2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871194

Estimating Reward Function from Medial Prefrontal Cortex Cortical Activity using Inverse Reinforcement Learning

Jieyuan Tan, Xiang Shen, Xiang Zhang, Zhiwei Song and Yiwen Wang, *Senior Member, IEEE*

***Abstract*—Reinforcement learning (RL)-based brain-machine interfaces (BMIs) learn the mapping from neural signals to subjects' intention using a reward signal. External rewards (water or food) or internal rewards extracted from neural activity are leveraged to update the parameters of decoders in the existing RL-based BMI framework. However, for complex tasks, the design of external reward could be difficult, which may not fully reflect the subject's own evaluation internally. It is important to obtain an internal reward model from neural activity to access subject's internal evaluation when the subject is performing the task through trial and error. In this paper, we propose to use an inverse reinforcement learning (IRL) method to estimate the internal reward function interpreted from the brain to assist the update of the decoders. Specifically, the inverse Q-learning (IQL) algorithm is applied to extract internal reward information from real data collected from medial prefrontal cortex (mPFC) when a rat was learning a two-lever-press discrimination task. Such an internal reward information is validated by checking whether it can guide the training of the RL decoder to complete movement task. Compared with the RL decoder trained with the external reward, our approach achieves a similar decoding performance. This preliminary result validates the effectiveness of using IRL to obtain the internal reward model. It reveals the potential of estimating internal reward model to improve the design of autonomous learning BMIs.**

***Index Terms*—brain-machine interface, internal reward, inverse reinforcement learning.**

I. Introduction

Brain-machine interfaces (BMIs) enable people with motor disabilities to better interact with the world through the communication between the brain and external devices [1]. Motor BMIs generally translate brain signals into movement intention using a decoder. Compared with decoders using supervised methods, reinforcement learning (RL)-based decoders do not require real limb movements for training, which is a significant advantage for paralyzed people [2]. In the RL-based BMI paradigm, the learning of the mapping from neural signals to actions is guided by a reward signal. To date, different designs of reward signals have been leveraged in RL-based decoding frameworks. External food or water reward are commonly used in BMI [3]–[5]. DiGiovanna *et al.* [3] presented an RL-based BMI using Q(λ)-learning algorithm using a water reward. In their experiment, the rats were trained to brain control a prosthetic arm to reach two targets. Wang *et al.* [6] proposed an incremental reward in the attention-gated reinforcement learning (AGREL)-based BMI framework when the primates were performing a center-out task. This reward is assigned based on the relative distance of the decoded cursor position to the target. Both DiGiovana and Wang *et al.* used external rewards to train the RL decoder. To develop autonomous learning BMIs, internal reward representation is proposed to replace the external reward in several studies. Shen *et al.* [7] proposed an internally rewarded RL-based decoder which extracts reward information from medial prefrontal cortex (mPFC) activity to guide the choice of movement. In their framework, the neural activity of the mPFC post the action duration is interpreted as the internal reward information. Mahmoudi *et al.* [8] replaced the external reward with the internal reward estimated from nucleus accumbens (NAcc) using a Multi-Layer Perceptron to develop an RL-based decoder. Both Shen and Mahmoudi *et al.* used the supervised methods to build an internal reward model that represents the mapping from neural activity to reward information, which still need the label of the external reward as the supervised signal for training. Since the external reward signals are artificially designed, it is not clear whether the learned reward model truly reflects of subjects' own evaluation. For some complex tasks, the design of external reward could be difficult and a more efficient approach to extract reward information from neural activity is needed.

Inverse reinforcement learning (IRL) method is a potential solution to address this reward design problem. The IRL methods aim to learn the reward function from observed behaviors through trial-and-error [9], allowing us to build a connection between neural activity and the recorded behaviors. Gabriel *et al.* [10] studied the behavior of rats in a response-preparation task and used the IRL method to map the neural spiking to instantaneous internal reward for RL decoding. But in their experiment, all recorded neurons were from motor areas, which may not be responsible for evaluation function. There have been many studies suggesting that the mPFC is involved in reward-guided learning [9]-[10]. Both human and animal studies have shown that mPFC neural signals are related to reward expectancy during the task learning [13]–[15]. The functionality of the mPFC makes it a good choice to be an internal critic to guide the learning of RL-based decoders.

We are interested in applying the IRL methods to estimate the internal reward function from mPFC cortical activity. A male Sprague Dawley (SD) rat was trained to learn a two-lever-press discrimination task according to audio cues. Neural

*This work was supported in part by the RGC of Hong Kong under GRF projects (GRF16213420), the National Natural Science Foundation of China (No.61836003), the special research support from Chao Hoi Shuen Foundation (R9051), the seed fund of the Big Data for Bio-Intelligence Laboratory(Z0428), and the HKUST-GZU Joint Research Collaboration Fund (GZU22EG01).

Jieyuan Tan, Xiang Shen, Xiang Zhang, Zhiwei Song and Yiwen Wang are with the department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology.

Yiwen Wang is also with the Department of Chemical and Biological Engineering, the Hong Kong University of Science and Technology. Yiwen Wang serves as the corresponding author (e-mail: eewangyw@ust.hk).

signals from primary motor cortex (M1) and mPFC were collected during the task learning. We apply Kalman filter to get the estimation of subject's kinematics. The estimated trajectories are then utilized as expert trajectories for inverse Q-learning (IQL) algorithm to learn the mapping from mPFC activity to instantaneous reward. To validate the effectiveness of reward function generated from IQL algorithm, we used the estimated internal reward to guide the training of an RL-based decoder and compare the decoding performance with the ones using external reward.

The rest of this paper is organized as follows: Section II shows the detail of experiment design, data collection, estimation of internal reward function and RL decoder validation. Section III shows the results of reward function estimation and decoding performance. Section IV gives the conclusions and future work.

II. METHOD

A. Experiment design and data collection

All animal handling procedures were approved by the Animal Care Committee of the Hong Kong University of Science and Technology, strictly complying with the Guide for Care and Use of Laboratory Animals. Six male Sprague Dawley (SD) rats were trained to perform a two-lever-press discrimination task. At the beginning of each trial, the rats would hear a high-pitch (10 kHz) or low-pitch (1.5 kHz) audio cue, which was randomly generated and lasted 900 ms. To get the water reward, the rats needed to press the corresponding lever (for example, high lever when hearing the high-pitch cue) within 5 s and hold it for 500 ms. If the rats accomplished the task successfully, a feedback cue (the same pitch as start cue) lasting 90 ms would be given to indicate the success and the rats would be rewarded with a water drop. If the rats pressed the wrong lever, released early or did not respond within 5 s, the trial would be considered unsuccessful, with neither the feedback cue nor water reward. The inter-trial interval was a random value ranging from 3 to 6 s.

To record neural signals, each rat was surgically implanted with two 16-channel microelectrode arrays in the areas of M1 and mPFC. Data acquisition and storage were accomplished by the neural recording system (Plexon Inc, Dalla, Texas). The raw signal was sampled at 40 kHz and digitally high-passed through a 500 Hz four-pole Butterworth filter. The spikes were detected with a -4 σ threshold (σ is the standard deviation of the histogram of the amplitudes). The single neuron of each channel was sorted on offline sorter (Plexon Inc, Dallas, Texas). Spike firing rates were counted with a nonoverlapping 100 ms time window. All the behavior events and their timings were recorded by the behavior recording system (Lafayette Instrument, USA) and synchronized with the neural recording system. In this paper, we used the data collected from one rat when the rat was well-trained with the task. On the selected day, the rat achieved a success ratio of 85.4% (205 successful trials among 240 trials).

B. Estimation of internal reward function using IQL

In this paper, we apply Inverse Q-learning (IQL) [16] algorithm to learn the mapping from mPFC activity to internal reward through trial and error. We validate the proposed internal reward model by checking whether the estimated reward can guide the training of RL-based decoders. The system is designed as shown in Fig. 1.

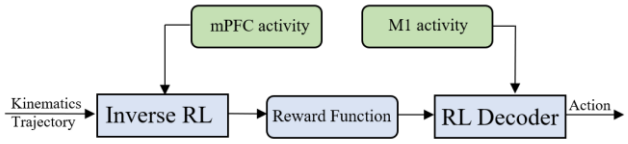

Fig 1. The design of internal reward model and its validation by decoding

The internal reward model takes the input from the 16-channel mPFC neural activity s, and current action a, and generates the instantaneous estimation on internal reward r. Since the neural firing state can be regarded as a continuous space, we use a neural network to build the reward function using approximation[16]

$$r = R(s, a, |\theta^r), \tag{1}$$

where θ^r denotes the parameter of the neural network, and the r denotes the instantaneous reward of taking action a in state s. The action a is defined as approaching or moving away from the target.

The IQL algorithm assumes that action should be chosen following a stochastic policy with an underlying Boltzmann distribution over optimal Q-values:

$$\pi(a|s) := \frac{\exp\big(Q^*(s,a)\big)}{\sum_{a' \in A} \exp\big(Q^*(s,a')\big)}. \tag{2}$$

where $\pi(a|s)$ denotes the probability of taking the action a from action space A at state s, and $Q^*(s,a)$ denotes the optimal Q-value for action a in state s. Here we also use neural networks to represent policy function $\pi(s,a|\theta^\pi)$ and Q-value function $Q(s,a|\theta^Q)$. Each neural network takes in the mPFC neural activity s, and current action a, and generates the instantaneous estimation on the probability and Q-value of taking the action a in state s of respectively. Under this assumption and from the derivation of Markov Decision Process (MDP) formulation, the outputs of the reward function, policy function $\pi(s,a|\theta^\pi)$ and Q-value function $Q(s,a|\theta^Q)$ follows the Inverse Action-value Iteration (IAVA) as in [16]:

$$r(s,a) = \eta_s^a + \frac{1}{n-1} \sum_{b \epsilon A_{\bar{a}}} r(s,b) - \eta_s^b, \tag{3}$$

$$\eta_s^a := \log\big(\pi(a|s)\big) - \gamma \max_{\tilde{a}} Q^*(\tilde{s}, \tilde{a}), \tag{4}$$

where γ is the discount factor and n is the number of actions. $Q^*(\tilde{s}, \tilde{a})$ denotes the optimal Q-value for action $\tilde{a}$ in next state $\tilde{s}$. This IAVA equation describes the relationship between policy $\pi(a|s)$ and reward $r(s,a)$. Given the expert's policy at each state, the corresponding reward distribution can be obtained by solving the system of linear equations.

We trained 3 neural networks to iteratively follows the IAVA. Firstly, we train the policy neural network $\pi(s,a|\theta^\pi)$ by observing the action taken in each state. Here we apply a Kalman filter to generate the expert trajectory from M1 activity for the IRL method. And the action at each time instance is chosen as directional velocity with fixed step size from the expert trajectory. We use the cross entropy as the loss

function and train the policy network $\pi(s,a|\theta^{\pi})$ using stochastic gradient descent. Then we fix the policy network $\pi(s,a|\theta^{\pi})$ and train the reward neural network $r(s,a|\theta^{r})$ and Q-value neural network $Q(s,a|\theta^{Q})$ iteratively. At each time instance, we randomly choose an action according to policy function, and update the parameters in the reward network $r(s,a|\theta^{r})$ using the mean square error (MSE) loss function, which is defined as the error between the value calculated from IAVA equation in eq.(xx) and the output of neural network. Then, we use Bellman optimality equation:

$$Q^{*}(s,a) = r(s,a) + \gamma \max_{\tilde{a}} Q^{*}(\tilde{s},\tilde{a}), \qquad (5)$$

to calculate the desired Q value and update the Q-value network $Q(s,a|\theta^{r})$ using MSE loss function. Both reward and Q-value neural networks are trained by stochastic gradient descent method. Each network has one hidden layer with optimal number of hidden PEs explored using 4-fold cross validation. The nonlinearity function of each PE is set as sigmoid function. And each neural network is trained with 10 times random initializations.

C. *Validation of the internal reward model*

We validate the reward model by using it as an internal critic for the RL decoder, which learns the mapping from M1 activity to movement intention. Here we use real data to simulate an online brain-machine interfaces scenario, in which subjects need to brain control the cursor to move towards the target. The RL decoder is based on the scheme of attention-gated reinforcement learning (AGREL) [17]. The input is 16-channel M1 neural activity, the output is the moving direction of the cursor. The update of the AGREL is using the internal reward information obtain from section II-B. The details of applying AGREL to movement decoding can be referred to [6].

In order to validate the effectiveness of the reward function we get from IQL algorithm, we compare the decoding performance of both using external and internal instantaneous reward. The external instantaneous reward is defined as:

$$r_{ex} = \begin{cases} 1, \Delta d(t) < 0 \\ 0, \Delta d(t) \geq 0. \end{cases} \qquad (6)$$

where $\Delta d(t) = dist(t-1) - dist(t)$ and $dist(t)$ is the distance between the target and the decoded position at time t. This reward function will give a positive reward to the decoder when the cursor is approaching the target, otherwise the reward is 0. The decoding performance is evaluated by the convergence rate of training and the testing decoding accuracy.

III. RESULT

In this section, we apply the IQL algorithm to estimate the reward function from mPFC activity and validate using RL decoder in II-C to check performance in BMI applications. We first present the trajectories generated by Kalman filter. Then we show the estimated results of reward function based on IQL algorithm. Finally, we compare the decoding performance of the RL decoder using our internal instantaneous reward with external reward.

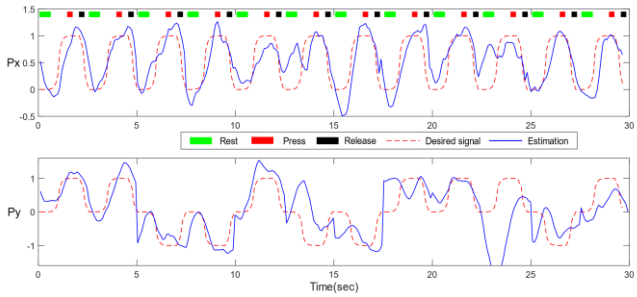

Fig 2. Generation of trajectory using Kalman Filter.

A segment of expert trajectory using Kalman filter is shown in Fig 2. The x-axis is the time and the y-axis is the coordinate of position. The red dash line is the desired movement signal obtained by smoothing the behavioral events, which are labelled in green, red and black bars on the top. The state rest, pressing high lever and pressing low lever are represented by 2D coordinates [0,0], [1,1] and [1, -1]. The blue solid line is the output of the Kalman filter. We can observe that the estimation can follow the change of desired signal very well, which allows the stochasticity in the expert trajectory. We perform the Kalman filter on 20 segments of data, each containing 250 data samples.

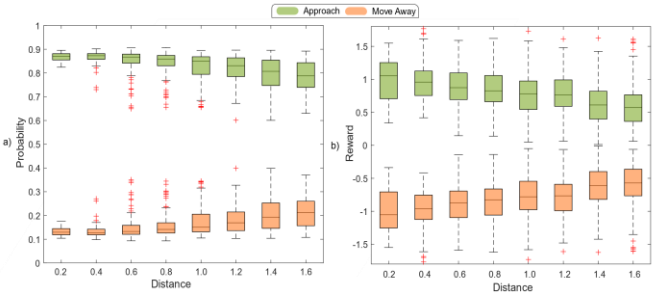

Fig 3. (a) Probability of taking each action at difference distance (b) Instantaneous reward of taking each action at difference distance

Given the estimated kinematics and corresponding mPFC activity of each trial, we apply the IQL algorithm to infer the internal reward function. We use 70% of the trials for training and 30% of the trials for testing. Here the testing is to find the best hyperparameter including learning rate and number of hidden units for each neural network. In Fig 3, we show the policy function and reward distribution obtained from IQL algorithm. We first label all the mPFC neural data with the distance between target and cursor and then calculate the output through policy and reward network. Fig. 3a shows the boxplot of policy value $\pi(a|s)$ and reward value $r(s,a)$ for different distances. The x-axis is the distance, and the y-axis is the probability (a) and instantaneous reward (b) of two action (green boxes represent approaching action and orange boxes represent moving away action from the target). Each distance value represents a distance interval (for example, 0.2 corresponds to the distance interval (0,0.2]). We can observe that, as the distance decreases, the probability of taking the action approach is increasing. This higher probability obtained from behaviors leads to a higher instantaneous reward in IQL algorithm. Compared with the artificial design reward, which could be a constant value across the distance, the estimated reward distribution matches better the subject's customized behavior patten and dynamics of neural activity.

We validate the estimated reward function by using as an internal critic to train the RL decoder. We choose 60% of the data for training, 20% for validation and the rest for testing. The learning rate β and number of hidden units M are explored to have

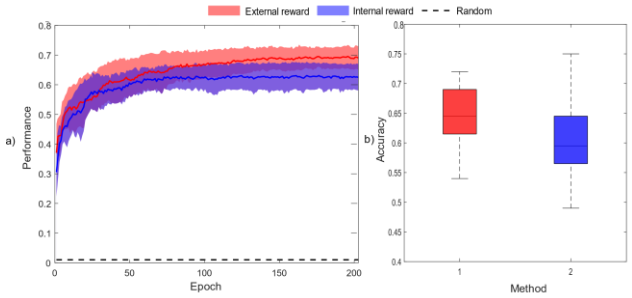

Figure 4. (a) Learning curve of the decoder using external and internal reward (b) Testing results of two methods.

the best performance in the validation set ($\beta = 0.001$ and $M = 5$). The weights of AGREL are randomly initialized for 20 times and the input data is shuffled in each initialization. As shown in Fig 4(a), we plot the learning curve of the RL decoder in the training stage. We train the decoder with external (red curve) and internal instantaneous reward (blue curve), respectively. The x-axis is the epoch, and the y-axis is the success rate of the trials for validation. Solid curve is the mean of success rates, shadow area is the standard deviation of success rates. We can observe that the learning curves of two methods are very close to each other and there is no significant difference in terms of the convergence rates. Both methods can achieve high performance after convergence, which is much higher than the chance level (black dash line). The chance level is testing by taking random action at each step, with a success rate about 0.1%. Fig. 4(b) plots the statistical decoding results in the testing stage across 20 random data shuffles. The average testing performance of decoder using external reward is $64.9\% \pm 5.0\%$, which is higher than that using internal reward with $60.1\% \pm 6.2\%$ (p=0.012). This slight drop of decoding performance can be explained by the noise embedded in mPFC activity recording which makes some estimation of reward not accurate. Overall, we conclude that the decoder using estimated reward function can achieve a high performance as well as that using external reward. All above results reveal that our internal reward model could extract internal critic from mPFC activity through trial and error, which ensures the high performance of the RL decoder.

IV. Conclusion And Discussion

To develop autonomous learning BMIs, it is important to extract internal reward from neural activity efficiently. In this paper, we apply the IRL method to estimate the internal reward function of mPFC cortical activity and verify the effectiveness of the estimated reward function on a RL movement decoder. The IQL algorithm is applied to learn the mapping from mPFC activity to instantaneous reward. The testing results demonstrate that the decoder using internal instantaneous reward can achieve a satisfactory performance compared with that using external instantaneous reward. This preliminary result shows the feasibility of applying IRL methods to estimate subjects' internal evaluation from neural activity. In the future work, we plan to explore more subjects and improve the design to develop a closed-loop autonomous BMI framework.

References

[1] M. A. Lebedev and M. A. L. Nicolelis, "Brain-machine interfaces: past, present and future," *Trends Neurosci.*, vol. 29, no. 9, pp. 536–546, 2006, doi: 10.1016/j.tins.2006.07.004.

[2] X. Zhang *et al.*, "Clustering Neural Patterns in Kernel Reinforcement Learning Assists Fast Brain Control in Brain-Machine Interfaces," vol. 27, no. 9, pp. 1684–1694, 2019.

[3] J. Digiovanna *et al.*, "Coadaptive Brain – Machine Interface via Reinforcement Learning," vol. 56, no. 1, pp. 54–64, 2009.

[4] L. R. Hochberg *et al.*, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–171, 2006, doi: 10.1038/nature04970.

[5] J. Wessberg *et al.*, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," no. 1, pp. 361–365, 2000.

[6] Y. Wang, F. Wang, K. Xu, Q. Zhang, S. Zhang, and X. Zheng, "Neural control of a tracking task via attention-gated reinforcement learning for brain-machine interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 3, pp. 458–467, 2015, doi: 10.1109/TNSRE.2014.2341275.

[7] X. Shen, X. Zhang, Y. Huang, S. Chen, and Y. Wang, "Task Learning over Multi-Day Recording via Internally Rewarded Reinforcement Learning Based Brain Machine Interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 3089–3099, 2020, doi: 10.1109/TNSRE.2020.3039970.

[8] B. Mahmoudi and J. C. Sanchez, "A symbiotic brain-machine interface through value-based decision making," *PLoS One*, vol. 6, no. 3, 2011, doi: 10.1371/journal.pone.0014760.

[9] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," *Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004*, pp. 1–8, 2004, doi: 10.1145/1015330.1015430.

[10] V. Marivate, "I Nverse R Einforcement L Earning and a Pprenticeship," *Learning*, no. June, pp. 128–135, 2015.

[11] T. U. Hauser *et al.*, "Temporally dissociable contributions of human medial prefrontal subregions to reward-guided learning," *J. Neurosci.*, vol. 35, no. 32, pp. 11209–11220, 2015, doi: 10.1523/JNEUROSCI.0560-15.2015.

[12] D. R. Euston, A. J. Gruber, and B. L. McNaughton, "The Role of Medial Prefrontal Cortex in Memory and Decision Making," *Neuron*, vol. 76, no. 6, pp. 1057–1070, 2012, doi: 10.1016/j.neuron.2012.12.002.

[13] S. W. Kennerley and J. D. Wallis, "Evaluating choices by single neurons in the frontal lobe: Outcome value encoded across multiple decision variables," *Eur. J. Neurosci.*, vol. 29, no. 10, pp. 2061–2073, 2009, doi: 10.1111/j.1460-9568.2009.06743.x.

[14] M. Shidara, B. J. Richmond, M. Shidaral, and B. J. Richmond, "Anterior Cingulate : Single Neuronal Signals Related to Degree of Reward Expectancy Published by : American Association for the Advancement of Science Linked references are available on JSTOR for this article : Anterior Cingulate : Si Neuronal Signals Rel," vol. 296, no. 5573, pp. 1709–1711, 2002.

[15] A. Jahn, D. E. Nee, and J. W. Brown, "The neural basis of predicting the outcomes of imagined actions," *Front. Neurosci.*, vol. 5, no. NOV, pp. 1–7, 2011, doi: 10.3389/fnins.2011.00128.

[16] G. Kalweit, M. Huegle, M. Werling, and J. Boedecker, "Deep inverse Q-learning with constraints," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, 2020.

[17] P. R. Roelfsema and A. Van Ooyen, "Attention-gated reinforcement learning of internal representations for classification," *Neural Comput.*, vol. 17, no. 10, pp. 2176–2214, 2005, doi: 10.1162/0899766054615699.