystem promise to outweigh the increased complexity and optimization challenges it entails? As discussed above, our results indicate a number of potential advantages to using separate prediction learning and control learning subsystems. These advantages include the capacity for unsupervised real-time adaptation, anticipatory movements, and improved control learning speeds. The use of a separate prediction learning system also allows temporally extended predictions to be made simultaneously at different timescales, providing additional controller state information that cannot be captured effectively in the standard implementation of an ACRL control learner or most conventional control approaches.

V. CONCLUSIONS

The comparisons presented in this article serve as a preliminary survey of different cases where it may be pragmatic to include a set of temporally extended predictions—learned and maintained in real time—in the input space of a direct or learned prosthesis controller. Building on other approaches where a predictive model is either hand engineered or learned in an offline setting, we here provide a way to acquire and maintain a temporally abstract predictive model during the ongoing operation of a human-robot system, and then use this predictive model to support real-time multi-joint control.

Our study contributes a first successful deployment of ACRL in concert with real-time prediction learning. We demonstrated this new control-learning approach during the live myoelectric control of a multi-joint robot limb, and validated our online observations with extended offline experiments. Our initial results suggest that coupling prediction learners and control learners may be a viable way to speed control policy acquisition, allow unsupervised adaptation in deployed myoelectric controllers, and drive synergistic actuators that extend the control dimensions available to a human user. Finally, our study shows that both learned and direct predictive controllers have the capacity to perform anticipatory and coordinated actuator movements. These results provide a starting place for research into more natural control interfaces for multi-actuator prostheses.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from the Alberta Innovates Centre for Machine Learning (AICML), the Natural Sciences and Engineering Research Council (NSERC), and Alberta Innovates – Technology Futures (AITF). We also thank Joseph Modayil for a number of helpful discussions regarding the use of predictions in control.

REFERENCES

[1] A. d'Avella, A. Portone, L. Fernandez, and F. Lacquaniti, "Control of fast-reaching movements by muscle synergy combinations," *J. Neurosci.*, vol. 25, no. 30, pp. 7791–7810, 2006.

[2] D. M. Wolpert, Z. Ghahramani, and J. R. Flanagan, "Perspectives and problems in motor learning," *Trends Cogn. Sci.*, vol. 5, no. 11, pp. 487–494, 2001.

[3] T. W. Williams, "Guest Editorial: Progress on stabilizing and controlling powered upper-limb prostheses," *J. Rehab. Res. Dev.*, vol. 48, no. 6, pp. ix–xix, 2011.

[4] S. Micera, J. Carpaneto, and S. Raspopovic, "Control of hand prostheses using peripheral information," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 48–68, 2010.

[5] B. Peerdeman, D. Boere, H. Witteveen, et al., "Myoelectric forearm prostheses: State of the art from a user-centered perspective," *J. Rehab. Res. Dev.*, vol. 48, no. 6, pp. 719–738, 2011.

[6] E. Scheme and K. B. Englehart,"Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *J. Rehab. Res. Dev.*, vol. 48, no. 6, pp. 643–660, 2011

[7] L. Resnik, M. R. Meucci, et al., "Advanced upper limb prosthetic devices: implications for upper limb prosthetic rehabilitation," *Arch. Phys. Med. Rehabil.*, vol. 93, no. 4, pp. 710–717, 2012.

[8] M. S. Johannes, J. D. Bigelow, J. M. Burck, et al., "An overview of the developmental process for the modular prosthetic limb," *Johns Hopkins APL Tech. Dig.*, vol. 30, no. 3, pp. 207–216, 2011.

[9] G. A. Dumanian, J. H. Ko, K. D. O'Shaughnessy, et al., "Targeted reinnervation for transhumeral amputees: current surgical technique and update on results," *Plast. Reconstr. Surg.*, 124(3), pp. 863–9, 2009.

[10] J. L. Collinger, B. Wodlinger, J. E. Downey, et al., "High-performance neuroprosthetic control by an individual with tetraplegia," *The Lancet*, 2012, article in press. DOI: 10.1016/S0140-6736(12)61816-9.

[11] J. R. Flanagan, P. Vetter, R. S. Johansson, and D. M. Wolpert, "Prediction precedes control in motor learning," *Current Biology*, vol. 13, no. 2, pp. 146–150, 2003.

[12] J. Zacks, C. Kurby, M. Eisenberg, and N. Haroutunian, "Prediction error associated with the perceptual segmentation of naturalistic events," *J. Cogn. Neurosci.*, vol. 23, no. 12, pp. 4057–4066, 2011.

[13] C. Pulliam, J. Lambrecht, and R. F. Kirsch, "Electromyogram-based neural network control of transhumeral prostheses," *J. Rehabil. Res. Dev.*, vol. 48, no. 6, pp. 739–754, 2011.

[14] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robot. Autom. Mag.*, vol. 4, no. 1, pp. 23–33, 1997.

[15] S. M. LaValle. *Planning Algorithms.* Cambridge Univ. Press. 2006.

[16] P. Abbeel, A. Coates, and A. Y. Ng, "Autonomous helicopter aerobatics through apprenticeship learning," *Int. J. Robotics Research*, vol. 29, no. 13, pp. 1608–1639, 2010.

[17] J. Modayil, A. White, P. M. Pilarski, and R. S. Sutton, "Acquiring a broad range of empirical knowledge in real time by temporal-difference learning," in *Proc. 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2012)*, Seoul, Korea, 2012, pp. 1903–1910.

[18] J. Modayil, A. White, and R. S. Sutton, "Multi-timescale nexting in a reinforcement learning robot," in *Proc. Int. Conf. Simulation of Adaptive Behaviour (SAB)*, Odense, Denmark, 2012, pp. 299–309.

[19] R. S. Sutton, J. Modayil, M. Delp, et al., "Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction," in *Proc. 10th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)*, Taipei, Taiwan, 2011, pp. 761–768.

[20] P. M. Pilarski, M. R. Dawson, T. Degris, et al., "Adaptive artificial limbs: a real-time approach to prediction and anticipation" *IEEE Robot. Autom. Mag.*, in press, 2013.

[21] P. M. Pilarski, M. R. Dawson, T. Degris, et al., "Dynamic switching and real-time machine learning for improved human Control of assistive biomedical robots," in *Proc. 4th IEEE RAS & EMBS Int. Conf. Biomedical Robotics and Biomechatronics (BioRob)*, Roma, Italy, 2012, pp. 296–302.

[22] M. L. Littman, R. S. Sutton, and S. Singh, "Predictive representations of state," in *Advances in Neural Information Processing Systems 14*, pp. 1555–1561, MIT Press, 2002.

[23] E. J. Rafols, M. B. Ring, R. S. Sutton, and B. Tanner, "Using predictive representations to improve generalization in reinforcement learning," *Proc. 2005 Int. Joint Conference on Artificial Intelligence (IJCAI)*, 2005, pp. 835–840.

[24] B. Boots and G. J. Gordon, "An online spectral learning algorithm for partially observable nonlinear dynamical systems", in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2011)*, San Francisco, California, USA, 2011.

[25] P. K. Artemiadis and K. J. Kyriakopoulos, "EMG-based control of a robot arm using low-dimensional embeddings," *IEEE Trans. Rob.*, vol. 26, no. 2, pp. 393–398, 2010.

[26] P. M. Pilarski, M. R. Dawson, T. Degris, et al., "Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning," in *Proc. IEEE Int. Conf. on Rehab. Robotics*, Zurich, Switzerland, 2011, pp. 134–140.

[27] T. Degris, P.M. Pilarski, and R.S. Sutton, "Model-free reinforcement learning with continuous action in practice," *Proc. of the 2012 American Control Conf. (ACC)*, 2012, Montreal, Canada, 2012, pp. 2177–2182.

[28] V. Mathiowetz, G. Volland, N. Kashman, and K. Weber, "Adult norms for the box and block test of manual dexterity," *The American Journal of Occupational Therapy*, vol. 39, no. 6, pp. 386–391, 1985.

[29] W. B. Knox and P. Stone, "Learning non-myopically from human-generated reward," in *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, March 2013.

[30] N. A. Vien and W. Ertel, "Learning via human feedback in continuous state and action spaces," in *2012 AAAI Fall Symposium Series, Robots Learning Interactively from Human Teachers (RLIHT)*, Arlington, USA, 2012, pp. 65–72.

[31] K. Muelling, J. Kober, O. Kroemer, and J. Peters, "Learning to select and generalize striking movements in robot table tennis," in *2012 AAAI Fall Symposium on Robots Learning Interactively from Human Teachers (RLIHT)*, Arlington, USA, 2012, pp. 38–45.