Control of a Center-Out Reaching Task using a Reinforcement Learning Brain-Machine Interface

Justin C. Sanchez, *Member IEEE,* Aditya Tarigoppula, John S. Choi, *Member IEEE*, Brandi T. Marsh, Pratik Y. Chhatbar, Babak Mahmoudi, *Member IEEE*, and Joseph T. Francis

***Abstract*— In this work, we develop an experimental primate test bed for a center-out reaching task to test the performance of reinforcement learning based decoders for Brain-Machine Interfaces. Neural recordings obtained from the primary motor cortex were used to adapt a decoder using only sequences of neuronal activation and reinforced interaction with the environment. From a naïve state, the system was able to achieve 100% of the targets without any a priori knowledge of the correct neural-to-motor mapping. Results show that the coupling of motor and reward information in an adaptive BMI decoder has the potential to create more realistic and functional models necessary for future BMI control.**

I. Introduction

BRAIN-machine interfaces (BMIs) offer tremendous promise as assistive systems for motor-impaired patients. At the core of a motor BMI, is an embedded computer that decodes in real-time the user's sensorimotor commands derived directly from the nervous system. Thus, two key problems of BMI research are (1) to establish the proper experimental interactive paradigm between user and decoder and (2) to model how the brain plans [1] and controls motion.

For patients with spinal cord injury, reaching and grasping is a desirable function to be regained after injury [2]. Moreover, the use of this function with high performance in complex environments during the activities of daily life will provide the greatest impact in terms of independence [2]. Many groups have conducted experiments that couple users and decoders in center-out reaching tasks to approximate those reaching behaviors desirable to BMI users. In these experiments, decoding has been performed with engineering "black-box" models that make simplifying assumptions about the nature of the neural-to-motor mapping and the complexities of the environments that they work in [3-11]. While these models have shown BMI proof-of-concept, they lack many components of realism of the true sensorimotor system. As a result, new studies are showing that they may not be general purpose in the activities of daily life because they lack high-level goals coupled with musculoskeletal dynamics, are not robust to the loss of neurons, and have difficulty dealing with environmental dynamics.

As a first step in rethinking the BMI decoding paradigm, we have developed a new framework for BMIs. We will use reinforcement learning (RL) as a guiding principle to link in vivo physiology with computational modeling during dexterous tasks in dynamical environments [12]. Unlike other supervised learning principles as is commonly used in BMI, RL enables our networks to learn and grow through experience as the natural network would [13]. This computational and biological framework offers a method of neural interfacing that uses goal-directed, experience-based learning to relate neural modulation to behavior through accumulation of rewards and interaction with the environment [13]. Collectively in this framework, sensorimotor subsystems contribute to forming a Perception-Action-Reward Cycle (PARC), which plays a critical role in organizing behavior in the nervous system [14]. In essence, by adding these components of real biological systems into the BMI we are shifting the decoding from "black-box" to "white-box" models. In this work, we develop an experimental primate test bed for a center-out reaching task to test the performance of reinforcement learning based decoders.

II. Methods

A. Center-Out Reaching Task and surgical implantation

A female bonnet macaque (*M. radiata*) was trained to perform a center-out reaching task (distance between the center of the start point and targets was 4 cm, target radius 0.8-1 cm) with her right arm attached to exoskeletal robotic manipulandum (KINARM, BKin Technologies, Kingston, ON, Canada.) Once the animal reached task proficiency level of ~80% success, she was implanted in left primary somatosensory cortex (S1, areas 3b and 1), motor cortex (M1) and dorsal premotor cortex (PMd) representing the right shoulder and elbow regions with multiple 'Utah' microelectrode arrays (10x10 electrode grid, tips covered with Platinum with 450 um inter-electrode distance at tip, 1.5 mm shank length, Blackrock Microsystems, Salt Lake City, UT.) We adapted to a new surgical implantation

Manuscript received January 14, 2011. This work was supported by DARPA project N66001-10-C-2008 awarded to J. Francis and J. Sanchez. Authors A. Tarigoppula and J. Sanchez had equal contribution in this work.

J. C..Sanchez, is with the Department of Biomedical Engineering and the Miami Project to Cure Paralysis, University of Miami, Coral Gables, FL 33146 USA (phone: 305-284-2330 e-mail: jcsanchez@miami.edu).

A. Tarigoppula, J. S, Choi, B. S. Marsh, and P. Y. Chhatbar are with the Department of Physiology and Pharmacology, SUNY Downstate Medical Center, Brooklyn, NY 11203, USA (email: [aditya30887, jschoi831, btm2002, pratikchhatbar]@gmail.com).

B. Mahmoudi is with the Department of Biomedical Engineering, University of Miami, Coral Gables, FL 33146 USA (e-mail: b.mahmoudi@gmail.com).

J. T. Francis is with the Department of Physiology and Pharmacology, Robert Furchgott Center for Neural and Behavioral Sciences, Program in Biomedical Engineering, SUNY Downstate Medical Center, Brooklyn, NY 11203, USA (email: joey199us@gmail.com).

technique utilizing a Nesting Platform to keep the infection rate and implantation costs low [15]. All the animal handling and surgical procedures were in compliance with Institutional Animal Care and Use Committee (IACUC) and were closely observed and assisted by the Division of Laboratory Animal Resources (DLAR) at SUNY Downstate Medical Center. For this paper, we will limit our discussions to neural recordings from M1.

B. Neurophysiological Recordings

After implantation surgery, the monkey was allowed to recover for 2-3 weeks after which, recordings of single unit activity were taken while the animal performed the center out reaching task. Recordings were made using externally synched multiple multichannel acquisition processor systems (MAPs). Neural signals were amplified, thresholded and single units were sorted based on their waveforms using principal-component-based methods in the Sort-Client software (Plexon Inc., Dallas, TX.) From the 96 channels we recorded from, we were able to sort anywhere between 190-240 units. We used 185 units for our simulations here.

C. Decoding Using Reinforcement Learning

The decoder used in this reaching task is based on the theory of reinforcement learning (RL) and is shown in Fig. 1 [12]. By assigning a reward to the agent, experience can be used to find the functional mapping between neural states, computer cursor actions, and rewards. This architecture was developed in [16, 17] and we present a brief synopsis of the same. Here, the adaptation is focused on maximizing rewards through successful completion of the trials by the agent. To illustrate this, we present a brief overview of exactly how the system learns in an open loop case (i.e. the data were first recorded while the monkey performed the manual task and then used to train the BMI decoder). The agent is trained with RL, which is a learning algorithm for decision making in goal-based tasks. RL is different from other paradigms because it learns through interaction with environment rather than a specific training signal [12]. We modeled the agent's cursor control problem as a Markov Decision Process (MDP), which is characterized by neural modulation as states *s* and discrete movements performed by the RL agent (BMI decoder) as actions *a* in Eqs. 1-2. Each action in a particular state will change the state of the environment with a certain probability. This probability in (1) is the transition probability. Additionally, the agent expects a reward *r* when taking an action given a state. This expected reward is expressed in (2). These transition probability functions are unknown; therefore we used RL to learn the approximate values of (1) and (2) based on observations. Once the agent has a good estimate of (2), it can choose the actions, which will maximize reward. Specifically, we used Q(λ) learning [12] to approximate (2). We implemented this learning with a multilayer perceptron (MLP) neural network to map state-action pairs to their expected value (derived from (2)) [17].

$$P^{a}_{ss'} = \Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\} \quad (1)$$

$$R^{a}_{ss'} = E\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\} \quad (2)$$

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)] \quad (3)$$

The network is trained trough (3). In this equation, Q is the state-action value function therefore the agent's control ability is a function of the MLP (Multi Layer Perceptron) training. The number of hidden units was 10. The network was updated using the method of back propagation [18]. The discounting factor γ was 0.9. The exploration rate, which defines how many times a random action is taken by the agent, was 0.01 (i.e. 1 random action per 100 actions). The parameter λ was assigned a value of 0.8. For this center-out reaching task, the reward distribution in the 2-D workspace is defined as 0.6 at the targets and -0.6 anywhere else [1].

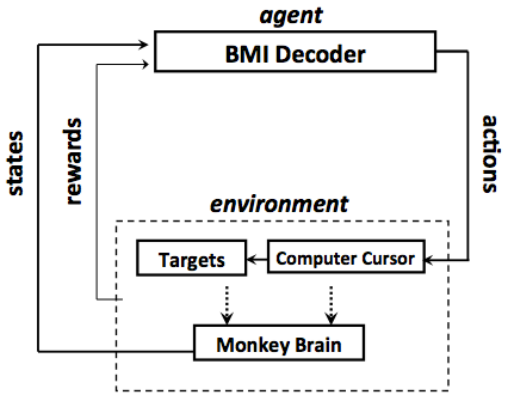

Figure 1: Decoding architecture based on reinforcement learning.

III. Results

For the neural data used in this study, the monkey performed manual reaches at a success rate of 90%. To train the RL network, only successful trials were used. The Reinforcement Learning (RL) agent was required to perform a two target center-out reaching task as shown in Fig. 2. The agent had eight possible actions (directions) to choose from, shown by arrows to reach the presented target. The targets presented to the RL agent were either collinear or non-collinear as shown in Fig. 2(a) and Fig. 2(b) respectively. From the moment the target was presented, monkey reached the target within a duration of about 700-800ms. The neural data corresponding to 185 units was given as inputs to the MLP, wherein, 700 ms of neural modulation history was embedded into each state (input).

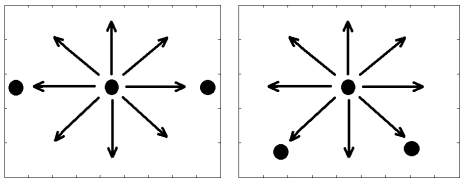

Figure 2: Center-out reaching task. The center circle indicates that start position. The black circles in the periphery represent the possible targets for the agent, whereas the arrows represent the possible actions that can be taken by the RL agent.

The state-action value (Q value) for each action is obtained as the output of the MLP. The RL agent is required to learn the optimal policy to complete the task. In this case, the agent needs to distinguish between the two correct actions that lead to a positive reward and six actions that provide a negative reward.

One of the hallmarks of RL decoding is that performace evolves and improves over time from the ramdomized initialization state. As shown in Fig. 3, the agent performing a task with collinear targets (Fig. 2(a)), began with an average success rate of 6.98% for the task completion during the first epoch. Each epoch consisted of 43 trials. By the end of the 15th epoch (trial 602-645) the success rate increased to 97.67%, whereas the agent performed the 18th epoch (trial 731-774) with a success rate of 100% and maintained its success rate for 17 more epochs as shown in Fig. 3. This figure is quite telling regarding the adaptation of the network over time because on epochs 1-14 we can see that one of the targets was learned but it took longer time to learn the second target through reinforcement.

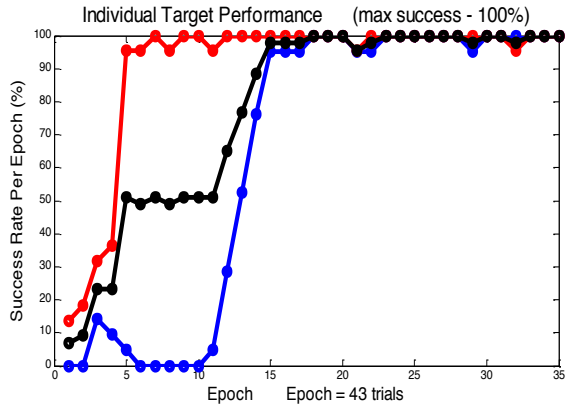

Figure 3: Inidividal target (right and left) performance along with the total success rate for a simulation performed on a 2 target plane arranged as shown in Fig 2(a). Simulation lasted for 35 epochs (i.e. above 1500 trials). Red indicates right target, blue indicates left target and black indicates total success rate of the agent.

In Fig. 4 we present a detailed analysis of the adaptation over trials to demonstrate how the variables of the system change over time. In subplot (a), we present the Q-values (State-Action value functions) for 8 possible actions. A clear differentiation of the right and left action Q-value from the Q-values of other possible actions is observed post 600th trial. In subplot (b), the targets selected by the agent are shown in 'red' whereas the actual targets given to the agent are shown in 'black'. Initially many wrong targets are selected, following which a bias towards an action is developed (trial 250-400), but over time through reinforcement the correct targets are selected and the bias is eliminated. In subplot (c), the MLP output weights indicate an increase in magnitude of only the parameters necessary for solving the task. The other weights remain as their small initialization values. Through adaptation, the agent learns to perform the trials correctly thus achieving more positive instantaneous rewards indicating a good solution. In subplot (e), the Q-values (State-Action value functions) for 8 possible actions are presented. Note the transitions between the Q-values of the left and right action which suggest the ability of the agent to select the correct action with respect to a given target. The agent was capable of acquiring 100% success for the organization of targets as illustrated in Fig. 2(a) and Fig. 2(b).

To compare against standard classification techniques used in the literature, a multinomial logistic regression model was used to similarly classify the 8 reach directions. The L2 regularization constant was found during the first fold of 4-fold cross validation, and the mean performance was taken using the neural data corresponding to the 2-target case. Since the RL agent chooses an end target in a single step, this linear multinomial classifier accomplishes the same goal, but with a supervised method. For this reason, the performance of the two algorithms can be compared as shown in Table I. Here, equivalent performance with a static classifier model can be achieved however in the case of the RL, a priori no training signal is required.

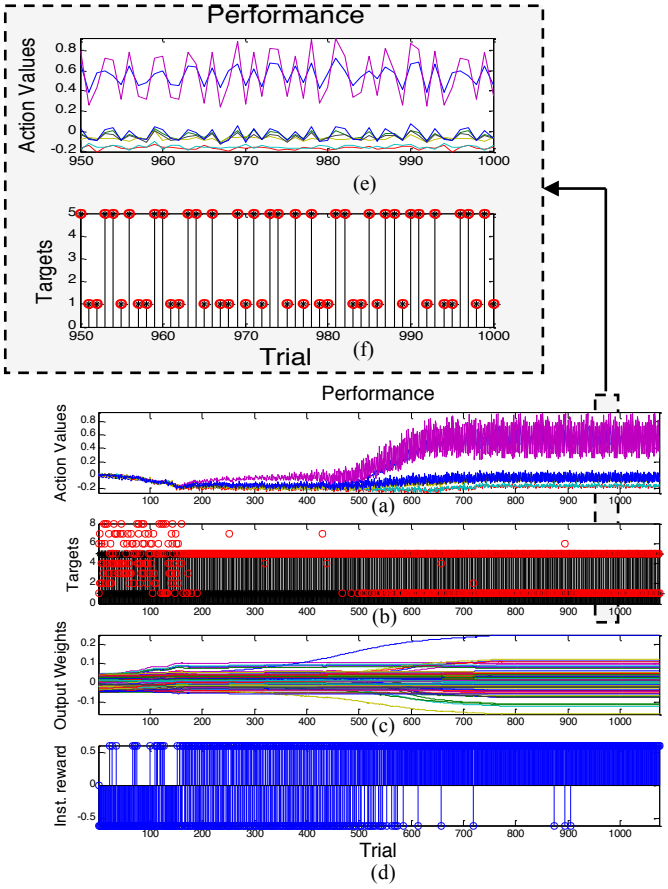

Figure 4: Adaptation of the Q-values, MLP weights, and instantaneous reward over trials. In (b) and (f), The targets selected by the agent are shown in 'red' whereas the actual targets given to the agent are shown in 'black'.

TABLE I Performance Comparison

Parameters	Random Actions (chance)	Non Random Actions (collinear and non-collinear targets)	Linear Regression
Hidden Layer learning rate	0.0009	0.0009	-
Output Layer learning rate	0.003	0.003	-
Success (%)	12.79	100	100

IV. Conclusions

In this work, a reinforcement learning based decoder for a center-out BMI task was developed. We have shown for a two-target task that the system could adapt from a naïve state to acquire the correct direction of movement with 100% success. Performance was on par with a standard BMI classifier however no a priori information is needed to train the system. Learning is achieved through interaction with the environment and the acquisition of rewards that are achieved through the production of brain states.

While the theory of RL based BMI decoding was first developed with rodent behavioral experiments, the work presented here is an extension to a center-out task with primate neural recordings. As it is the first step, the trajectories formed were quite simple and only consisted of one step to achieve the target. This foundation will be expanded in the future to include multi-step trajectories as were derived in the initial rodent experiments.

Lastly, the experiments conducted here were performed offline to facilitate systematic study of the system convergence, parameters, and performance. This initial prototyping is a necessary step towards closed-loop decoding with such a paradigm. Now that the foundations have been developed, our goal is to carryout closed-loop experiments to study how engaging the motor and sensory cortices with reinforcement learning affect the performance.

References

[1] K. V. Shenoy, D. Meeker, S. Cao, S. A. Kureshi, B. Pesaran, C. A. Buneo, A. P. Batista, P. P. Mitra, J. W. Burdick, and R. A. Andersen, "Neural prosthetic control signals from plan activity," *NeuroReport,* vol. 14, pp. 591-597, 2003.

[2] K. D. Anderson, "Targeting recovery: Priorities of the spinal cord-injured population," *Journal of Neurotrauma,* vol. 21, pp. 1371-1383, Oct 2004.

[3] S. P. Kim, J. C. Sanchez, Y. N. Rao, D. Erdogmus, J. C. Principe, J. M. Carmena, M. A. Lebedev, and M. A. L. Nicolelis, "A Comparison of Optimal MIMO Linear and Nonlinear Models for Brain-Machine Interfaces," *J. Neural Engineering,* vol. 3, pp. 145-161, 2006.

[4] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple Neural Spike Train Data Analysis: State-of-the-art and Future Challenges," *Nature Neuroscience,* vol. 7, pp. 456-461, 2004.

[5] M. D. Serruya, N. G. Hatsopoulos, L. Paninski, M. R. Fellows, and J. P. Donoghue, "Brain-machine interface: Instant neural control of a movement signal," *Nature,* vol. 416, pp. 141-142, 2002.

[6] S. I. Helms Tillery, D. M. Taylor, and A. B. Schwartz, "Training in cortical control of neuroprosthetic devices improves signal extraction from small neuronal ensembles," *Reviews in the Neurosciences,* vol. 14, pp. 107-119, 2003.

[7] D. W. Moran and A. B. Schwartz, "Motor cortical representation of speed and direction during reaching," *Journal of Neurophysiology,* vol. 82, pp. 2676-2692, 1999/11 1999.

[8] D. M. Taylor, S. I. H. Tillery, and A. B. Schwartz, "Direct cortical control of 3D neuroprosthetic devices," *Science,* vol. 296, pp. 1829-1832, 2002.

[9] W. Wu, M. J. Black, Y. Gao, E. Bienenstock, M. Serruya, and J. P. Donoghue, "Inferring hand motion from multi-cell recordings in motor cortex using a Kalman filter," in *SAB Workshop on Motor Control in Humans and Robots: on the Interplay of Real Brains and Artificial Devices*, University of Edinburgh, Scotland, 2002, pp. 66-73.

[10] Y. Gao, M. J. Black, E. Bienenstock, W. Wu, and J. P. Donoghue, "A quantitative comparison of linear and non-linear models of motor cortical activity for the encoding and decoding of arm motions," in *The 1st International IEEE EMBS Conference on Neural Engineering*, Capri, Italy, 2003.

[11] J. C. Sanchez, S. P. Kim, D. Erdogmus, Y. N. Rao, J. C. Principe, J. Wessberg, and M. A. L. Nicolelis, "Input-output mapping performance of linear and nonlinear models for estimating hand trajectories from cortical neuronal firing patterns," in *International Work on Neural Networks for Signal Processing*, Martigny, Switzerland, 2002, pp. 139-148.

[12] R. S. Sutton, Andrew G. Barto, *Reinforcement learning: an introduction*. Cambridge: The MIT Press, 1998.

[13] J. C. Sanchez, B. Mahmoudi, J. DiGiovanna, and J. C. Principe, "Exploiting co-adaptation for the design of symbiotic neuroprosthetic assistants," *Neural Networks special issue on Goal-Directed Neural Systems,* vol. 22, pp. 305-315, 2009.

[14] J. M. Fuster, "Upper processing stages of the perception-action cycle," *Trends in Cognitive Sciences,* vol. 8, pp. 143-145, 2004.

[15] P. Y. Chhatbar, L. M. von Kraus, M. Semework, and J. T. Francis, "A bio-friendly and economical technique for chronic implantation of multiple microelectrode arrays," *J Neurosci Methods,* vol. 188, pp. 187-94, May 15 2010.

[16] S. I. H. Tillery, D. M. Taylor, and A. B. Schwartz, "Training in cortical control of neuroprosthetic devices improves signal extraction from small neuronal ensembles," *Reviews in the Neurosciences,* vol. 14, pp. 107-119, 2003.

[17] J. DiGiovanna, B. Mahmoudi, J. Fortes, J. C. Principe, and J. C. Sanchez, "Co-adaptive Brain-Machine Interface via Reinforcement Learning," *IEEE Trans. Biomed. Eng.,* 2008.

[18] R. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning,* vol. 3, pp. 9-44, 1988.