

# Co-Adaptive Learning in Brain-Machine Interfaces

Babak Mahmoudi<sup>1</sup>, Jack DiGiovanna<sup>2</sup>, Jose C. Principe<sup>3</sup>, and Justin C. Sanchez<sup>4</sup>

<sup>1,2</sup>*Department of Biomedical Engineering, University of Florida, Gainesville, FL 32608 USA*  
babakm@ufl.edu, jfd134@ufl.edu

<sup>3</sup>*Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA*  
principe@cnel.ufl.edu

<sup>4</sup>*Department of Pediatrics, Division of Neurology, University of Florida, Gainesville, FL 32610 USA*  
jcs77@ufl.edu

**Abstract-** This paper studies the cooperation between an artificial agent and a biological organism to accomplish a goal directed task in context of Reinforcement Learning Brain-machine Interface (RLBMI). The artificial agent's learning is based on Temporal Difference (TD) learning to adapt model parameters whereas the biological agent's learning is represented in the temporally specific modulation of brain neural activity. We observed that these two completely different learning mechanisms coexist and contribute to the overall learning of the system in achieving reward which is characterized by a behavioral learning curve criterion. This learning paradigm may introduce a new framework to specify the engineering principles for designing next generation adaptive systems.

**Index Terms**— Brain-Machine Interface, Neuroprosthetic, Reinforcement Learning, Co-adaptation.

## I. INTRODUCTION

Learning through reinforcement is a critical adaptive mechanism that allows animals to shape their behaviors to maximize rewards from the environment [1]. The remarkable aspects of the neural process that support such behavior is that they are capable of responding almost instantaneously in a variety of changing environments. Discovery of the underlying principles of how animals use neural representation and timing has the great potential to specify the engineering architecture of the next generation of adaptive systems. For artificial systems, there have been many developments in the machine learning paradigm known as reinforcement learning (RL) [2-5] in this direction. In the RL framework, the learner (which is called the agent) continually interacts with its environment and after each interaction the agent receives a reward from its environment. The agent tries to maximize earned rewards over time. While RL has been an influential computational theory in neuroscience there are still several aspects that need to be investigated - namely its speed of adaptation, realism of the modelling of brain learning, and its ability to explain a variety of the neural systems involved [6].

One approach for determining in more detail the design principles for cooperative adaptation is to study the direct

interaction between brain and machine as is done in neuroprosthetics. The beauty of Brain-Machine Interface (BMI) technology is that adaptive models serve as surrogate communication channels for neural systems. This technology provides a framework to study theories of interactive learning both from engineering and neuroscience perspectives. Much work has already been done in BMIs however from primarily a static input-output modeling framework [7-9] and the concept of co-adaptation [10, 11] has yet to be fully realized.

In the pursuit of co-adaptive neural interfaces, we have proposed a BMI paradigm which is based on a modified interpretation of Reinforcement Learning (RLBMI) [12, 13]. In this paradigm an artificial, intelligent agent through interaction with the user's brain learns how to map neural activity to taking actions which are desirable for the user. Through this process, the artificial agent receives rewards or punishments based upon satisfaction of the user. The user also tries to adapt their neural modulation to achieve a common goal based on the user's understanding of the agent's behavior. This co-adaptation opens a new field in interactive learning where synergy among adaptive components can facilitate the learning. The idea of multi agent learning is the intersection between machine learning and multi-agent systems where all of the learners are artificial. The novelty in our approach is the interaction among artificial and biological agents. This paradigm can provide a platform to study the machine learning and biological learning as well as the mutual learning that happens in their interaction. In this paper, we present the co-adaptive learning in the context of an experimental BMI that requires coordination between artificial and biological intelligence to solve a motor task for reaching and grasping. We quantify here the relative speed of adaptation of both the computational model and neural modulation to better specify the engineering of next generation co-adaptive BMIs.

## II. METHODS

### A. Computational framework

Our new closed-loop framework for studying causation between biological networks and computational models is shown in Fig. 1. This framework tests theories of goal-based learning and decision making through experience using a robotic arm in 3-D workspace during a reaching task. Here, the interaction between a computational agent (BMI Algorithm) and user’s brain (Rat’s Brain) occurs through the generation of a sequence brain states that are mapped by the agent to a series of actions of a robotic arm. This action sequence must move the robotic arm from a central location to one of the two targets (see Fig. 2) in the robot’s workspace to achieve a goal. Upon completing the task, the animal will receive a water reward. The agent and user must learn co-adaptively (based on actions and observed states and rewards) which strategies will maximize the earned reward.

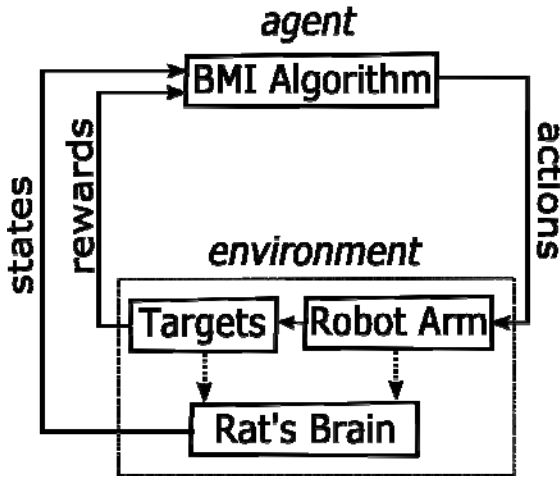


Fig. 1 RLBMI architecture

In our paradigm, the rat’s brain and agent participate in a dialogue to maximize their cumulative rewards. To achieve this, the artificial agent is trained through reinforcement learning. It is assumed that the environment is stochastic and modeled as Markov Decision Process (MDP) which is characterized by a set of states and actions. Each action in a particular state will change the state of the environment with a certain probability. This probability is known as the transition probability.

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

where  $s$  and  $a$  represent state and action respectively. The Markov property of MDP implies that transition from the current state to the next state is independent of all previous transitions and states which may be a strong assumption for BMI applications. The agent expects a reward (negative reward can be respected as punishment) by visiting a new state. This expected reward can be expressed in the form of Eq. (2) [2].

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

In addition to the state and action sequences, the RLBMI is composed of three main elements; reward function, value function and policy. The reward function determines the immediate agent reward based on endpoint position of the robot in the 3-D workspace. The reward function provides a scalar reward value to the sequence of state-action pairs that have caused the robot to reach its current location. The value function  $Q$  determines the value of state or state-action pair in terms of expected return. The return is defined in (3) as a discounted sum of all rewards that would be earned in future, where for  $\gamma \leq 1$ , smaller values of  $\gamma$  give more weight to immediate rewards. The policy determines which action the agent will select given  $Q$ .

$$R_t = \sum_{n=t+1}^{\infty} \gamma^{n-t+1} r_n \quad (3)$$

Temporal Difference (TD) methods can estimate the value function through interaction with environment without need for a model of environment. In RLBMI the transition probability of states is unknown therefore TD learning is a good choice for value function estimation. Another feature of TD methods is that they can be implemented incrementally; therefore, it is suitable for real time applications. The one-step update equation for the value function estimation (VFE) is defined by:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (4)$$

where the  $Q$  function approximates optimal action-value function independent of the policy [2]. Eq. (4) uses TD error to update  $Q$ ; TD error is given in (5)

$$\delta_t = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}^*) - Q(s_t, a_t) \quad (5)$$

We have used  $Q(\lambda)$ -learning, an advanced version of  $Q$ -learning, which is an off policy TD control algorithm [2]. A Multi Layer Perceptron (MLP) neural network is trained with  $TD(\lambda)$  error backpropagation to estimate the  $Q$  function (state-action value function) [16]. In our paradigm, the artificial agent uses this neural network to assign a value for each action given the neural states. The 3-layer network is defined by the multidimensional neural input and a 3-tap gamma memory structure [14] for an effective tap delay of 600 ms. The hidden layer consists of three hyperbolic tangent processing elements whose number was determined using pruning [15,13]. The output consists of twenty-seven linear Processing Elements (PE) each corresponding to one of the actions (directions) that the robot can move at each time step (27 PE: 26 actions + 1 NO action).

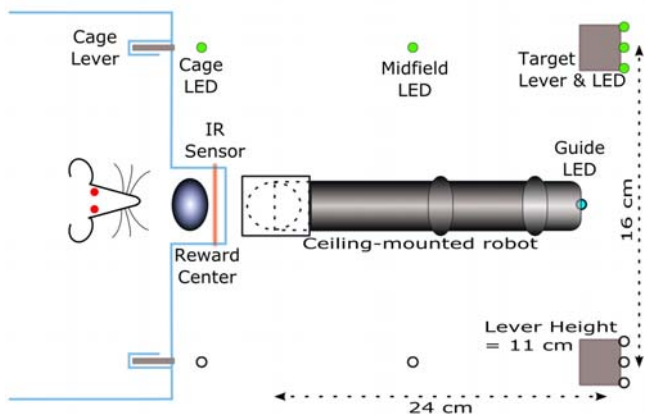


Fig.2 RLBMI experiment setup

**B. Experimental paradigm**

In this paradigm, we have trained three Sprague-Dawley rats in a two-target choice task. To obtain a water reward, the rats were required to manually press one of the two cage levers concurrently with a five degree of freedom robotic arm pressing one of two target levers (left and right positions), cued by an LED light (Fig. 2). After training, the rats were bilaterally implanted with two electrode arrays in the forelimb region of primary motor cortex (16 electrodes in each M1) [17]. Electrophysiological recordings are performed with commercial neural recording hardware (TDT, Alachua FL). A TDT system (one RX5 and two RP2 modules) operates synchronously at 24414.06 Hz to record neuronal action potentials from both microelectrode arrays. The neuronal potentials are band-pass filtered (0.5 - 6 kHz) and spike sorted using template matching [18, 19]. During brain-control of the robotic arm, single unit activity of cortical cells was chronically recorded and their firing rates (100ms bins) created the environmental state for artificial agent in RLBMI. In this architecture, the agent controls the robot and at each time step it maneuvers the robot by evaluating the environmental state and action pairs and selecting one of 27 action directions (e.g. left, right, forward-left-up) in the 3-D space. To test the neural response to the co-adaptation, the shaping of complex behaviors was enabled with an adjustable threshold which sets the necessary target proximity to earn reward for both the agent and animal. This threshold was iteratively adjusted from close to the robot starting position to far away where the range of probabilities of randomly intersecting the target was 25% - 4.4%. In other words, by increasing the threshold, animal had to maneuver the robotic arm for a longer distance to reach the target hence increase in difficulty level. Each animal was assigned a single VFE model and it was co-adapted over multiple sessions (1-2 hours) spanning multiple days.

**III. RESULTS**

Using this architecture, we found that the three animals were able to maintain an average task performance (earn rewards) of at least 450% over chance despite the increasing

task difficulty [ref TBME]. Here, we study two adaptive elements of the architecture: the agent model and the user’s neuromodulation. First, in the agent’s VFE network we observed the weight tracks to be smoothly varying (no discontinuities) over multiple days indicating that past experience was being maintained by the agent for each difficulty level. For the user’s neuromodulation, 60% of the neurons had a decrease, 30% had an increase and 10% had no significant change in the mean firing rate as a function of the difficulty level when compared to the first session of co-adaptation. Interestingly, of the neurons with a decrease in the mean and variance of their firing rate, 73% had an increase in their Coefficient of Variation (CV) of firing, which is the ratio of variance to mean, and the rest had no significant change in their CV. 86% of the neurons which had increase in their mean firing had no significant change in their CV. The remaining 10% of the neurons that had no significant change in their mean firing rate also did not show a significant change in their CV. All metrics were tested for significance using ANOVA at 95%. Based on the results, the animal’s contribution to co-adaptation does not primarily occur as a general up-regulation in neuronal firing but as an increase in temporally specific neuromodulation of the ensemble related to sub goals of the complete reaching task.

**A. Within Session Model Adaptation**

In order to demonstrate this trend, we now show the model and neural co-adaptation for one particular session (animal 3 - 4.4% probability of randomly reaching the target). To measure the co-adaptation, we define here three metrics in the system: the first measures the overall learning rate of the system. The second reflects the change in the behavior of the artificial agent and the third defines the change in the behavior of the biologic agent.

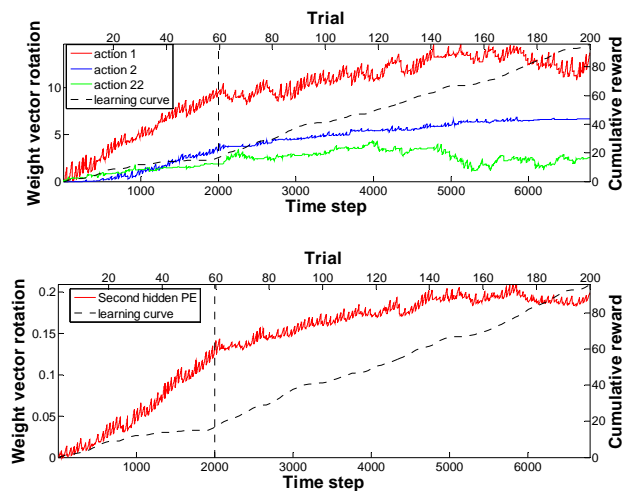


Fig. 3 Relationship between the overall learning rate and model adaptation. At time 2000, the model adaptation plateaus and there is a change in slope of the learning rate. A) Weight adaptation of three most significant actions. B) Input weight adaptation connected to the second hidden PE.

The criteria for behavioral learning (criteria 1) was defined as the rate of reward returns over time [20]. It is assumed that as the behavioral task is mastered the rate of rewards will increase [22]. The time between two successful trials where the animal has earned reward is used as a measure of overall learning in the system. Fig. 3 (A, B) shows the learning curve of the system. In this figure, at each time a reward is earned the cumulative reward is incremented by one. Once the task is learned, the reward should be earned more frequently therefore the slope of the curve would increase. From the learning curve of the system we can see that at time step 2000 there is a change in the slope of the learning curve that indicates a change in the performance of the system.

In our experimental paradigm, in each session the network was initialized with the weights from the previous session; therefore, learning through a session manifests itself in the changes in the network parameters from the initial condition. As a measure of the artificial agent's learning through a real-time RLBMI experiment session we have computed the directional cosine between the output layer weights, input layer weights and their initial values at each intra trial time step. Fig. 3 shows the instantaneous rotation of the output layer weight vectors that correspond to three most significant actions and the second hidden PE weight vector with respect to its initial condition. (For the other hidden PEs weight vectors did not significantly rotate from their initial condition.) Compared to the learning curve of the system we can see that at the time the weight rotation has reached a knee point (time step 2000) the inter reward time interval has decreased significantly (knee point of the learning curve). In other words, the adaptation in the artificial agent's parameter correlates with the overall learning of the system.

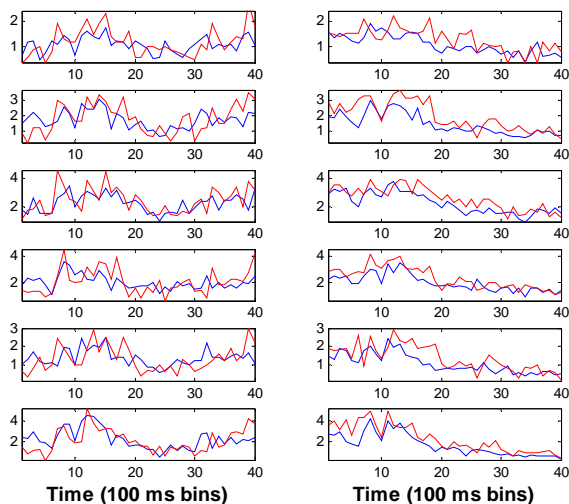


Fig. 4 Left and right trial peri-event time histograms for six neurons. Column 1 corresponds to averages over the time before the knee in the learning curve. Column 2 corresponds to averages after the knee. The y-axes are spike counts. Blue averages are for left trials, red are for right trials. For each neuron the scale on y-axes for left and right plots is the same.

### B. Within Session Neural Adaptation

In the RLBMI we also measure the changes in neural representation associated with co-adapting with the agent to solve the task. To assess the learning of the biologic agent from its neural activity we used the neurophysiologic approach of computing the peri-event time histogram [21] of the neuronal firing rates over the trials before and after the knee point of learning curve (in criteria 1). The goal is to determine if there is any statistically significant change in the representation that occurs before and after the knee of the behavioral learning curve.

Fig. 4 shows the peri-event histogram of firing rate of 6 neurons before and after the knee point of the learning curve. Within each subplot the neuromodulation in 100 ms bins is separated into left and right successful trials. The average trial time before and after learning is 3.6 seconds. Time  $t=0$  is the start of trial. From these plots we can see that there are changes in the overall neural representation during animal adaptation. After the knee of the learning curve, there is an increase in the neural modulation in bins 1-5 which can be used to support an initial selection of movement direction. Up to bin 20, the increased neuromodulation also becomes more temporally specific with a substantial reduction in the firing during the later times of the trial.

## IV. CONCLUSIONS

In this paper, we have investigated the co-adaptation between two learning systems; an artificial agent and a rat's brain, in the context of co-adaptive BMI. We have found that in this paradigm both the artificial agent and the brain of biological organism adapt their modulation and parameters through mutually beneficial interaction to earn reward. Learning in the artificial agent is represented in the adaptation of a set of basis functions for the animal to optimally project its neural activity into action space. The change in rotation of the weight vectors of the artificial agent before and after the knee point of system learning curve, implies that the artificial agent has changed the projection such that particular actions were selected that result in more rewards. However earning rewards is not the sole responsibility of the agent. Through behavioral interaction with artificial agent's actuator (robotic arm), the animal also tries to earn reward by changing its neural modulation. By analyzing the firing rate of ensembles of neurons and comparing the peri-event histogram of spikes before and after the knee point of learning curve of the system, we can see that learning in the system coincides with a significant change in the temporal pattern of animal's neural modulation. These results suggest that animal has learned how to change its neural modulation in a specific way to earn reward. In other words, the overall learning in the system is a result of co-adaptation of both the artificial agent and rat's brain. While the rat is learning which neural modulations result in water rewards, the BMI agent must adapt to more effectively respond to the rat's brain states. Reinforcement Learning provides a computational framework for quantifying

this interaction. This concept might be used in developing a new generation of neural interfaces through integration of biological and artificial intelligence.

### Acknowledgement

This work was supported in part by the U.S. National Science Foundation under Grant #CNS-0540304, the Children's Miracle Network, and the UF Alumni Association Fellowship

### REFERENCES

- [1] G. H. Bower, *Theories of Learning*, 5th ed. Englewood Cliffs: Prentice-Hall, Inc., 1981.
- [2] R. S. Sutton, Andrew G. Barto, *Reinforcement learning: an introduction*. Cambridge: The MIT Press, 1998.
- [3] K. Doya, Samejima, K, Katagiri, K, and Kawato, M., "Multiple model-based reinforcement learning," *Neural Computation*, vol. 14, pp. 1347-1369, 2002.
- [4] F. Rivest, Y. Bengio, and K. J., "Brain Inspired reinforcement learning," in *NIPS*, Vancouver, CA, 2004.
- [5] N. Jong and P. Stone, "Kernel Based models for reinforcement learning," in *Workshop on Kernel machines for Reinforcement Learning, Proc. ICML Pittsburgh, PA, 2006*.
- [6] M. Kawato and K. Samejima, "Efficient reinforcement learning: computational theories, neuroscience and robotics," *Current Opinion in Neurobiology*, vol. 17, pp. 205-212, 2007.
- [7] J. C. Sanchez, J. C. Principe, T. Nishida, R. Bashirullah, J. G. Harris, and J. Fortes, "Technology and Signal Processing for Brain-Machine Interfaces: The need for beyond the state-of-the-art tools," *IEEE Signal Processing Magazine*, vol. 25, pp. 29-40, 2008.
- [8] J. C. Sanchez and J. C. Principe, *Brain Machine Interface Engineering*. San Rafael: Morgan and Claypool, 2007.
- [9] 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.
- [10] D. M. Taylor, S. I. Helms Tillery, and A. B. Schwartz, "Information conveyed through brain-control: Cursor versus robot," *IEEE Trans. Neural Systems and Rehabilitation Engineering*, vol. 11, pp. 195-199, 2003.
- [11] 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.
- [12] J. DiGiovanna, B. Mahmoudi, J. Fortes, J. C. Principe, and J. C. Sanchez, "Co-adaptive Brain Machine Interface via Reinforcement Learning," *IEEE Transactions on Biomedical Engineering (Special issue on Hybrid Bionics)*, vol. submitted, 2007.
- [13] J. DiGiovanna, B. Mahmoudi, J. Mitzelfelt, J. C. Sanchez, and J. C. Principe, "Brain-Machine Interface Control via Reinforcement Learning," in *3rd International IEEE EMBS Conference on Neural Engineering*, Kohala Coast, Hawaii, 2007.
- [14] B. De Vries and J. C. Principe, "The gamma model: a new neural network model for temporal processing," *Neural Networks*, vol. 5, pp. 565-576, 1993.
- [15] G. Orr and K.-R. Müller, *Neural Networks: Tricks of the Trade* vol. 1524. Berlin; New York: Springer, 1998.
- [16] R. S. Sutton, "Implementation details of the TD( $\lambda$ ) procedure for the case of vector predictions and backpropagation," 1989.
- [17] J. P. Donoghue and S. P. Wise, "The motor cortex of the rat: cytoarchitecture and microstimulation mapping," *J. Comp. Neurol.*, vol. 212, pp. 76-88, 1982.
- [18] M. A. L. Nicolelis, D. Dimitrov, J. M. Carmena, R. Crist, G. Lehew, J. D. Kralik, and S. P. Wise, "Chronic, multi-site, multi-electrode recordings in macaque monkeys," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 11041-11046, 2003.
- [19] M. A. L. Nicolelis, *Methods for Neural Ensemble Recordings*. Boca Raton: CRC Press, 1999.
- [20] Z. M. Williams and E. N. Eskandar, "Selective enhancement of associative learning by microstimulation of the anterior caudate," *Nature Neuroscience*, vol. 9, pp. 562-568, 2006.
- [21] E. E. Fetz, "Are movement parameters recognizably coded in the activity of single neurons," *Behavioral and Brain Sciences*, vol. 15, pp. 679-690, 1992 1992.
- [22] Stephen B. Klein, *Learning: Principles and Applications*, McGraw-Hill, 2001