Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777693

MATHEMATICAL MODELING OF NEURAL ACTIVITY BASED ON OPTIMAL REINFORCEMENT LEARNING

V. Krishnan
PG & Research Department of Mathematics, Jamal Mohamed College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli -620 020, Tamilnadu, India
Email ID: vkrishnan1987@gmail.com

Abstract:

The brain is a very complex persistent neural network. The Reinforcement Learning (RL) theory has become one such approach applied in studies of brain-and-machine interfaces. We design experiment-based neural models to enable information processing system. In present study we proposed a well-organized learning technique, such as attention-gated (AG) reinforcement learning, to use a three-layer neural network to instantly understand the neuronal position at each time of action compilation. Three models discussed in this study had similar neural (firing) inputs, and similar neural network structures (nonlinear), but dissimilar policies to pick the weights and actions. The TARs of the three models demonstrated that attention-gated (AG) reinforcement learning has higher TAR values as compared to Q-greedy, and Q-softmax. The decoders begin to track the non-stationary fresh neural information every day, after the adaptation of one data segments, it showed better performance. Attention-gated (AG) reinforcement learning shows decoding ability while maintaining the performance of non-stationary nerve activity over several days of recording. The RL-based BMI architecture is an effective reinforcement learning method for designing adaptive neural decoders in a sophisticated process space that accelerates performance and reliably improves performance in complex artificial control tasks.

Key words: Neural control, Brain-machine interfaces (BMIs); Trajectory tracking; Attention-gated reinforcement learning

Introduction:

The brain is a very complex persistent neural network. In contrast, the information-processing paradigms that dominate computational neuroscience are shallow structures that perform simple mathematical operations. In neuroscience, the model describes how the nervous system is physically organized and/or how its function changes dynamically over time. The Reinforcement Learning (RL) theory has become one such approach applied in studies of brain-and-machine interfaces (BMI). Prior knowledge of the environment model influences the RL system to operate in an unrestricted environment. Empowerment learning (EL) is a process of adaptive nature that uses animal models to explore past experiences to develop the outcome of future options. Chavarriaga and Millán used errors (ErrP) linked to EEG capabilities that were developed as reward signals to reduce the risk of a defect [1]. Digiovanna was the first to

discover the experimental RL-based BMI model in which mice were trained to control the brain with an artificial arm in two targeted options using Q (λ) Learning [2]. Sanchez Q (λ) cautioned against expanding the co-adaptive architecture of Primate Testbed, which performs an egress center task [3]. However, these studies simplify decoding motions into which classification arrives within an experiment. Kaelbling et al. in his study discussed three models based on how the reward for optimism is considered [4]. We design experiment-based neural models to enable information processing system. In present study we propose a well-organized learning technique, attention-gated (AG) reinforcement learning, to use a three-layer network of neural to instantly understand the neural location at every time index into a rich action compilation.

Reinforcement Learning Theories:

Mathematical modeling of reinforcement learning (RL) plays a crucial role in the budding fields of *Neurotechnology*. Prior studies demonstrated RL as the "Markov Decision Process" (MDP) with "Q-Learning" process. This is explained by a fixed set of states (St)**,** actions (Ac); and a transition function (Tr) which allocated to every state and action pair a possibility distribution over the state, and reward functions (Rw).

$$\mathrm{Tr}(\mathrm{Tr}: \mathrm{St} \times \mathrm{Ac} \rightarrow \Pi\,(\mathrm{St}))$$

$$\mathrm{Rw}(\mathrm{Rw}: \mathrm{St} \times \mathrm{Ac} \rightarrow \mathrm{Rw})$$

In line with the theories, the Markov Q-Learning algorithm combines the possibility of the Q-learning algorithm with the maximum action value Q* to model the decision-making process for consistent learning difficulty. Make a map accordingly and, more precisely, a Q-study computes a list of all values, called the Q-Stable, a constant estimate of Q(St, Ac). Q (St, A) is defined as a prediction under the hypothesis that agents execute activity among state, and then the activities that are most rewarding are always selected. Each function pair, Q(St, Ac) is started with arbitrary value and then reorganized at every steps t($t > 0$), the action (a_t) executed in state (s_t) according to the following equation:

$$Q(s_{t-1}, a_{t-1}) \leftarrow Q(s_{t-1}, a_{t-1}) + \delta(\gamma \max_{a \varepsilon A} Q(s_t, a) - Q(s_{t-1}, a_{t-1}) \quad (1)$$

The term β is the learning rate $1>\delta>0$; γ was the discount factor that influence how many rewards are counted; RW is the reward gained for action performing at−1 in state s_{t-1}; s_t is the state meet after performing action at−1 in s_{t-1}. Q-learning process is straightforward to execute however it is impracticable for problems of attention since the dimension of the Q-table (|S|×|A|) go up rapidly for problem sampling, thus, require oversimplification of the process. The study learning performance is measured using three main dimensions: (i) ultimate convergence to optimal- many algorithms come with a guarantee of asymptomatic conversion for optimal behavior, (ii) Rate of conversion with optimality - the closest maximum completion speed, rather than good behavior from the start, is the expected reduction in the reward of implementing a learning algorithm that errors occur somewhere during the race.

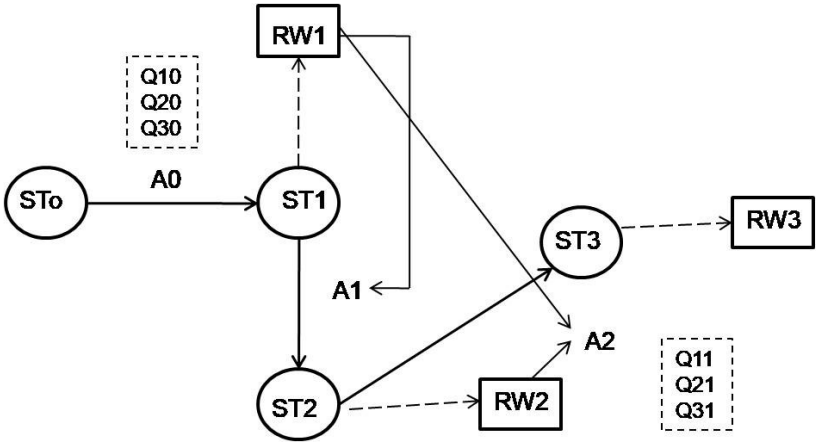

Figure 1: Reinforcement Learning model, ST= state; RW =Reward; A=Action

1. **Model-based reinforcement learning:** Several reinforcement learning (RL) techniques adopted from the default global model. Model-based RL understand the uncertainty of the sequence of events and activities in a task (resulting in the processes that follow, e.g., different paths in a given way) which could be use dynamically and adaptively to decide perfect activities by simulating their events of consequences. The model-based approach first used the calculation transfer (p) and cost functions (c) and used them to calculate the value (v). In contrast, a model-free approach can formulate responses regardless of the transfer activity or reward function.
2. **Value-based:** In this architecture, the decision-maker maintains reasonable value estimates starting in each state in the environment, and updates those estimates when they have a new experience. The maximum value is calculated by repeating the value. Any such algorithm has been shown to better integrate into the maximum value estimation, which can be used to generate maximum behavior. Technically, value-based RL can be applied to model-based RL. Convergence theory is very important when using value function based RL. Asynchronous dynamic programming integrates into the maximum value function. Value-based model could be restructured based on the various forms of information and data. They adjust information based on the penalty or reward collected by the subjects after every action. Value-based RL will be not change, so there is no need for learning if current selection actions always accurately predict current function values. Or else, current function values should be customized to reduce errors in reward predictions. The signature difference between the actual reward and expected reward by the current value functions and is called the error of the reward prediction. Error of reward prediction is the main way of changing class reinforcement learning algorithms and value functions, which is called learning simple or model-free empowerment. Specifically, the value activity of the animal activity or action is upgraded on the basis of the error of the prediction, while the activity values for all the actions and states remains unaffected or remains inactive.
3. **Policy-based:** A policy is applied for function estimate. The policy is a greedy policy like neural network, for example actor-critic model. In value-based model, a minute variation in value can influence action selected. However, this is not the case in greedy

policy. Apart from these, in off-policy model, an agent cannot pick actions and learns from experts and sessions (recorded data). Dual Q-learning process is an off-policy reinforcing learning algorithm where another policy is used for evaluation rather than selecting the next action. In practice, two different value functions are performed in parallel with each other using specific experiments.

Experimental Setup in a RL Diagram

In present study, we planned to approve an competent reinforcement learning system for neural study, for instant attention-gated (AG) reinforcement learning that used to utilize a multi-layer neural network to instantly infer the neural activities arrived at every time into a affluent action assembly. Attention-gated reinforcement learning approach learns the efficient mapping from the communication on the basis of an instant reward other than the actual movements, which is clinically obtainable to the brain-and-machine interfaces (BMI). The neural system facilitate a easy-coding to discover sequential actions based on the probability system, also describe as rigorous function value according to the prediction error to strengthen the learning system by evaluating the unpredicted rewards, which add value to improve performance [2]. To test the competence and efficacy of the proposed method, we compare attention-gated (AG) reinforcement learning against Q-learning process with "α-greedy policy" (Q-greedy) implementation and policy of softmax for decoding accuracy, stability and convergence. All these models are useful to rebuild the path of a real brain-and-machine interfaces task with neural data, and additionally confirmed the adaptively based on multiple days's observation and recording.

We used attention-gated reinforcement learning to predict the path straightly from the recorded neural data. The neural data were collected from experimental setup. According to the characteristics of the path recorded in the brain-and-machine interfaces (BMI) task, the two dimensional behavioral data were clustered into seven actions, (left, right, up, down, Y axis position holding, X axis position holding and resting). The two holdings in X and Y directions were corresponding to the reflex actions. The direction of the neural actions in the accordance with neural states is accomplished by a three-layer neural network [5].

The unseen layer uses sigmoidal nonlinear functions. The output of the unseen layer, Z_k is express as:

$$Z_k = \frac{1}{1 + \exp(-\sum_{i=0}^{n-1} y_{ik}\ x_i)} \tag{2}$$

Here, y_{ik} is the weights; 'n' is no. of the input nodes; unseen layer surrounds a bias value: Z_0=1. Function states of path direct the no. of the output nodes. For every output node W_j responds to diverse actions. AG reinforcement learning approves the softmax stochastic rule to evaluate the neural activity. The prediction of choosing action is express as:

$$Ro(W_j = 1) = \frac{\exp(\sum_{k=0}^{m-1} g_{kj} Z_k)}{\sum_{k'=1'}^{B} \exp(\sum_{k=0}^{m-1} g_{kj'} Z_k)}$$

(3)

Here, g_{kj} the weight, whereas m is the no. of nodes in the unseen layer, that evaluated by exchanging computational values and the model performance. For Q-learning compare and implement two policies such as "Q-greedy" and "Q-softmax" [2]. However, all of the methods have similar neural system (nonlinear network), neural firing inputs, and the similar action ensemble productivity, however dissimilar policies to pick the errors and action. The 'Q-greedy policy' expressed as follow:

$$\mu(s_t) = \begin{cases} Ag \max Q(s_t), r(1-\varepsilon) \\ actionA(s), r(\varepsilon) \end{cases} \quad (4)$$

Where,

$$Qj(s_t, a_t) = \frac{1}{1 + \exp(-\sum_{k=0}^{m-1} g_{kj} Z_k)} \quad (5)$$

Where, ε = 0.01. The actions having maximum Q-value is selected with likelihood (1-ε). For Q-softmax, the activities are chosen based on the probability systems of the neural network.

$$Q_t(W_j = 1) = \frac{\exp\left(\left(\sum_{k=0}^{m-1} g_{kj} Z_k\right) \div \psi\right)}{\sum_{j'=1}^{B} \left(\left(\sum_{k=0}^{m-1} g_{kj} Z_k\right) \div \psi\right)} \quad (6)$$

Where, ψ is the temperature parameter, here ψ=0.05

$$\begin{aligned} X_{\Pr}(t) &= X_{pr}(t-1) + W(a_t) \\ Y_{pr}(t) &= Y_{pr}(t-1) + W(a_t) \end{aligned} \quad (7)$$

Where, $X_{Pr}(t)$=X(1) and $Y_{pr(t)}$=Y(1), are the primary point of real trajectories on the direction of X and Y.

The network weights, Y_{ik} & g_{kj}, were initiated arbitrarily between -0.1 to +0.1, and then adjusted until the error predictions meet to maximize reward signal. After the system obtains an instant reward, AG reinforcement learning represent a universal error signal which is computed as below:

$$\alpha_t = [2 - R_o(Wc(t) = 1]r(t+1) - 1 \quad (8)$$

Where, 'c' represent winning unit. Q-learning was used to evaluate the error using eligibility trace, λ:

$$\alpha_t = r_{t+1} + \lambda \max Q(S_{t+1,} a') - Q(S_t, a_t)$$

Here, λ is the discount factor

$$\Delta y_{ik} = \delta X_i Y_j f(\alpha)(1 - Z_k) \sum_{k=1}^{B} W_j g_{kj}) \quad (9)$$

Where, δ is the learning rates. For Q-learning, the network is taking the chronological dissimilarity error. The performance of brain-and-machine interfaces decoders is determined with the mean square error (MSE) and correlation coefficient (CC) between the actual and the predicted paths, which are extensively used in. Target acquisition rate (TAR) is utilized to evaluate skill of the decoders which understand patterns of the firing neural assembly to the projection activities [6, 7].

RESULTS

Attention reinforcement learning is used to evaluate the successive actions from actual neuron firing rates. The two dimensional data of behavior are assemble into seven activities or actions, (up-down, left-right, X and Y axis position holding, and resting). Each time neural activity is decoded immediately in an action or activity, and reward is determined according to the relative distance of the current to the concave crease position. The load is started arbitrarily at each stage and is upgraded regularly. After gaining weight at the end of the opening day, then set them as the initial values of the data class entered in the following days and keep updating. If the MSE is below the usual average of 0.1 with the desired path strength, detection is considered consistent.

Day 1 consists of four data classes and two data classes lasting 40 minutes in 2 days, 3 days and 6 days 20 minutes, respectively. We investigated the genetic parameters to evaluate the optimal combination of RL and Q-Learning (Q-Softmax, Q-Greed). The parameters were selected according to the correctness of the average performance selection to confirm the data in 50 convergent organizations. Note that the Q- learning indicates a delayed return only when each test is completed, but each time we decode the immediate motion [8, 9].

Study of the function of the brain and machine interfaces (BMI), we depicted the transformation process to describe how all the three models compute the neural tracking task. Q-learning and AG reinforcement learning decoders may not be able to perform the task at the start of the learning phase due to arbitrary initializations. In contrast, attention-gated (AG) reinforcement learning is able to shift to accurate directional from the holding or latent action timely, as learning can be intensified with unpredicted rewards. In final state of learning phase, all these methods learn to define neural positions for accurate actions, but Q-greedy processes still have limited amplitude of path rebuilding.

We found transition between the values of the output activities, which indicating that the weights develop to set in the input systems and the correct action learn policy. The sequential multi-units' actions of Channels that modulate signal variation for RIGHT, LEFT, UP and DOWN activities independently throughout 5 days. Table 1 lists the average target acquisition

rates (TAR) in 5 days for the corresponding actions. AG reinforcement learning beat two other decoders models at all commands represented by the right-tail paired t-test at p-value 0.05 significant level (AG reinforcement learning against Q-greedy: p= 0.010, and for AG reinforcement learning against Q-softmax: p= 0.0211).

Table 1: The mean of the target acquisition rates (TAR) of prediction movement in all the three models such as AG reinforcement learning, Q-greedy, and Q-softmax

	RIGHT	**LEFT**
Reinforcement learning	0.8528±0.1928	0.8989±0.1508
Q-greedy	0.5583±0.3109	0.7832±0.3255
Q-softmax	0.7705±0.0773	0.8159±0.1257
	UP	**DOWN**
Reinforcement learning	0.5193±0.2377	0.7596±0.1564
Q-greedy	0.7985±0.0987	0.8183±0.1560
Q-softmax	0.8418±0.1899	0.8943±0.0859

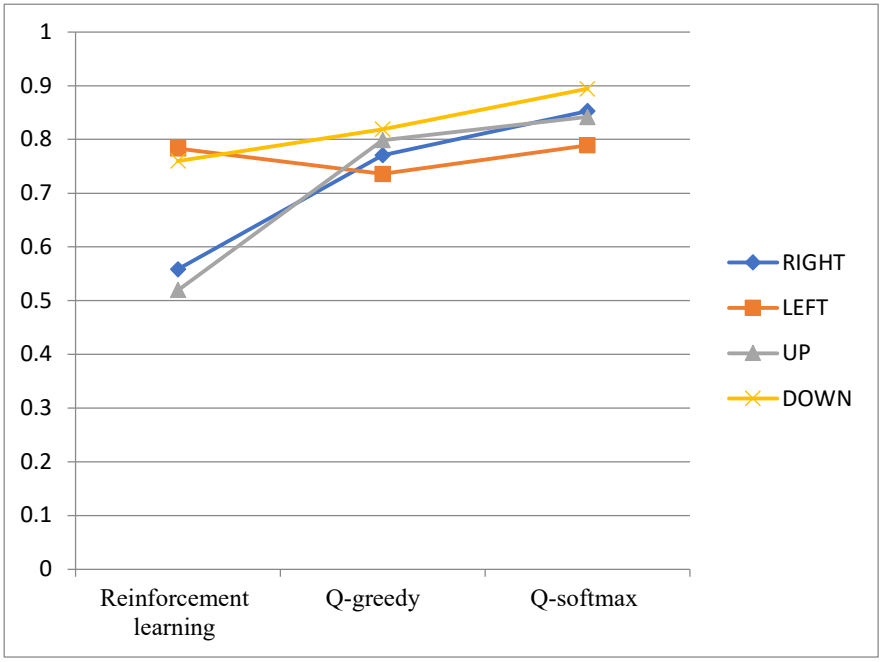

Figure 1: The TAR of prediction movement among the Reinforcement learning, Q-greedy, & Q-softmax

We recorded performance of the all three decoders data during all 5 day's setting. The values of the decoders found by the close of the earlier day are set as the starting values in the subsequent data sections and carry on updating process. The means and standard derivations were determined for Reinforcement learning, Q-greedy, and Q-softmax throughout 60 convergent initializations. Among all the models reinforcement learning was showing the best convergence and least initial effects. The compassion to the Q-learning initialization may outcome from the sequential variation between two consecutive assessments and total of predictions gradients. The data required for parameters upgrade of Reinforcement learning is available locally, that make reinforcement learning less reliant on the starting conditions and more competent. The standard correctness of X and Y axis by reinforcement learning enhance quickly than those of Q-softmax and Q-greedy representing the quicker convergence in neural learning. In this study we found that performance of decoding falls slightly at the initial phase of every five days of the experiments, but not like first day. Every decoder need less data segments (up to 3) to converge to high-quality performance in subsequent days. The average TAR achieved by Q-greedy, Q-softmax and AG Reinforcement learning throughout all 5 days demonstrated that attention-gated (AG) reinforcement learning has higher TAR values as compared to Q-greedy, and Q-softmax. The decoders begin to go after the movable form of the latest neural data for daily update and provide good performance after every adaptation of data segments.

Discussion and Conclusion:

Attention-gated (AG) Reinforcement learning represent model-based and flexible reinforcement learning, where we can altered value functions. These algorithms can control the value functions on the basis of the movement state of animals and its information of the surroundings without direct penalty or reward. Prior study demonstrated that animal can quickly change their behavior according to their current incentive status after the value of certain foods is reduced. It is often used as a test of goal-oriented behavior and implies that animals are indeed able to model-based flexible learning. Humans can also mimic the outcome of possible actions they can choose. This is known as paradoxical thinking, and the value of imaginative outcomes can be incorporated into the value functions by unselected measures when the results are predicted by the current value function. The difference between a hypothetical and a resultant error equal to an error is called a mythic or counter-prediction error.

To examine the benefits of AG reinforcement learning over other learning models, we unite Q-learning with softmax decision rule, and performed a comparison analysis of the three models (AG reinforcement learning, Q-greedy, and Q-softmax). All models combine the same neural firing input, the identical non-linear neural network configuration, and similar actions and functions ensemble but share different principles for choosing the rules for upgrading weights and actions. We found the similar phenomenon where learning is easily biased when a

particular action gives the highest reward in the early stages of Q-greedy [10]. However this bias can be removed by reinforcement learning on constantly choosing the accurate actions, it need more time for decoder to learn a new accurate action, which may lead to slow down performance. When we did comparison analysis, AG reinforcement learning approves a softmax policy to pick the activities based on the probability allocation of every possible action. Although the best actions are not possible to selected, the sub-optimal actions can be selected with a higher probability than the others, which helps to prevent sudden changes in performance. To additionally validate the benefits of the Softmax policy, we integrate Q-learning and focus on improving the effectiveness of decoding. However, attention-gated (AG) reinforcement learning still outshines Q-Softmax in term of TAR values, since attention-gated (AG) reinforcement learning also can evaluate a global error; it is actually calculated by dopamine neurons [5, 11, 12]. Extensive action defined with global error intensifies the reinforcement learning by evaluating unexpected rewards that contributes to the improvement of attention-gated (AG) reinforcement learning performance. For the above two reasons, attention-gated (AG) reinforcement learning demonstrates the ability to decode by maintaining movable neural activity after a number of days of recording. Attention-gated (AG) reinforcement learning based BMI design is a sophisticated process to design an adaptive neural decoder adapted to the space, an efficient reinforcement learning technique that accelerates implementation and consistently enhances performance in multifaceted artificial control functions.

It may be recognized that attention-gated (AG) reinforcement learning requires exploring a larger state-of-the-art space to collect appropriate mapping that is less successful than learning to supervise. Existing reinforcement learning methods adapt to the brain environment, resulting in neural activity. While the animal experimental model is doing the BMI task, the decoder is learning parallel to the brain function to output the accurate action to perform the task instantly. In future study the brain and machine interfaces (BMI) will observe the robotic arm without actual limbs, which imposes more dynamics of brain activity than everyday alteration. This challenge can be solved by optimistic attention-gated (AG) reinforcement learning and sophisticated RL algorithms as the decoder continue to update model-parameters based on trial-and-error method. Furthermore, neural activity adaptation can be modeled to predict future neural conditions and environmental dynamics can be used to improve the proficiency of brain and machine interfaces (BMI) applications.

Reference:

1. Chavarriaga, R. and J.d.R. Millán, *Learning from EEG error-related potentials in noninvasive brain-computer interfaces.* IEEE transactions on neural systems and rehabilitation engineering, 2010. **18**(4): p. 381-388.
2. DiGiovanna, J., et al., *Coadaptive brain–machine interface via reinforcement learning.* IEEE transactions on biomedical engineering, 2008. **56**(1): p. 54-64.
3. Sanchez, J.C., et al. *Control of a center-out reaching task using a reinforcement learning brain-machine interface*. in *2011 5th International IEEE/EMBS Conference on Neural Engineering*. 2011. IEEE.

4. Kaelbling, L.P., M.L. Littman, and A.W. Moore, *Reinforcement learning: A survey.* Journal of artificial intelligence research, 1996. **4**: p. 237-285.
5. Roelfsema, P.R. and A.v. Ooyen, *Attention-gated reinforcement learning of internal representations for classification.* Neural computation, 2005. **17**(10): p. 2176-2214.
6. Moritz, C.T., S.I. Perlmutter, and E.E. Fetz, *Direct control of paralysed muscles by cortical neurons.* Nature, 2008. **456**(7222): p. 639-642.
7. Millan, J.R. *On the need for on-line learning in brain-computer interfaces*. in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*. 2004. IEEE.
8. Tesauro, G., *Temporal difference learning and TD-Gammon.* Communications of the ACM, 1995. **38**(3): p. 58-68.
9. Zhang, W., *Reinforcement learning for job-shop scheduling.* 1996.
10. Belkhode, P.N., *Development of mathematical model and artificial neural network simulation to predict the performance of manual loading operation of underground mines.* Journal of Materials Research and Technology, 2019. **8**(2): p. 2309-2315.
11. Schultz, W., P. Dayan, and P.R. Montague, *A neural substrate of prediction and reward.* Science, 1997. **275**(5306): p. 1593-1599.
12. Schultz, W. and A. Dickinson, *Neuronal coding of prediction errors.* Annual review of neuroscience, 2000. **23**(1): p. 473-500.